Extraction et Complétion de Terminologies Multilingues

Valérie Hanoka

Université Paris Diderot (Paris 7)
École doctorale 132 : Sciences du Langage
U.F.R. de Linguistique

THÈSE DE DOCTORAT

Nouveau régime

Pour obtenir le grade de
Docteur en Sciences du Langage
Discipline : Linguistique Générale

En vue d'une soutenance publique
par

Valérie HANOKA-MAITENAZ

Le 6 juillet 2015

Extraction et Complétion de Terminologies Multilingues

Thèse sous la direction de
Laurence Danlos et Benoît Sagot

Jury

Pr.	Marie-Claude L'Homme	(rapporteur)	Université de Montréal
Dr. HDR	Mathieu Lafourcade	(rapporteur)	Université de Montpellier 2
Dr.	Bruno Gaume	(examinateur)	CLLE (ERSS) CNRS
Dr.	Benoît Sagot	(co-directeur)	Inria
Pr.	Laurence Danlos	(directrice)	Université Paris Diderot

Table des matières

Table des figures

Liste des tableaux

Résumé

Les processus d'extraction terminologique automatique ont été jusqu'ici majoritairement conçus pour être appliqués à des corpus monolingues et dans des registres de langue uniformes. Cette thèse, réalisée dans le cadre d'une convention CIFRE, prolonge cet objectif pour une application à des données textuelles bruitées et issues de langues de plus en plus variées, pour l'extraction de « termes de terrain ».

Ce travail s'inscrit dans le cadre de l'analyse de verbatim issus d'enquêtes internes au sein de multinationales traitées par l'entreprise *Verbatim Analysis - VERA* ; il consiste à élaborer une séquence de traitements pour l'extraction automatique de terminologies qui soit faiblement dépendante de la langue, du registre de langue ou du domaine.

Suivant une réflexion fondée sur différents aspects de typologie linguistique appliquée à sept langues, nous proposons des prétraitements textuels préliminaires à l'entraînement de modèles. Ces derniers sont soit indispensables (segmentation en tokens), soit optionnels (amputation d'une partie de l'information morphologique). Sur l'ensemble des données ainsi produites, nous calculons des traits numériques (statistiques ou fréquentiels) pour l'entraînement des modèles statistiques de type CRF. Nous sélectionnons un ensemble de meilleurs modèles grâce à une évaluation automatisée, au moyen d'une métrique adaptée, des termes extraits par les modèles produits pour l'ensemble des cadres expérimentaux envisagés pour chaque langue. Nous réalisons alors une seconde série d'évaluations pour étudier l'exploitabilité de ces modèles pour d'autres langues que celles sur lesquelles ils ont été entraînés. Il ressort de ces expériences que cette méthode aboutit à une extraction de termes de terrain de qualité satisfaisante. Les meilleurs scores obtenus (pour une évaluation monolingue des modèles) se situent, pour la majorité des langues, au-dessus de l'isoligne de f-score $0,9$. Ces scores peuvent même être améliorés pour certaines langues

grâce à l'application trans-lingue des meilleurs modèles d'autres langues ; il en ressort que notre approche constitue potentiellement un bon levier à des extractions terminologiques pour des langues ne disposant pas de leurs propres modèles.

La seconde partie de notre travail présente nos travaux relatifs à la complétion automatique de terminologies structurées multilingues. Nous avons proposé et évalué deux algorithmes de complétion qui prennent en entrée un graphe de traduction multilingue (que nous construisons à partir de ressources libres) et une terminologie multilingue structurée. Ils proposent alors de nouveaux candidats termes pour cette dernière. Notre approche permet de compléter la terminologie structurée dans une langue qu'elle couvre déjà, mais également d'étendre sa couverture à de nouvelles langue. L'un de ces algorithmes est également appliqué au wordnet du français WOLF, ce qui en permet une amélioration importante de la couverture.

Mots-clefs

Extraction terminologique multilingue, extension de terminologie multilingue, graphe de traduction, traitement automatique des langues.

Abstract

Until now, automatic terminology extraction techniques have been often targeted towards monolingual corpora that are homogeneous from a language register point of view. This work, carried out in the context of a CIFRE convention, extends this objective to non-edited textual data written in typologically diverse languages, in order to extract « field terms ».

This work focuses on the analysis of verbatim produced in the context of employee surveys carried out within multinational companies and processed by the *Verbatim Analysis - VERA* company. It involves the design and development of a processing pipeline for automatically extracting terminologies in a virtually language-independent, register-independent and domain-independent way.

Based on an assessment of the typological properties of seven diverse languages, we propose a preliminary text pre-processing step prepares the training of models. This step is partly necessary (tokenization) and partly optional (removal of part of the morphological information). We compute from the resulting data a series of numerical features (statistical and frequency-based) used for training statistical models (CRFs). We select a first set of best models by means of an automatic dedicated evaluation of the extracted terms produced in each of the experimental settings considered for each languages. We then carry out a second series of evaluations for assessing the usability of these models on languages that differ from their training languages. Our results tend to demonstrate that the quality of the field terms that we extract is satisfying. The best scores we obtain (in a monolingual setting) are above $0,9$ for most languages. These scores can even be further improved for several languages by using some of the best models trained on other languages ; as a result, our approach could prove useful for extracting terminologies in languages for which such models are not available.

In the second part of our work, we describe our experiments on automatic extension of multilingual structured terminologies. We have introduced and evaluated two extension algorithms that take as an input a multilingual translation graph (which we created based on free resources) and a multilingual structured terminology. They produce new candidate terms. Our approach allows to extend the coverage of the structured terminology both in a language it already covers and to new languages not yet covered. We also apply one of these algorithms to the extension of the French wordnet WOLF, which led to a significative increase of its coverage.

Keywords

Multilingual Terminology Extraction, Multilingual Terminology Extension, Translation Graph, Natural Language Processing.

Remerciements

Je remercie chaleureusement :

Benoît Sagot, mon Directeur de thèse, pour son enthousiasme permanent, sa curiosité naturelle débordante, sa disponibilité et son aide efficace.

Dimitri Tcherniak, président co-fondateur de l'entreprise *Verbatim Alanysis - VERA* pour son accueil chaleureux, sa patience et la grande liberté qu'il m'a accordé pour mon travail de recherche.

Marie-Claude L'Homme et Mathieu Lafourcade qui m'ont fait l'honneur d'être rapporteurs de cette thèse.

Bruno Gaume pour l'intérêt qu'il a témoigné envers ce travail de thèse en acceptant d'être examinateur, et pour son sympathique accueil dans le laboratoire CLLE-ERSS lors de ma visite en juin 2013. Je remercie également toute son équipe, et en particulier Emmanuel Navarro pour ses conseils.

Laurence Danlos, et l'ensemble de l'équipe d'Alpage, pour leur accueil. Au sein de ce laboratoire, j'ai notamment eu la chance de pouvoir échanger avec Pierre Magistry et Djamé Seddah, que je remercie pour leurs conseils avisés. Dans des styles très différents, vous avez souvent été des bouffées d'air frais pour l'animal stressé que je suis ! Merci également à Juliette Thuilier pour son aide méticuleuse et bienveillante, et surtout son soutien moral discret mais efficace.

Danielle Candel, Corentin Ribeyre, François Guérin et Emmanuel Costa-Drolon pour des pistes bibliographiques, résolutions de problèmes et relectures.

Les jeunes gens que j'ai eu le plaisir de côtoyer durant ces quelques années passées à P7 : Chloé, Sarra, Eve, Rachid, Luc, Charlotte, Diego, Rosa et Enrique.

L'irremplaçable Marianne Djemaa, pour être qui elle est, et faire ce qu'elle fait. Tout simplement.

Marion Baranes pour sa bonne humeur communicative, son support constant et ses multiples qualités humaines qui ont illuminé des périodes parfois sombres, et ses dons royaux de nourriture qui ont si bien consolé mon estomac chafouin.

Fanny Grandry et Mathieu Huin, des potes, des vrais. Merci.

Brebiche et toute ma famille, pour leur amour et leur soutien indéfectible.

À Sylvie, ma mère.

Introduction Générale

Problématique

Il arrive régulièrement que des travaux menés en linguistique pour une langue donnée ou plusieurs langues apparentées soient généralisés à un ensemble bien plus vaste de langues. Or, comme le fait remarquer Gil (2001), la linguistique contemporaine a été développée et pratiquée principalement par des locuteurs de langues européennes. À l'heure actuelle, la plupart des grandes théories et applications linguistiques sont toujours dévolues de façon disproportionnée aux principales langues d'Europe, et singulièrement à l'anglais, même si cette tendance tend à s'estomper. Cet eurocentrisme est également très prégnant dans le domaine de la terminologie, ce qui a eu des répercussions sur les pratiques mises en œuvre en extraction terminologique automatique.

La majorité des outils de traitement automatique des langues développés à l'heure actuelle se concentrent d'une part sur ces quelques langues dites « bien traitées » et d'autre part sur des données textuelles bien formées (grammaticales, correctement orthographiées et contenant un minimum d'erreurs). Or ce genre de données constitue une faible proportion des données textuelles pouvant être recueillies, notamment dans le cadre professionnel (enquêtes clients, notes internes, réponses aux questions ouvertes dans le cadre de sondages etc.), ou même plus généralement sur Internet et dans le cadre de communications interpersonnelles (Ittoo & Bouma, 2013). La prolifération de données textuelles bruitées dans des ensembles de langues de plus en plus variées est un défi pour le traitement automatique des langues.

Ces constatations sont également valables dans le cadre de la terminologie et de l'extraction terminologique automatique. L'objectif de cette thèse consiste, dans un premier temps, à explorer une approche pour l'extraction des éléments « terminologiques » pertinents à partir de verbatim dans des langues typologiquement diverses. Ces derniers sont issus de réponses à des questions ouvertes dans des enquêtes internes que des multinationales ont proposées à leurs

employés. Dans un second temps, il s'agira de développer et d'utiliser un gros graphe multilingue pour l'extension de la terminologie extraite, aussi bien pour améliorer sa couverture pour des langues déjà prises en compte que pour l'ajout de langues supplémentaires.

Contexte de cette recherche

Cette thèse s'est déroulée dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE) en partenariat avec la société *Verbatim Analysis – VERA*. Cette dernière est spécialisée dans des systèmes multilingues de traitement automatique d'enquêtes d'opinion.

Les sondages d'opinion peuvent concerner la satisfaction de clients ou d'employés, sonder des opinions politiques ou autre. En règle générale, ils se présentent sous la forme d'un questionnaire en trois parties :

- renseignements de données signalétiques (age, sexe, lieu de résidence ou de travail...) ;
- réponse à des questions d'opinion de type QCM ;
- facultativement, une ou plusieurs questions ouvertes dont la réponse est rédigée par le répondant dans la langue de son choix.

Ces verbatim sont une source de renseignements précieux car il s'agit non plus d'informations *quantitatives*, comme dans les questions fermées, mais de données *qualitatives* couvrant souvent un nombre important de thématiques variées qui ne sont pas abordées dans les autres parties de l'enquête. Par ailleurs, ces verbatim permettent souvent d'expliquer les réponses aux questions fermées et d'insister sur les problématiques d'importance.

Avant le développement de solutions dédiées à la fouille de texte, il était impossible aux spécialistes des enquêtes d'opinion d'extraire des informations riches sur de grandes quantités de données qualitatives autrement que manuellement. Bien qu'effectuée par des experts, cette activité restait coûteuse en temps et ne débouchait pas sur un apport d'information entièrement fiable (biais de subjectivité) : d'une part, il était impossible de lire l'ensemble des commentaires pour dégager toutes les thématiques pertinentes, les commentaires étant souvent bien trop nombreux [1] ; d'autre part, l'identification dans chaque commentaire des thématiques abordées était réalisée par de multiples analystes. La cohérence des annotations était alors un problème important, en particulier pour les commentaires abordant plusieurs sujets ou des sujets moins courants ; enfin, comme pour toute annotation manuelle, le taux d'erreur était significatif, quoique difficile à évaluer. Or dans le milieu industriel, et malgré l'ancienneté des travaux dans ce domaine (Lebart & Salem, 1994), ce travail d'analyse reste souvent manuel.

1. Typiquement, les enquêtes internes traitées par Verbatim Analysis - VERA vont de quelques centaines à quelques centaines de milliers de verbatim, rédigés jusque dans plusieurs dizaines de langues

Ces tâches étaient et restent aujourd'hui encore très difficiles à mettre en œuvre dans le cadre des enquêtes internes. Ces enquêtes, comparées par exemple aux enquêtes client, comportent en effet des questions fermées relatives à des données signalétiques complexes (position dans l'entreprise, lieu de travail, etc), dont la structure, souvent hiérarchique ou matricielle, est aussi complexe que celle de l'entreprise elle-même. De plus, le champ des sujets abordés est bien plus vaste (bien qu'il soit relativement comparable d'une enquête à une autre) et les problématiques étudiées sont très variables d'une équipe à une autre, d'un pays à un autre, d'un site à un autre.

Afin de pouvoir classifier les thématiques, et détecter les problématiques d'intérêt pour chaque enquête, la stratégie utilisée par VERA consiste à tenir à jour un lexique structuré de termes multilingues, que l'on peut qualifier d'ontologie légère. Cette ressource étant la pierre angulaire du reste des traitements, sa mise à jour et les possibilités d'extension à de nouvelles langues est un enjeu crucial pour la société. L'objectif consiste à développer un système d'extraction terminologique pouvant être appliqué sur toutes les langues susceptibles d'être traitées dans le cadre d'enquêtes internes pour des multinationales. Dans un second temps, notre but est de proposer un moyen de rattacher des termes dans une structure de lexique multilingue préexistante, qu'il s'agisse de termes nouveaux dans une langue déjà couverte ou de termes issus d'une langue non encore couverte.

Motivations et Contributions

Comme nous l'avons déjà évoqué, la tradition terminologique a jusqu'ici été appliquée sur des textes plus ou moins normalisés, dans des langues uniformes. Or, avant même de viser un fort multilinguisme, on peut déjà constater que lorsque un traitement est appliqué à une langue en particulier, il n'est appliqué souvent qu'à la version « standard » de cette langue. Cette approche est appropriée dans le cadre d'une extraction terminologique normative, mais ne l'est plus dès lors qu'il s'agit de traiter des données de terrain.

La finalité de cette recherche est d'offrir une séquence de traitements flexibles pour l'extraction automatique d'une terminologie faiblement dépendante de la langue, des registres de langue ou du domaine. Par ailleurs, nous proposerons l'utilisation d'un graphe de traduction fortement multilingue pour permettre de suggérer de nouveaux candidats termes dans différentes langues.

Structure de la thèse

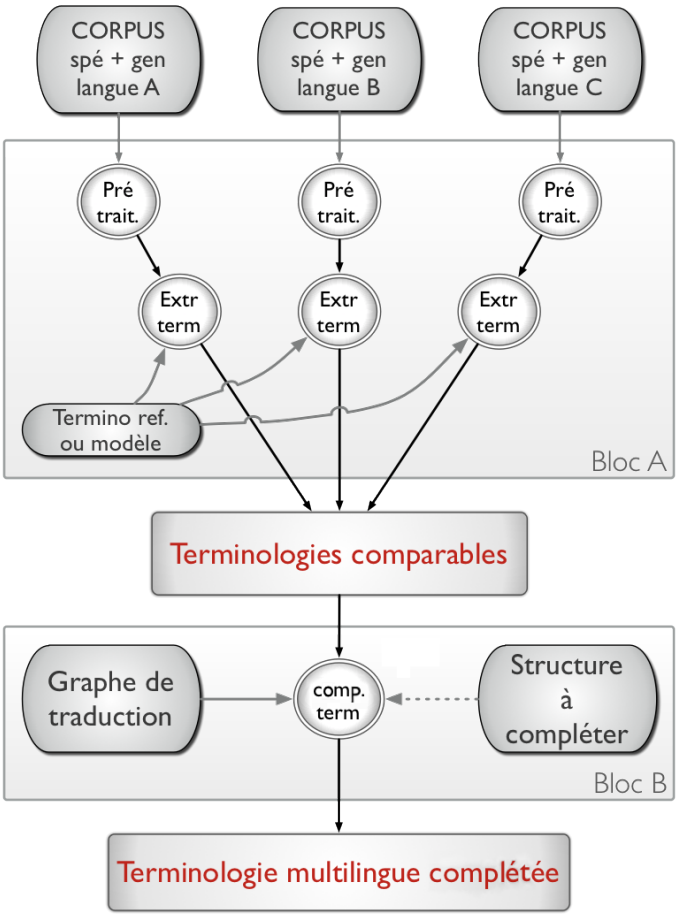
Cette thèse est organisée en 11 chapitres, regroupés en 3 parties.

La première partie présente un état de l'art de la théorie terminologique et des différentes techniques utilisées en terminologie computationnelle. Dans cette partie, nous annonçons le détour « typologique » sur laquelle notre méthodologie s'appuie.
Le *chapitre 1* propose d'aborder les principes fondateurs de la terminologie wüsterienne (Wüster, 1931 ; Wüster & Bauer, 1979) et déroule les principales critiques qui lui ont été adressées, notamment du point de vue sociocognitiviste (Temmerman, 2000). Puis un aperçu de différents types d'unités pouvant former des termes posera la question de l'universalité de la notion de mot. Cela nous conduira ultérieurement aux problématiques relatives au choix d'unités élémentaires de traitement à travers les langues.
Le *chapitre 2* passe en revue des procédés courants utilisés en terminologie computationnelle. Il y est constaté que la majorité des procédés d'extraction terminologique automatiques sont pour une large part monolingues et majoritairement centrés sur le traitement de langues européennes.
Le *chapitre 3* introduit le raisonnement que nous avons souhaité suivre pour l'élargissement de la stratégie d'extraction terminologique automatique à un plus grand nombre de langues : ce dernier, selon les recommandations de Bender (2011), repose notamment sur des informations typologiques. Un rapide aperçu typologique concernant la morphologie des langues et l'ordre des constituants dans la phrase y est proposé.

La suite de ce manuscrit est divisée en deux parties, qui correspondent aux deux tâches centrales de cette thèse : l'extraction terminologique « indépendante de la langue » à partir de corpus quasi-comparables [2] et la complétion de terminologie multilingue *via* un large graphe de traductions fortement multilingue. La figure 0.1 présente un schéma simplifié de l'architecture de la chaîne de traitement appliquée en vue d'obtenir la ressource finale.

2. Fung & Cheung (2004) définissent un corpus *quasi-comparable* comme étant un corpus bilingue non-aligné, non-traduit, pouvant faire référence ou pas aux mêmes thèmes.

FIGURE 0.1 – Schéma global (hors évaluation) du processus d'extraction et de complétion de terminologies multilingues.



Les éléments constitutifs du bloc A de la figure 0.1, qui sont relatifs à l'extraction terminologique, seront décrits dans la partie II, en 5 chapitres comme suit :

Le *chapitre 4* présente les sept langues sur lesquelles nous avons procédé à nos expérimentations (allemand, anglais, arabe, chinois, français, polonais et turc), ainsi que les corpus utilisés pour ces dernières.

Le *chapitre 5* décrit les pré-traitements textuels auxquels nous avons soumis les corpus en fonction des langues sur la base des éléments de typologie décrits dans le chapitre 3. Ces pré-traitements ont deux objectifs distincts. En premier lieu, une tâche de « découpage » appliquée aux textes bruts (tokenisation ou segmentation, selon les langues) consiste à isoler des unités de type mot ayant une taille comparable entre les langues. En second lieu, et afin d'affiner ce premier choix d'unités de traitement, nous proposons d'escamoter une partie de l'in-

formation morphologique [3] des unités lexicales issues de langues à morphologie complexe. Cette opération, assimilée ici à une sous-spécification sémantique, verra son influence testée lors des expérimentations. A l'issue de ce chapitre, trois cadres expérimentaux reposant sur le type d'unités lexicales manipulées sont proposés : le premier fait directement usage des tokens, les deuxièmes et troisièmes reposent sur des unités sous-spécifiées avec une stratégie respectivement modérée ou (iii) franche [4] .

Le *chapitre 6* décrit la méthode que nous avons adoptée pour la mise en place de l'extraction terminologique. Pour rester indépendants de la langue et des domaines terminologiques considérés, et afin de pouvoir utiliser un modèle appris sur une langue donnée sur les données d'une autre langue, nous avons adopté une approche reposant sur l'entraînement de modèles CRF à partir de traits numériques statistiques et fréquentiels. Après une présentation du principe de fonctionnement de l'apprentissage avec des champs markoviens conditionnels (CRF), nous avons listé les caractéristiques utilisées pour l'entraînement des modèles, et les pré-traitements numériques qui y ont été appliqués (normalisation et discrétisation). Ce chapitre présente également le jeu d'étiquettes et les données de référence utilisées pour l'annotation des corpus d'entraînement et de développement.

Le *chapitre 7* détaille le protocole d'évaluation employé pour l'ensemble des expériences. Il y est question de ré-échantillonnage et de ré-équilibrage des corpus pour procéder à une évaluation par validation croisée. Ce chapitre aborde également la problématique de l'évaluation non-binaire automatisée de candidats termes dans un cadre fortement multilingue. Enfin, la stratégie employée pour la sélection des meilleurs modèles lors de l'évaluation y est décrite.

Le *chapitre 8* présente les résultats obtenus pour les différentes expériences. En premier lieu, en ce qui concerne l'application de modèles entraînés sur une même langue, pour les sept langues de notre échantillon de test. Une seconde série d'expériences investigue les possibilités d'utiliser des modèles entraînés dans une langue pour proposer des termes dans une autre langue. Ces résultats sont discutés en fin de chapitre.

La partie III détaille ensuite le traitement schématisé au sein du bloc B de la figure 0.1 concernant la complétion de terminologies multilingues structurées.

Cette dernière s'effectue grâce à un graphe de traduction fortement multilingue, baptisé YaMTG, dont la construction est décrite dans le *chapitre 9*. Dans ce chapitre, les sources des traductions compilées dans YaMTG sont énumérées et les processus d'extraction de liens de traduction sont détaillés si besoin. Une heuristique de nettoyage du graphe est proposée et

3. Cette information est identifiée à l'aide d'un analyseur morphologique non-supervisé pour que l'indépendance de la langue soit préservée, sans que des outils spécifiques à chaque langue ne soient nécessaires.

4. La différence entre ces deux stratégies réside dans la productivité des affixes supprimées. La stratégie franche supprime l'ensemble des affixes identifiées par l'analyseur morphologique, alors que la stratégie modérée n'élimine que la partie la plus productive des affixes. Cette dernière est identifiée à l'aide d'informations typologiques.

évaluée sur un sous-ensemble de traductions.

La méthodologie dévolue à la complétion multilingue est quant à elle exposée dans le *chapitre 10*. Deux algorithmes y sont mis en avant : un premier, très local, utilise des contre-traductions pondérées pour la proposition de candidats termes ; le second, utilisant des frontières de *clusters* déterminées à l'aide d'un algorithme de clustering local reposant sur le recuit simulé, propose un plus large spectre de candidats termes. Ces deux méthodes sont évaluées sur des données différentes.

Enfin, le *chapitre 11* conclut cette thèse en récapitulant les bénéfices apportés par la forte composante multilingue maintenue pour les différentes parties. Un résumé des contributions suscitées par ces travaux, ainsi que des perspectives envisageables pour leur prolongement, est également proposé.

Première partie

État de l'Art

Chapitre 1

La théorie terminologique

Sommaire

1.1 Philosophie de la terminologie

L'intérêt philosophique porté aux « catégories des choses existantes » s'inscrit dans une longue tradition de spéculations métaphysiques remontant à l'Antiquité et poursuivie durant le Moyen Âge à travers un questionnement épistémologique (Alexeeva, 2006). La réflexion terminologique a été conçue ultérieurement pour catégoriser les connaissances techniques reposant sur des savoirs philosophiques, logiques et linguistiques. Les premières allusions à la terminologie en tant que domaine d'étude datent de la seconde moitié du XVIII^{e} siècle en Europe occidentale et en Russie (Rey, 1979 ; Rondeau, 1984). En particulier, les premières pièces du puzzle intellectuel ayant abouti à la terminologie telle que nous la connaissons aujourd'hui ont été assemblées par le linguiste et théoricien Eugène Wüster au début des années 30 (Wüster, 1931 ; Cabré, 1998). Ce sont tout à la fois les pressions pour l'abandon de dialectes régionaux, le colonialisme, puis les progrès techniques et l'industrialisation qui ont donné naissance à la terminologie moderne. La création décuplée de néologismes associée à une circulation de plus en plus rapide de l'information a rendu nécessaire l'émergence de la terminologie en tant qu'outil d'homogénéisation linguistique préalable, selon les termes de Rey (1979), à un « réglage social ».

Les principes mis en avant par Wüster (fondateur de l'École de Vienne) édictent qu'un terme doit être une unité, mono ou multi-mot(s) [1], inclues dans un système terminologique satisfaisant aux recommandations suivantes :

1. L'onomasiologie étant le point de départ idéal de toute analyse terminologique, la notion de concept est au centre de la théorie. Un terme est considéré comme étant uniquement la désignation d'un concept.
2. Tous les concepts du système terminologique doivent être clairement définis.
3. L'ensemble des concepts doivent être associés à des définitions terminologiques idéalement intentionnelles.
4. Les fonctions reliant un concept à un terme sont bijectives : un concept donné est dénoté par un et un seul terme (interdisant de ce fait la *synonymie*) et un terme donné ne dénote qu'un seul concept (proscrivant alors l'*homonymie* et la *polysémie*).
5. Un système terminologique ne s'envisage qu'en synchronie.

Dans ce schéma normatif, la terminologie est vue comme une taxonomie de concepts instanciés de façon univoque par des « objets simplifiés (taxons) » (Alexeeva, 2006) qui suffisent à décrire le monde. Leitchik & Shelov (2006) expliquent qu'il est néanmoins difficile de donner une définition globale du *concept* puisque cette notion peut faire référence à différentes acceptions pouvant dépendre du degré d'abstraction et de clarté, des relations avec la logique et des

1. Nous discuterons de la notion de *mot* dans la section 1.2.1.

liens avec un domaine de connaissances spécifique. La totalité de l'arrière-plan philosophique adopté pour mettre sur pied une terminologie conditionne la portée du *concept*.

Cependant, quelles que soient les nuances doctrinales qui peuvent faire varier cette définition, cette approche de la terminologie est impraticable telle quelle. Wüster lui-même avait admis la rigidité inhérente de ses principes (Candel, 2004). Il ne paraît y avoir qu'un contexte dans lequel cette approche peut porter ses fruits : celui de la normalisation technique. La création de l'Organisation Internationale de Normalisation (ISO) au lendemain de la seconde guerre mondiale, en 1946, a conduit à la formation de Comités Techniques chargés de la normalisation des terminologies, en utilisant les principes de Wüster.
L'École de Vienne a été imitée par d'autres (École de Prague, École soviétique). Toutes ont en commun d'être plus prescriptives que descriptives, et d'être orientées vers la normalisation. À leur sujet, Temmerman (2000, p. 15) déplore qu'elles confondent les principes, c'est-à-dire les objectifs que l'on souhaite atteindre, avec les faits, autrement dit les fondements d'une science.

Campenhoudt (2006) propose de considérer les détails de la théorie wüsterienne comme faisant partie de l'épistémologie de la linguistique plutôt que comme des spécifications pratiques, cette théorie étant irréaliste en ce qui concerne les applications modernes. À ce titre, Sager (1990, chap. 1) refuse même de considérer la terminologie comme une discipline indépendante. Il l'envisage au contraire comme « un nombre de pratiques qui ont évolué autour de la création, la collecte et l'explication des termes, et de leur présentation », refusant de ce fait toute tentative de théorisation. La *lexicographie*, qui était autrefois clairement distincte de la terminologie pour des raisons conceptuelles, a vu un certain nombre de ses procédés utilisés pour l'extraction terminologique. De la même façon, d'autres approches, issues des sciences de l'information, ne postulant aucune connaissance sémantique ou linguistique, ont également été intégrées et adaptées aux pratiques terminologiques (voir section 2.1). Dans cette acception, la terminologie consiste en une grande variété d'applications pratiques.
Cette *tabula rasa* a permis aux terminologues de s'affranchir de l'inertie qui pesait encore sur la discipline. Néanmoins, l'amplitude du séisme que représente le passage d'une théorie rigide à un chaos applicatif pose la question de l'existence d'une voie médiane.

Temmerman (2000) a questionné la possibilité de conserver un cadre théorique à la terminologie, compatible avec ses applications variées. Elle passe en revue les critiques faites par les terminologues des années 90 au sujet des principes traditionnels mentionnés plus haut. Tout d'abord, les concepts peuvent s'avérer inadéquats pour la description des termes (Sager, 1990 ; Zawada & Swanepoel, 1994 ; Weissenhofer, 1995 ; Kageura, 1995). Ils ne sont pas nécessairement strictement délimités (Sager, 1990 ; Meyer, 1993 ; Weissenhofer, 1995 ; Cabré, 1995 ; Kageura, 1995) et leur donner une définition intentionnelle peut ne pas constituer le

meilleur choix (Sager, 1990 ; Meyer, 1993 ; Zawada & Swanepoel, 1994). Par ailleurs, on peut vouloir rendre compte de l'ambiguïté (Sager, 1990 ; Meyer, 1993 ; Weissenhofer, 1995 ; Cabré, 1995) et la contrainte de la synchronie ne permet pas de rendre compte de l'évolution de la formation des termes, pourtant utile (Sager, 1990 ; Cabré, 1995). En parallèle, les années 90 virent l'émergence de la *socioterminologie*, une branche étudiant la terminologie *in vivo*, afin de re-contextualiser l'étude des termes.

Forte de ces observations et de ces influences, Temmerman (2000) a proposé une nouvelle ébauche de cadre théorique, baptisé *Théorie Sociocognitive de la Terminologie*.

> Our proposition for an alternative theory of Terminology starts from the insight that words cannot « mean » objectively, but rather that they can be understood in a linguistic communication process about a reality outside language which has to be understood as well.
>
> Temmerman (2000, p. 42)

Dans cet ouvrage, elle adopte un point de vue résolument sémasiologique (qui préfère partir des termes pour définir ensuite des concepts, par opposition à l'approche onomasiologique) pour réfuter ou modérer chacun des cinq principes de la terminologie traditionnelle comme suit :

1. Au lieu de *concepts*, elle propose de considérer des *unités de compréhension* (*units of understanding*, UC ci-après). Alors que le concept est supposé exister objectivement, ce n'est pas le cas de l'unité de compréhension [2].
2. Dans la Terminologie Traditionnelle, chaque concept est clairement défini et trouve sa place dans une structure (onto)logique (liens *is-a* ou *is-part-of*) ; dans la Terminologie Sociocognitive, une UC correspond à un prototype cognitif, c'est-à-dire à une signification expérimentale, qui est en quelque sorte « le meilleur représentant de sa catégorie ». Les UC peuvent être structurées à la fois de façon intra- et inter-catégorielle, et les frontières de catégories peuvent être floues. Dans ce cas, Temmerman parle de « chunks of knowledge ».
3. Alors que les concepts doivent être définis en intention, « la description du sens peut avoir des unités d'information plus ou moins essentielles » qui vont dépendre de plusieurs facteurs. À la lumière de ces facteurs, la définition pourra être modelée à travers

2. Ces UC se rapprochent des concepts décrits par Rosch (1975) dans sa théorie des prototypes ; les concepts sont alors des appareils qui mettent à disposition un maximum d'informations avec le moins d'efforts scognitifs ; Il s'agit de prototypes décrivant des items du monde redondants tels qu'ils sont perçus par le plus grand nombre plutôt que la connaissance métaphysique d'items équiprobables (Rosch, 1978).

un *patron de compréhension* (*template of understanding* en anglais), qui contient un ensemble de traits à instancier.

4. L'isomorphisme entre les concepts et les termes est artificiel. Il est important de rendre compte des phénomènes de synonymie et de polysémie parce qu'ils constituent une indication sur le sens d'une catégorie et peuvent refléter des différences de perspectives.
5. La synchronie et son corollaire – la relation arbitraire entre les termes et les concepts – ne fait aucun cas de la productivité des langues. Or c'est justement le développement des idées nouvelles, qu'il faut nommer, qui justifie dans une certaine mesure, l'utilité de la terminologie.

En ce qui concerne ce dernier point, Moravcsik (2007) constate que les seules explications causales possibles pour des systèmes linguistiques se font à l'aune de l'histoire, c'est-à-dire en tenant compte de la manière dont un système donné a évolué à partir d'un état antérieur. Étant donné que les langues et les dialectes qui nous intéresseront par la suite peuvent avoir des liens de parenté forts, et qu'en typologie linguistique, les similarités lexicales constituent un indicateur possible de liens de parenté entre langues (Bakker *et al.*, 2009), il serait dommage de refuser d'envisager d'éventuelles considérations diachroniques.

Les bases théoriques jetées par Temmerman semblent mieux adaptées aux nombreuses applications pratiques envisagées en terminologie computationnelle, en particulier dans le cadre de langues non contrôlées. Nous nous intéresserons ici aux questions entourant l'extraction automatique de termes dans différentes langues, contenus dans des textes bruités et relevant de domaines non-techniques dont la terminologie évolue rapidement, et pouvant comporter des nuances culturelles. L'approche sociocognitive parait compatible avec cette conception utilitaire du *terme* et de la *terminologie*, notamment dans le cadre de production de ressources mono-fonctionelles (c'est-à-dire développées pour une tâche spécifique) comme c'est le cas ici. Même si les travaux présentés dans cette thèse feront l'économie de discussions théoriques poussées sur la nature essentielle des termes, comme toujours dépendante du cadre applicatif (L'Homme, 2005), nous regarderons les principes génériques de la terminologie sociocognitive comme une approche théoriquement conciliable avec nos objectifs.

1.2 Objets de la terminologie

Dans la section 1.1, nous avons évoqué plusieurs courants terminologiques : l'approche traditionnelle, onomasiologique, dans laquelle un terme désigne un concept ; l'approche agnostique de Sager, dans laquelle les considérations théoriques sont avant tout soumises à des considérations pratiques ; et le point de vue sociocognitiviste qui envisage le terme comme une unité de compréhension. Les approches agnostique et sociocognitiviste favorisent toutes

deux une démarche sémasiologique.

Pour toutes ces approches, il existe un dénominateur commun : les *termes* sont caractérisés avant tout par leur spécialisation. La différence entre les termes et les autres unités porteuses de sens tient au fait que les termes relèvent d'un domaine spécifique. D'ailleurs, la notion de *termhood* est aujourd'hui largement utilisée, notamment en terminologie computationnelle, pour décrire cette relation que les termes entretiennent avec leur domaine (Kageura & Umino, 1996).

Concernant la nature grammaticale ou le statut phraséologique qu'un terme peut avoir, il existe un choix de possibilités selon la philosophie qui prévaut et les perspectives d'utilisation de la terminologie que l'on souhaite construire (Wright, 1997). La nature d'un terme peut être plus ou moins codifiée en fonction des langues et des domaines considérés pour l'extraction. Nous aborderons cette question dans la section 2. Pour ce qui est de la taille des unités qui peuvent former un terme, il peut s'agir aussi bien d'un mot simple (on parle alors de *terme simple*) que d'une unité phraséologique (il s'agit alors d'un *terme complexe*).

Nous avons déjà utilisé plusieurs fois le terme *mot* sans le définir. Ceci a été fait à dessein, car le locuteur du français a une compréhension intuitive de ce qu'il considère être un mot. Cette intuition, influencée par l'environnement linguistique du locuteur, peut toutefois varier d'une personne à l'autre, et ce même si des connaissances linguistique sur les questions relatives au mot entrent en jeu (Allwood *et al.*, 2010). Cette notion est encore plus ambiguë si on l'envisage depuis des langues très différentes : un Inuit et un Chinois auront probablement de grandes difficultés à comprendre intuitivement quel est *le mot* de l'autre.

Les sections suivantes présentent un aperçu de différents types d'unités pouvant former des termes, en commençant notamment par esquisser une définition de la notion de mot.

1.2.1 À propos du mot

L'analyse linguistique structurale est souvent envisagée comme une pyramide à cinq couches (de la base au sommet : phonologie, morphologie, syntaxe, sémantique, pragmatique). Lorsque l'on souhaite faire un traitement linguistique des couches supérieures, il est souvent préférable d'avoir déjà réalisé l'analyse des strates inférieures. Cette représentation permet d'avoir une image globale des étapes à franchir pour le traitement des langues. Néanmoins, elle reste trop simpliste car elle ne rend pas compte des interdépendances entre les différents niveaux. Typiquement, la définition choisie pour la segmentation en mots en début d'analyse peut être remise en cause par chacune des couches supérieures, et essentiellement par celle qui prévaut pour la finalité de l'analyse.

La plupart des concepts fondamentaux, et pourtant ambigus, discutés en linguistique sont liés aux unités de base autour desquelles se structurent les langues. En fonction de la pers-

pective linguistique ou de la langue considérée, il peut arriver qu'il n'y ait pas de consensus concernant la question des unités les plus basiques. Cependant, tout processus d'extraction terminologique repose sur une définition préliminaire de l'unité visée qui influence la qualité et la reproductibilité entre les langues. Dans le cadre de l'extraction terminologique, discipline historiquement occidentale, on a souvent tendance à considérer que la typographie suffit à distinguer les unités de base. Or comme nous le verrons à la section 3.1, les systèmes d'écritures utilisés par les langues ont un impact considérable sur la segmentation. La plupart des écritures alphabétiques disposent de caractères de segmentation évidents pour permettre au lecteur d'identifier confortablement les éléments du lexique [3]. Les espaces et la ponctuation sont couramment utilisés comme des séparateurs de *mots typographiques* dans les systèmes syllabographiques et phonographiques.

À l'inverse, d'autres langues n'utilisent pas systématiquement des indices explicites pour séparer les mots. Considérons par exemple l'énoncé suivant, en mandarin :

(1.1) 他 特 别 强
tā tè bié qiáng
'Il est particulièrement fort'

Dans cet exemple, la chaîne de caractères « 特别强 » (« *particulièrement fort* ») peut être segmentée en deux mots (« 特别 » et « 强 ») mais sera plus vraisemblablement considérée comme un seul mot par la plupart des locuteurs chinois lettrés, même si Liu *et al.* (2013) indiquent que pour la plupart des locuteurs chinois lettrés, la notion de mot reste vague et inadaptée.

Considérons les exemples d'école suivants : le vietnamien (exemple 1.2, Comrie (1989)) ainsi que le yupik sibérien transcrit en alphabet latin, et dans lequel des tirets ont été insérés pour séparer les morphèmes [4] (exemple 1.3, Comrie (1989)). Ces deux exemples exhibent des types de morphologie différents qui seront décrits à la section 3.2.

(1.2) *Khi tôi đến nhà bạn tôi, chúng tôi bắt đầu làm bài.*
quand moi venir maison ami moi, ⟨PL⟩ moi commencer faire leçon
'Quand je suis arrivé chez mon ami, nous avons commencé la leçon'

(1.3) *angya-ghlla-ng-yug-tuq.*
bateau-⟨AUGMENTATIF⟩-acquérir-⟨DÉSIDÉRATIF⟩-⟨3.SG⟩.
'Il veut acquérir un gros bateau.'

3. Cette tendance n'est pas systématique. Par exemple, les systèmes d'écritures employés par des langues comme le thaï (abugida), le grec primitif ou latin classique (alphabets) sont ou ont été *scriptio continua*. À l'inverse, il peut exister une « sur-segmentation ». C'est la cas en vietnamien, comme cela sera évoqué par la suite.

4. Nous donnons cet exemple uniquement pour illustrer un cas limite (polysynthétisme) régulièrement mis en avant dans la littérature. En pratique, nous ne sommes intéressés dans cette thèse que par le traitement des langues disposant d'un système d'écriture stable et d'utilisation courante chez des locuteurs natifs, ce qui n'est pas le cas de la plupart des langues polysynthétiques.

Bien que ces exemples puissent être tous deux considérés comme des phrases, on observe que des espaces séparent les morphèmes dans le premier cas et non dans le deuxième. Toutefois en vietnamien, les frontières entre morphèmes ne sont pas immuables : les deux morphèmes « *bắt* » et « *đầu* », qui, pris isolément signifient respectivement « attraper » et « tête » prennent un sens différent (« commencer ») s'ils sont contigus comme dans l'exemple 1.2. De la même façon, les tokens « *chúng* » et « *tôi* » fonctionnent ensemble dans la phrase, et l'on peut souhaiter leur regroupement.

Similairement, la question des unités préliminaires à considérer pour une analyse linguistique n'est pas triviale non plus pour les systèmes d'écriture qui utilisent un séparateur tel que l'espace. Avant même de faire entrer en jeu des considérations linguistiques, force est de constater que la typographie comme indicateur de frontière de mots révèle des unités de taille parfois difficilement comparables entre les langues.

Aucune définition du *mot* satisfaisante à tous les égards n'a jamais été proposée. Packard (2000) a recensé plusieurs définitions possibles du *mot* pour le chinois mandarin. Parmi elles, les *mots lexicaux*, les *mots sémantiques*, les *mots morphologiques* et les *mots syntaxiques*. Ces définitions se situent à différents niveaux d'analyse linguistique.

Au niveau syntaxique, les mots sont des entités qui peuvent apparaître librement dans le discours et être remplacées par des entités similaires au même emplacement sur l'axe syntagmatique. Ces entités, lorsqu'elles se situent sur un même axe paradigmatique, possèdent la même partie du discours.

Au niveau morphologique, les mots sont « les sorties correctes produites par les règles morphologiques de la langue » (Packard, 2000). Il est relativement aisé de les détecter dans du texte. Cependant, ils ont pour inconvénient, dans les langues analytiques, d'être des variantes d'un même mot sémantique. Une analyse morphologique sur ces mots, selon leur complexité, permet d'identifier des *morphes*, des *racines* ou même des *lemmes*, ainsi que des catégories flexionnelles.

Les mots lexicaux correspondent aux associations « idiosyncratiques et arbitraires de sons et de sens » ne pouvant pas être déduits de règles de formations (Packard, 2000, p.9). Typiquement, ce sont des unités susceptibles d'apparaître dans des listes lexicales, à l'instar d'entrées de dictionnaires. Les trouver de façon automatique est une tâche complexe car il s'agit d'identifier des unités de type lemme (dans les langues exhibant des phénomènes de flexion et de dérivation), mais aussi des locutions, des termes, voire des entités nommées.

Les mots sémantiques font quand à eux référence à des concepts unitaires, qui en appellent plus aux propriétés ontologiques qu'aux propriétés fonctionnelles d'un mot. On peut considérer qu'il existe une équivalence grossière entre le *mot sémantique* et le *mot lexical*.

S'il est complexe de définir la notion de mot dans une langue donnée comme l'a fait (Pa-

ckard, 2000) pour le chinois, il est encore plus difficile d'avoir une notion de mot valable à travers les langues. À ce titre, beaucoup d'auteurs en sont arrivés à la conclusion que la réalité du mot n'est pas universelle (pour une discussion plus fournie à ce sujet, nous invitons à consulter Magistry (2013)).

La norme ISO MAF (*Morphosyntactic Annotation Framework*, (Clément & Villemonte de La Clergerie, 2005)) a été proposée de manière à pouvoir considérer aussi bien les paramètres typographiques, morphologiques, lexicaux que syntaxiques. Dans cette norme, un *token* est une séquence contiguë non vide de caractères, isolée de ses voisines par des séparateurs. Ces derniers peuvent être soit des espaces, soit des marques typographiques de ponctuation. Les séparateurs peuvent être implicites comme en chinois, où l'on peut considérer chaque caractère comme un token (Magistry, 2013 ; Dong *et al.*, 2010). L'idée de token correspond à la notion de *mot typographique*, sans vernis théorique.

Le MAF autorise le regroupement de tokens (« *token attachment* ») ou la division d'un token (« *granularity* », par exemple pour le mot « auquel » analysé en « à lequel ») dans des mots-formes. Dans ce cas, les lemmes et les parties du discours peuvent être indiqués sur des attributs XML associés au token en question. Les définitions plus abstraites du mot sont diluées dans ce que le MAF appelle *mot-forme*. Les mots-formes sont « des entités, contiguës ou non, issues d'un discours ou d'une séquence textuelle qui sont identifiées comme telles dans le cadre de relations associatives » (Clément & Villemonte de La Clergerie, 2005). Dans le cadre de cette thèse, nous nous appuierons initialement sur la notion de token, comme définie dans le MAF, afin de trouver une notion intermédiaire entre mots lexicaux et mots morphologiques convenant au plus grand nombre de langues.

Clarifier le statut théorique des différents cas de figure pour lesquels plusieurs mots typographiques peuvent être envisagés conjointement va nous permettre de déterminer quels sont les statuts phraséologiques possibles pour les termes complexes : mots composés (section 1.2.2.1), collocations (section 1.2.2.2), idiomes (section 1.2.2.3) et entités nommées (section 1.2.2.4).

1.2.2 Unités polylexicales

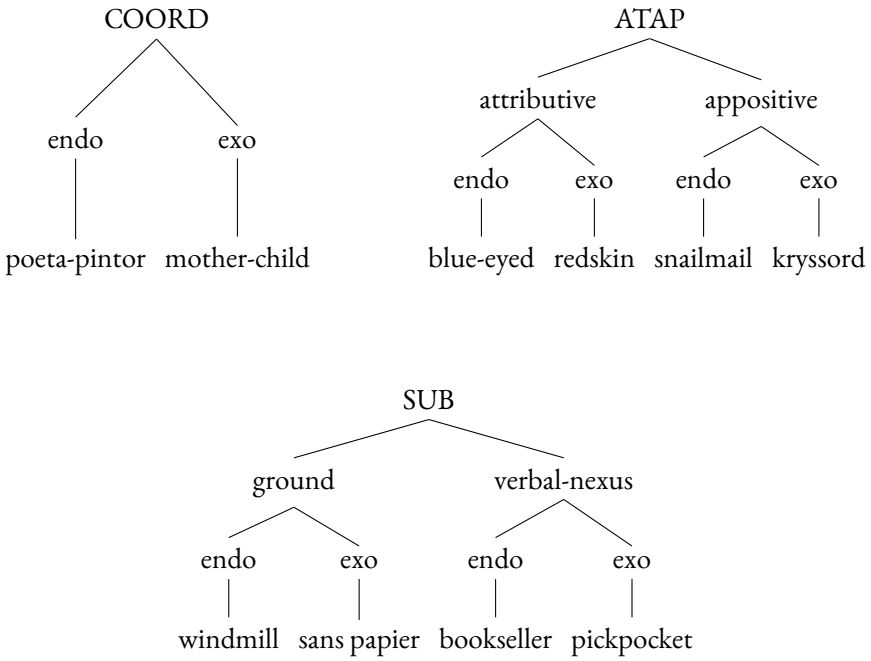
1.2.2.1 Mots composés

La composition est un phénomène répandu, mais pas universel (Štekauer *et al.*, 2012, p. 309). En termes pratiques, la composition est la combinaison (possiblement récursive) de deux mots existants dans le lexique d'une langue. Les mots composés sont donc des mots ou des unités multiples qui peuvent être analysés à la fois dans le lexique et dans la grammaire (Lieber *et al.*, 2009). Ils partagent les propriétés suivantes (Lieber *et al.*, 2009, chap. 8) :

— ils contiennent plus d'un constituant ;
— ils appartiennent à des domaines syntaxiques opaques (certaines opérations syntaxiques ne s'y appliquent pas) ;
— leur sémantique n'est pas nécessairement compositionnelle ;
— phonologiquement, leurs accents toniques peuvent ne pas coïncider avec ceux des mots ou des phrases.

Bisetto & Scalise (2009) proposent une classification hiérarchique des types de mots composés applicable dans un grand nombre de langues (partiellement reproduite en figure 1.1). Ils distinguent trois niveaux de discrimination pour l'interprétation sémantique des mots composés : coordination (COORD), attribution/apposition (ATAP) et subordination (SUB).

FIGURE 1.1 – Classification des types de mots composés proposée par Bisetto & Scalise (2009). Les traductions des termes étrangers sont proposées en notes de bas de page [5].



Ces trois niveaux correspondent aux relations implicites entre deux constituants d'un mot composé. Dans chaque mot composé, il existe généralement un élément qui est plus important que les autres. Cet élément s'appelle la *tête*. La description proposée par Bisetto & Sca-

5. Espagnol : *poeta-pintor* 'poète-peintre' ; Anglais : *blue-eyed* 'aux yeux bleus' ; Anglais : *pickpocket* 'piquerpoche = pickpocket' ; Anglais : *windmill* 'ventmoulin = moulin à vent' ; Anglais : *snailmail* 'escargotcourrier = courrier papier' ; Anglais : *mother-child* 'mère-enfant' ; Anglais :*redskin* 'peaurouge = peau-rouge' ; Norvégien :*kryssord* 'croixmot = mots croisés' ; Anglais : *bookseller* 'livrevendeur = libraire'

lise (2009) permet de localiser la tête dans une langue donnée pour ces différents types de constructions.

Toutefois, cette classification décrit principalement des constructions binaires non-productives. Or il existe des langues pour lesquelles la compositionnalité est un procédé très productif qui va au-delà des trois structures proposées par Bisetto & Scalise (2009). Dans les langues germaniques occidentales, les langues scandinaves ou encore en grec moderne, la composition concaténative est une construction non-binaire ne pouvant pas être analysée comme une application récursive de la composition binaire (van Huyssteen & Verhoeven, 2014). La plupart du temps, le sens de ces mots composés n'est pas idiomatique.

1.2.2.2 Collocations

Le terme de *collocation* a été introduit par Firth dans les années 30. Il désigne un groupe de mots apparaissant ensemble plus souvent qu'attendu. Cette association est en théorie (1) habituelle, (2) lexicalement transparente, (3) arbitraire et (4) exhibe des rapports syntagmatiques bien formés (Williams, 2001).

Il peut être difficile de décider si des mots co-occurrents forment une collocation ou une expression idiomatique, à partir des quatre caractéristiques énoncées ci-dessus. Ces dernières constituent des propriétés non pas catégorielles mais plutôt gradientes, de telle sorte qu'elles ne font pas l'unanimité auprès des linguistes lorsqu'il s'agit de les accepter toutes ensemble (Williams, 2001). Comme le dit Williams, le concept de collocation correspond à un prototype en ce sens qu'il est formalisé en fonction de son contexte d'application. Il existe néanmoins deux familles de conception pour l'analyse des collocations (Williams, 2001 ; Evert, 2005).

Les héritiers de Firth, aussi appelé « néo-firthiens », défendent une approche distributionnelle dans laquelle les collocations sont considérées comme étant des co-occurrences directement observables dans des fenêtres textuelles. Manning & Schütze (1999) et Evert (2005) proposent, dans ce cas, de ne pas parler de collocations mais bien de *co-occurrences*.

L'autre conception favorise une approche intentionnelle. Sous sa forme la plus restrictive, les collocations sont formalisées comme des entrées lexicales. Elles sont considérées comme « des paires semi-compositionnelles de mots, comportant un élément « libre » (la « base ») et un autre élément lexical non déterminé (le « colloqué ») (Evert, 2005).

> Word combinations that are considered as collocations range from compound nouns (*black box*), over semantically opaque idiomatic expressions (*kick the bucket*), to fully compositional combinations that are only lexically restricted (*handsome man* vs. *beautiful woman*). This variability in definition is mirrored by a large number of alternative terms that are used almost interchangeably, such as multi-word expressions (MWE), multi-word units (MWU), bigrams and idioms.
>
> Evert (2005, p.16)

Néanmoins, cette définition de la collocation couvre plusieurs sous-classes que l'on peut trouver le long d'un gradient d'autonomie, allant des combinaisons semi-compositionnelles aux idiomes fixes. Une modélisation plus fine des phénomènes collocatifs a été proposée par Mel'čuk (1998), et reformulée par Polguère (2003). L'analyse qui en découle fait usage du formalisme des fonctions lexicales de la Théorie Sens-Texte (Mel'čuk, 1997) pour codifier la nature et l'organisation des relations lexicales entre base et collocatifs.
Nous allons aborder rapidement les unités multi-mots remarquables que sont les *idiomes* et les *entités nommées*.

1.2.2.3 Idiomes

Les idiomes sont des expressions conventionnelles dont le sens global n'est pas fonction du sens des éléments qui le composent. Ils peut s'agir de lexèmes non compositionnels (comme le mot anglais « *hot dog* »), auquel cas on les appelle « mots composés » (Makkai, 1972). Plus précisément, tous les mots composés sont idiomatiques à divers degrés (Benczes, 2006, p.78). Les idiomes peuvent aussi être des phrases métaphoriques (par exemple *coûter les yeux de la tête* pour indiquer la cherté) ou métonymiques (par exemple *de nouveaux visages* pour signifier « de nouvelles personnes ») (Knowles & Moon, 2006).

1.2.2.4 Entités nommées

Ce terme a été consacré à l'occasion de la sixième *Message Understanding Conference* (MUC, 1995), reconnaissant l'apport qualitatif que présentait l'identification et la classification d'entités nommées pour les tâches d'extraction d'information. Néanmoins, il n'existe pas de définition précise étant donné que la nature des entités nommées dépend de son champ d'utilisation.

> Étant donnés un modèle applicatif et un corpus, on appelle *entité nommée* toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus.
>
> Ehrmann (2008, p. 168)

Ehrmann définit le modèle comme une typologie de données intéressantes au regard d'une application. Cette typologie modélise le monde comme nous le percevons, à la fois partiellement et avec un biais applicatif. Les unités linguistiques qui (1) trouvent une place dans le modèle applicatif et (2) ont un référentiel unique dans la typologie peuvent prétendre au statut d'entités nommées. En ce sens, les entités nommées sont souvent définies par extension, en particulier par le moyen de typologies, plutôt que par intention à travers une définition stable et unique.

Les entités nommées peuvent être aussi génériques que des noms (de personnes, de lieux, d'organisation...) ou des expressions numériques, mais leur niveau de granularité peut varier au sein du modèle applicatif. Par exemple, on peut identifier plus de niveaux de sous-catégorisation pour le type *personne* (par exemple humain, animal, groupe, imaginaire) (Ehrmann, 2008, p. 55).

Que les entités nommées puissent être considérées comme des termes sous certaines conditions n'est pas évident. Leur usage référentiel est différent de celui des termes. Cependant certaines entités nommées peuvent être sémantiquement liées à un domaine spécifique. Cela peut être le cas pour des entités nommées particulières comme les noms de médicaments, de composés chimiques, de minéraux ou de protéines par exemple. Pour répondre à cette interrogation, on peut envisager la différence entre un terme et une entité nommée le long d'un axe dénotant un degré de référence qui dépend du modèle applicatif choisi. De ce fait, il n'existe pas de frontière stricte entre les deux notions.

1.2.3 Termes

Les termes peuvent emprunter à tous les statuts phraséologiques listés ci-dessus. Sur un plan théorique, chaque couche conceptuelle ajoutée aux notions de *mots composés*, *collocation* et *idiomes* est étroitement liée à ses cadres d'usages. Il existe donc un degré de subjectivité qui empêche d'établir un accord clair sur la nature exacte des unités multi-mots.

Xu *et al.* (2009) ont abordé le problème sous un autre angle. Selon eux, la notion de compositionnalité n'a pas grand sens dans le référentiel chinois. Ils proposent simplement de mesurer l'étroitesse du lien entre deux unités (ici, des caractères) sur un *continuum*. Sur ce dernier, la compositionnalité serait en partie fonction de l'attraction que deux unités entretiendraient

entre elles. Ainsi, les unités multi-mots non compositionnelles (c'est-à-dire les mots composés, les idiomes et les translittérations d'un nom propre) peuvent être regroupées.
Cette intuition existait déjà dans le champ de l'extraction terminologique avec la notion de *unithood*.

> "Unithood" refers to the degree of strenght or stability of syntagmatic combinations or collocations. Thus the concept of "unithood" is not only relevant to simple and complex terms, but potentially to other complex units as grammatical collocations or idiomatic expressions.
>
> Kageura & Umino (1996, p. 260)

Cette notion de *unithood* est en général associée à celle de *termhood* (qui désigne quant à elle la spécificité d'un terme relativement à un domaine). Elles constituent la base de nombreuses techniques d'extraction terminologique (Zhang *et al.*, 2008). À elles deux, ces notions résument les qualités essentielles qu'un terme doit satisfaire en théorie. Les concepts philosophiques qui peuvent s'ajouter à ces deux notions, permettent, au besoin, de sélectionner les candidats potentiels pour les intégrer dans un système terminologique conforme à une utilisation particulière.

1.3 Conclusion

Les prémices de la discipline ont longuement bridé la pratique terminologique, la cantonnant au domaine limité du normatif. La remise en question du cadre théorique définissant les objets, leurs propriétés, leurs relations et même leur finalité, aura permis de développer de nombreuses terminologies extraites *ad hoc*, ne se conformant pas aux principes originaux mais se rapprochant pourtant intuitivement de l'idée d'une terminologie.

Afin de satisfaire aux intuitions dictées par ces utilisations pratiques ne pouvant, ni parfois ne voulant, se définir formellement, il a été nécessaire de relaxer la plupart des contraintes philosophiques. La terminologie sociocognitive, formalisée par Temmerman (2000), est la reformulation de la théorie terminologique la plus adaptée aux besoins de faisabilité et d'utilité globale. Ainsi, en consacrant notamment le droit à la sous-spécification, inhérente aux pratiques de terrain, cette approche théorique a ouvert la porte à une réconciliation entre théorie et applications. Une enjambée si considérable qu'il aura tout de même paru nécessaire à Temmerman de renommer le *concept*, en *unité de compréhension*. Cette notion d'unité de compréhension est d'autant plus commode l'on travaille dans un cadre multilingue (voir section 3.4)

S'il existe un consensus qui a résisté aux joutes entre terminologues, c'est au sujet de la spécificité du terme (*termhood*) et de son degré de cohésion interne (*unithood*). Y compris en l'absence de théorie terminologique, ce sont ces deux caractéristiques qui ont permis de développer des applications d'extraction automatique. La notion d'*unithood* est indissociable, notamment dans le cadre multilingue, de la question de l'atomicité des éléments constitutifs d'un terme. Or, le point de départ de tout traitement textuel concerne le choix, non trivial, des unités élémentaires (mots morphologiques, mots typographiques, mots sémantiques, etc.), qui peut ne pas être transposable d'une langue à une autre. C'est une question que nous aborderons de nouveau dans le chapitre 3 et plus concrètement dans le chapitre 5.

Chapitre 2

La pratique terminologique : terminologie*s* computationnelle*s*

Sommaire

La conception de la notion de *terme* et ses perspectives d'inclusion au sein d'une structure de connaissances influent de façon considérable sur les processus d'extraction terminologique. Depuis la remise en question de la théorie traditionnelle, la terminologie peut être construite de façon à accepter un degré d'incertitude plus ou moins grand et être intégrée dans un système sémantique plus ou moins organisé.

Au sens strict, un terme est un élément d'une terminologie. En pratique, il existe un panel diversifié de structures de connaissances pouvant inclure des termes. Ces dernières peuvent aller des listes simples (dictionnaire, glossaire, lexique, etc.), à des systèmes plus ou moins structurés comme les *wordnets* (voir Fellbaum (1998) ; Vossen (1998) ; Stamou *et al.* (2002) ; Bhat *et al.* (2013) entre autres) , les réseaux sémantiques, les ontologies, les taxonomies ou d'autres systèmes de classification. Les modèles et les contraintes de ces différents systèmes de représentation des connaissances se chevauchent parfois, mais rarement totalement. De plus, la description partielle de la sémantique du monde réel a lieu le plus souvent dans un contexte d'utilisation particulier. À ce titre, la plupart de ressources créées sont monofonctionnelles.

2.1 Étapes d'un processus d'extraction de terminologie

Le simple fait de s'affranchir de la doctrine terminologique pour la création de ressources terminologiques a permis de faire évoluer les techniques d'extraction vers une attitude plus pragmatique. À ce titre, l'extraction terminologique en tant qu'application comporte de nombreux points communs avec d'autres domaines traitant de l'information ayant vu le jour de façon concomitante (Sager, 1990). Ces applications issues des sciences de l'information ont différentes finalités : indexation, fouille de textes (*text mining*), moteurs de recherche, classification de textes, etc.

Les traitements d'extraction terminologique à proprement parler consistent souvent en l'enchaînement de quatre grandes étapes : (I) le pré-traitement des corpus, (II) l'application de procédés pour l'identification de candidats termes, (III) le classement et la catégorisation des candidats termes et (IV) la validation des candidats termes.

L'intégralité des traitements et mesures que nous allons présenter dans la section suivante est appliquée à des corpus textuels. Nous reprenons les termes de Foo (2012) pour la description des corpus utilisés dans le processus d'extraction terminologique. Ceux dont on souhaite extraire les termes sont appelés *corpus internes*. Les autres, utilisés pour mettre en exergue des contrastes, sont appelés *corpus externes*. Ces derniers peuvent concerner le même domaine de spécialité que le corpus interne, des domaines apparentés, d'autres domaines de spécialité ou bien être des corpus génériques.

2.1.1 Pré-traitement des corpus

La phase de pré-traitement dépend en grande partie de la qualité et de la nature des corpus utilisés. Si les corpus sont composés de textes normalisés, écrits dans une langue uniforme, alors la tache de pré-traitement peut comporter les pré-traitements linguistiques standards. Parmi eux, il y a la segmentation (en tokens et/ou en phrases), la lemmatisation, la racinisation, ou encore l'étiquetage morpho-syntaxique.
Dans le cas où les corpus sont de qualité moindre (divergences typographiques, erreur d'orthographe ou de syntaxe, registres de langue différent) mais où l'on n'admet que des termes ressortant de la langue normée, il faut rajouter des étapes de pré-traitement de surface comme la normalisation (détection et suppression du bruit), voire de correction orthographique.

2.1.2 Identification de candidats termes

On distingue traditionnellement trois grands types d'approches pour l'identification de candidats termes : les approches linguistiques, les approches statistiques et probabilistes, et les approches hybrides. L'objet de cette section est de présenter les grandes lignes des différentes approches

2.1.2.1 Approches linguistiques

Nous avons évoqué à la section 1.2 la possibilité d'appliquer des contraintes quant à la nature morpho-syntaxique d'un terme. Dans la majorité des publications et des ouvrages traitant d'extraction terminologique, il est communément admis que la plupart des termes sont de nature nominale (Sager, 1990 ; Bourigault, 1992 ; Daille, 1994 ; Justeson & Katz, 1995).
Ainsi, en s'appuyant sur les informations morpho-syntaxiques obtenues en phase de pré-traitement, il est possible de créer un ensemble de patrons linguistiques identifiant des syntagmes nominaux ou des sous parties de tels syntagmes pouvant être des termes. Ces patrons, également appelés règles, sont des expressions régulières sur les étiquettes morpho-syntaxiques. Ils peuvent être aussi bien créés manuellement par un expert qu'induits à partir d'exemples avec des systèmes d'apprentissage automatique. Ce genre de grammaire est régulièrement accompagné de listes d'éléments interdits (*stop-lists*) qui permettent de favoriser la précision (filtre strict) ou le rappel (filtre permissif).

Dans certains cas, il est également possible de mettre à profit une analyse morphologique. C'est ce qu'a fait par exemple Ananiadou (1994). Partant du constat que la plupart des termes médicaux en anglais relevaient en grande majorité de morphologie latine ou grecque, elle a implémenté un système de grammaire morphologique pour améliorer la reconnaissance de termes médicaux. De la même manière, Heid (1999) a identifié un certain nombre d'affixes

(majoritairement latins) relevant du domaine technique, dans le but de parfaire l'extraction terminologique en allemand. En coréen, Oh *et al.* (2000) ont également mis à profit ce genre d'indices sur une base syllabique pour identifier des translittérations [1].
Bien entendu, ces approches dépendent fortement de la langue, du domaine d'extraction, de la qualité des corpus (qui doit être irréprochable) et de l'outil d'étiquetage morpho-syntaxique utilisé.

2.1.2.2 Approches statistiques et probabilistes

Les approches statistiques reposent généralement sur la fréquence des tokens ou de groupes de tokens appelés n-grammes (n étant le nombre de tokens de la séquence). Par souci de simplicité, la discussion suivante n'illustrera les différentes approches qu'avec les fréquences de tokens.
Cette fréquence d'apparition peut être calculée de différentes manières. La plus évidente consiste à récupérer le nombre d'occurrences de tous les tokens. Nous noterons, $occ(t)$ le nombre d'occurrences d'un token t [2]. Cette mesure est parfois appelée *fréquence* par abus de langage.
Il est généralement admis que le spectre du nombre d'occurrences des mots dans les langues humaines suit une distribution de Zipf-Mandelbrot (Mandelbrot, 1953), dans laquelle la fréquence d'un mot dans un corpus décroît inversement proportionnellement au rang de ce mot. Plus formellement, la loi de Zipf-Mandelbrot peut être traduite par l'équation suivante (Bentz *et al.*, 2014) :

$$occ(r_i) = \frac{C}{(\beta + r_i)^\alpha}, C > 0, \alpha > 1, \beta > -1, 1 \leq i \leq n$$

où $occ(r_i)$ est le nombre d'occurrences d'un mot r_i de rang i, n est le nombre de rangs, C est un coefficient de normalisation, et α et β sont des paramètres. Les paramètres C, β et α varient en fonction de la langue ou des textes envisagés (Popescu & Altmann, 2008 ; Serrano *et al.*, 2009 ; Bentz *et al.*, 2014). Cette distribution, typique des lois de puissance, implique que la plupart des valeurs se trouvent aux extrêmes de la distribution. Afin de réduire l'effet du petit nombre de valeurs extrêmement fréquentes de la distribution, il est possible d'effectuer une transformation logarithmique sur le nombre d'occurrences brut. Ainsi, le nombre d'occurrences logarithmique est défini comme suit :

1. Partant du constat que beaucoup de termes techniques utilisés mondialement se sont vu assigner des translitérations plutôt que des traductions en coréen, et que ces dernières étaient parfois trop variables pour être répertoriées dans des dictionnaires de domaines bilingues, Oh *et al.* (2000) exploitent des différences syllabiques systématiques entre les mots coréens et les translittérations de termes occidentaux afin de les reconnaître.
2. Le signe générique « $*$ » en argument d'une fonction désigne n'importe quel token.

$$OccLog(t) = \begin{cases} \log(occ(t)+1) & \text{si } occ(t) > 0 \\ 0 & \text{sinon.} \end{cases}$$

Le nombre d'occurrences peut également être normalisé par rapport au nombre maximal d'occurrences du corpus $MaxOcc$. On parle alors de *fréquence normalisée* :

$$FreqNorm(t) = \frac{occ(t)}{MaxOcc}$$

D'autres heuristiques de normalisation plus élaborées sont possibles (Manning *et al.*, 2008).

Au-delà des comptes de fréquence, qui constituent la mesure statistique la plus élémentaire, il existe un nombre considérable d'autres métriques. Elles peuvent être appliquées aussi bien au corpus lui-même qu'à d'éventuels corpus externes pour calculer la spécificité des termes de façon contrastive. Dans la section 1.2.3, nous avons vu que les qualités essentielles qu'un terme devait satisfaire en théorie relevait de deux notions : *termhood* (qui donne un aperçu de la spécificité de terme) et *unithood* (qui indique le degré de cohésion interne, dans le cas des termes complexes).

◇ Mesures liées à l'*unithood* : Les métriques d'*unithood* sont réservées aux candidats termes constitués de plusieurs tokens. Un large spectre de calculs liés à l'*unithood* sont regroupés sous la dénomination de « mesures d'association ». Ces dernières sont destinées à formuler des modèles statistiques pouvant prédire si les co-occurrences observées sont le fruit du hasard ou, au contraire, si elles correspondent à des associations de mots plus fréquentes qu'attendu (Evert, 2005). Evert (2005) présente les mesures d'association les plus usitées sur des paires de tokens et les catégorise en 4 groupes d'approches :

— celles mesurant l'importance de l'association sur la base de tests statistiques (mesures de vraisemblance et tests d'hypothèse),
— celles reposant sur le calcul de coefficient pouvant dénoter une association positive ou négative,
— celles se basant sur la théorie de l'information et les concepts probabilistes d'entropie, d'entropie mutuelle et d'information mutuelle,
— et des approches dites heuristiques, qui combinent ou modifient des mesures.

La prolifération des mesures possibles atteste du fait qu'aucune ne rend vraiment compte d'une réalité linguistique sous-jacente qui soit universelle. Wermter (2009, p. 62-65) a recensé de nombreux travaux ayant proposé des évaluations comparatives de différentes mesures d'association. Il déplore le fait que les conditions d'évaluation des mesures d'association dans ces recherches soient tributaires d'un trop grand nombre de paramètres (taille des corpus, types

de collocations à extraire, sélection des données de référence etc.), rendant *in fine* l'évaluation « subjective et superficielle ». En d'autres termes, même si certaines mesures ont été identifiées comme plus performantes dans un contexte donné, il serait dangereux de reprendre à son compte ces résultats pour en faire une généralisation.

À défaut de pouvoir déterminer quelles sont les mesures d'association les plus performantes dans tous les contextes, il convient d'identifier quels sont les cas de figure pour lesquels aucune mesure d'association ne peut donner de résultats fiables. La distribution zipfienne des éléments de la langue implique qu'il existe un grand nombre de mots rares. Or, il suffit que des données se trouvent en dessous d'un certain seuil de fréquence ($f \leq 2$) pour que certaines mesures ne soient plus fiables (Evert, 2005). Ces *hapax* ou *dis legomena*, situés en queue de distribution, sont donc à exclure des données pour de tels calculs.

Par ailleurs, les mesures d'association sont conçues pour fonctionner sur des bigrammes. De nombreux termes ne se limitent pas à deux tokens. Petrović *et al.* (2009) ont recensé et proposé des motifs d'extension capable d'élargir le calcul de mesures d'association à des n-grammes de longueur arbitraire en généralisant des approches proposées auparavant par Tadić & Šojat (2003), da Silva & Lopes (1999) et McInnes (2004). Sur le principe, il s'agit de segmenter le n-gramme en plusieurs « bigrammes virtuels » et de proposer une éventuelle normalisation sur ces calculs de mesures d'association résultant des découpes choisies. Leur article propose cinq motifs génériques applicables étant donnés une mesure d'association g et un n-gramme $w_1, w_2 \cdots w_n$:

$$G_1(g, w_1 \cdots w_n) = \frac{g(w_1, w_2 \cdots w_n) + g(w_1 \cdots w_{n-1}, w_n)}{2}$$

$$G_2(g, w_1 \cdots w_n) = \frac{g(w_1 \cdots w_{[n/2]}, w_{[n/2]} \cdots w_n) + g(w_1 \cdots w_{[n/2+1]}, w_{[n/2+1]} \cdots w_n)}{2}$$

$$G_3(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_i, w_{i+1})$$

$$G_4(g, w_1 \cdots w_n) = g(w_1 \cdots w_{n-1}, w_2 \cdots w_n)$$

$$G_5(g, w_1 \cdots w_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} g(w_1 \cdots w_i, w_{i+1} \cdots w_n)$$

Le motif G_1 (Tadić & Šojat, 2003) calcule la moyenne de la mesure d'association g entre d'une part le token initial et le (n-1)-gramme, et d'autre part le (n-1)-gramme initial et le token final. Le motif G_2 (McInnes, 2004) découpe également le n-gramme en bigramme mais applique le

découpage au milieu du n-gramme. G_3 (Petrović *et al.*, 2009) va plus loin en moyennant les résultats de la mesure d'association g appliquée à tous les bigrammes consécutifs du n-gramme. Néanmoins pour les n-grammes longs, cette approche désavantage le score d'association final donné au n-gramme si ce dernier n'est pas lui même composé de bigrammes très collocatifs. G_4 (Petrović *et al.*, 2009) applique la mesure d'association g aux (n-1)-grammes initiaux et finaux du n-gramme considéré. Le motif G_5 (da Silva & Lopes, 1999) divise le n-gramme en deux parties, contenant tous les x-grammes ($x < n$) adjacents possibles pour y appliquer g, et renvoie la moyenne de tous ces résultats.
Pour le traitement des trigrammes, alors que le motif G_4 calcule :

$$G_4 = g(w_1w_2, w_2w_3)$$

les motifs d'extension G_1, G_2 et G_5 sont équivalents :

$$G_{1,2,5} = \frac{g(w_1, w_2w_3) + g(w_1w_2, w_3)}{2}$$

Ce traitement a tendance à sous-estimer le score des collocations comprenant au moins un mot grammatical (souvent assimilé à un *stopword*). Ceux-ci sont en effet très fréquents dans les corpus, au moins dans certaines langues, et réduisent les scores des collocations qui les contiennent. Petrović *et al.* (2009) proposent donc des heuristiques pour supprimer les *stopwords* du calcul des scores d'association.

◇ Mesures liées au *termhood* : L'appartenance d'un terme à un domaine spécifique est souvent déterminée à l'aide de mesures contrastives.
Ces dernières peuvent comparer les fréquences relatives de candidats termes entre un corpus interne et un ou des corpus externe(s). Ahmad *et al.* (1999) ont proposé la mesure de *weirdness*, qui calcule le ratio entre la fréquence normalisée d'un token t dans un corpus externe générique G (possédant n_G tokens) à la fréquence normalisée du même token dans un corpus interne spécifique S (contenant n_S tokens) :

$$Weirdness = \frac{occ_S(t)}{n_S} \times \frac{n_G}{occ_G(t)}$$

À quelques paramètres méthodologiques près, cette mesure se rapproche de celles proposées par Park *et al.* (2002) et Velardi & Sclano (2007), relatives à la pertinence et à la spécificité d'un terme vis-à-vis d'un domaine.
Une mesure très utilisée en sciences de l'information, appelée TF∗IDF (*Term Frequency* ∗ *Inverse Document Frequency*), peut également être exploitée en extraction terminologique. Au lieu de considérer un corpus interne et un ou des corpus externes, les textes sont segmentés

en *documents*. Ce n'est pas sur l'ensemble des termes des documents que cette approche se focalise, mais sur les termes qui décrivent le mieux un ou des document(s) donné(s). TF est le nombre d'occurrences (ou toute autre variante possible pour mesurer la fréquence) d'un terme t donné dans un document. La fréquence inverse de document IDF (Jones, 1972), mesure la dispersion du terme t (apparaissant dans n_t documents) sur l'ensemble des documents. N est le nombre total de documents examinés. L'IDF, sous sa forme la plus simple, est calculé *via* la formule :

$$IDF(t) = \log \frac{N}{n_t}$$

Les termes qui apparaissent souvent dans un petit sous-ensemble de documents mais pas dans les autres (TF fort, IDF faible) possèdent un score TF∗IDF plus élevé. Il existe de nombreuses variations *ad hoc* du TF∗IDF. Bien que pertinent pour l'indexation, ce score peut perdre de son intérêt si les termes que l'on souhaite extraire ne sont pas très fréquents. Il existe de nombreuses variations sur le calcul de l'IDF. Un exemple notable en extraction terminologique est la fréquence inverse de mot (*inverse word frequency* ou IWF) proposé par Basili *et al.* (1999) qui considère le nombre de mots plutôt que le nombre de documents :

$$IWF(t) = \log \frac{\sum_{s+g} occ_{s+g}(*)}{\sum_{s} occ_s(t)}$$

où $\sum_{s+g} occ_{s+g}(*)$ est le nombre de mots de l'ensemble des documents et $\sum_{s} occ_s(t)$ désigne le nombre d'occurrences du terme t dans l'ensemble des documents de spécialité.

Termhood désigne la spécificité d'un terme, mais peut parfois aussi subsumer la notion d'*unithood* en ce sens que le *termhood* peut tout simplement indiquer qu'une séquence de mots est un terme. Ainsi, il est possible d'envisager des approches distributionnelles proches de celles présentées dans le cadre des mesures d'*unithood* comme mesurant également la spécificité d'un terme (Zhang *et al.*, 2008). Justeson & Katz (1995) ou Caraballo & Charniak (1999) ont observé que les items nominaux complexes spécifiques à un domaine étaient plus rarement modifiés que les autres. Partant de ce constat, plusieurs mesures de « modifiabilité » ont été suggérées. Une mesure proposée (Wermter & Hahn, 2005 ; Wermter, 2009), appelée *LPM* (pour *Limited Paradigmatic Modifiability*), consiste à envisager la probabilité avec laquelle un (ou plusieurs) emplacement(s) d'un n-gramme ne peut (peuvent) pas être rempli(s) par un (ou plusieurs) autre(s) token(s) que celui (ou ceux) effectivement présent(s) dans le n-gramme de base. Plus simplement dit, pour un n-gramme donné, plus les axes paradigmatiques de chacun de ses élément lexicaux offrent des choix limités, plus il y a de chance que ce n-gramme soit un terme ; c'est ce que traduit le score *LMP*.

Ces mesures, à l'instar de la plupart des mesures d'association (Daille, 1994), semblent bien adaptées pour la détection de termes complexes peu fréquents.

◇ Approches liées à l'apprentissage automatique : Étant donné le caractère monofonctionnel des outils d'extraction terminologique, de plus en plus de recherches s'appliquent à développer des techniques d'extraction qui soient beaucoup plus paramétrables. En ce sens, l'apprentissage automatique, qui est une branche de l'intelligence artificielle étudiant les moyens d'apprendre à reconnaître des motifs à partir de données, offre de grandes perspectives pour l'extraction terminologique automatique. Il existe beaucoup de configurations possibles pour l'apprentissage, selon que l'on souhaite obtenir une classification binaire ou multi-classes, une estimation structurée, un rapport de régression, etc. Selon les buts à atteindre et les données disponibles, les techniques utilisées sont différentes. Tout d'abord, l'apprentissage automatique peut être supervisé ou non-supervisé :

— Dans le cadre de l'apprentissage supervisé, l'algorithme apprend, à partir de données étiquetées et validées manuellement (également appelées données *Gold*), une fonction à même de générer une sortie sur des données nouvelles. L'ensemble des données Gold doit être représentatif de tous les cas de figure possibles que l'on souhaite gérer. Ces données sont souvent décrites au moyen de vecteurs de traits capables de décrire suffisamment les objets qu'ils représentent. Ensuite, à l'aide d'un algorithme, un ensemble de fonctions est entraîné sur le modèle Gold. Dès qu'une représentation optimale est obtenue, elle peut être testée sur des données dites de développement, c'est-à-dire des données correctement étiquetées qui n'ont jamais été utilisées lors de la phase d'apprentissage. L'utilisateur peut faire varier des paramètres de l'algorithme afin d'améliorer les résultats de la prédiction sur ces données de validation. Enfin, l'efficacité globale de l'algorithme d'apprentissage automatique ainsi entraîné et paramétré peut être évalué sur des données de test, distinctes des données d'entraînement et de développement.

— Dans le cadre de l'apprentissage non supervisé, un algorithme essaye de découvrir une structure dans des données non étiquetées. Les méthodes d'apprentissage non supervisé peuvent être classées en deux familles : les techniques d'estimation de densité, qui construisent des modèles statistiques, et les techniques d'extraction de traits, capables d'exploiter des (ir)régularités statistiques dans les données (Dayan, 1999). Ces méthodes utilisent par exemple des algorithmes de *clustering* ou des modèles de Markov cachés.

L'apprentissage automatique peut être appliqué *a posteriori* sur les résultats de différentes métriques d'extraction (linguistiques, statistiques ou contextuelles). C'est ce qu'ont fait Vivaldi *et al.* (2001) sur l'espagnol, en utilisant la méthode AdaBoost. Cette dernière permet de déterminer des règles de classification performantes en combinant plusieurs méthodes de classification de qualité moyenne, qui sont les métriques d'extraction terminologique précédemment calculées, traduites en plusieurs arbres de décision binaires de profondeur 1. Cette combinaison obtient systématiquement de meilleurs résultats que les mesures d'extraction de termes les plus performantes utilisées en première intention.

La méthode proposée par Foo & Merkel (2010) consiste à employer un algorithme d'induction de règles capable de produire des règles facilement déchiffrables par les utilisateurs et exportables. Plusieurs expériences ont ainsi été menées pour l'extraction de termes de longueurs différentes. L'ensemble des traits adoptés contenait aussi bien des traits statistiques que linguistiques. Enfin, les données utilisées pour l'entraînement ont été rééquilibrées de différentes manières concernant les ratios d'exemples positifs et négatifs lors de la phase de test.

Loukachevitch (2012) a également appliqué une technique d'apprentissage automatique, la régression logistique, pour la reconnaissance automatique de termes dans des corpus en russe. Sa contribution est surtout axée sur la sélection des traits sur lesquels le modèle s'appuie en fonction des domaines d'utilisation.

Lee *et al.* (2012) ont eu recours à une machine à vecteur de support (SVM) (Vapnik, 1995) de façon itérative sur le chinois. L'originalité de leurs recherches par rapport à celles précédemment menées tient à la modicité des connaissances utilisées pour l'entraînement du modèle. Leur méthode s'appuie uniquement sur du texte. Initialement, un seul terme est donné pour l'entraînement. À chaque étape de l'itération, le modèle propose de nouveaux termes qui sont eux aussi employés pour trouver d'autres termes à l'étape suivante.

Conrado *et al.* (2013) ont fait usage de différents paradigmes d'apprentissage basés sur induction de règles (JRip), les arbres de décision (J18) et l'apprentissage probabiliste (classification naïve bayésienne), tous décrits par Witten & Frank (2005), pour la détection de termes simples dans des corpus en portugais brésilien. Les traits conservés pour l'entraînement de leurs modèles vont des mesures statistiques et linguistiques simples, comme la fréquence et les parties du discours, à des connaissances plus sophistiquées relatives par exemple à l'analyse des contextes dans lesquels apparaît le terme.

D'autres publications font état de l'utilisation de modèles graphiques comme Cen *et al.* (2008), avec des modèles de Markov cachés, ou encore Zhang *et al.* (2010), Daille & Blancafort (2013) et Li *et al.* (2012), avec des champs aléatoires conditionnels (CRF).

Les algorithmes non-supervisés sont peu utilisés dans le domaine de l'extraction terminologique automatique, car moins efficaces. À notre connaissance, les propositions existantes

utilisent majoritairement des métriques dérivées du TF-IDF (Kim *et al.*, 2009), éventuellement associées à du clustering sur les tokens ou à une approche de renforcement itératif (Liu *et al.*, 2009 ; Wan *et al.*, 2007).

2.1.2.3 Approches hybrides

Les approches hybrides combinent les approches linguistiques et statistiques. Il peut s'agir de sélectionner des candidats avec un traitement statistique puis de leur appliquer un filtre morpho-syntaxique (Smadja (1993) entre autres), ou bien d'appliquer des mesures statistiques sur des candidats extraits avec une approche linguistique (Justeson & Katz (1995) ; Daille (1994) entre autres).
L'approche hybride la plus populaire en extraction terminologique est la *C/NC Value* proposée par Frantzi *et al.* (1998). Cette heuristique combine deux mesures qui s'appliquent à un ensemble de termes candidats sélectionnés au moyen d'une approche purement linguistique :

— La *C-Value* permet d'extraire des termes complexes imbriqués. Pour un candidat terme a contenant $\mid a \mid$ mots et apparaissant dans un ensemble T_a de candidats termes plus grands, la formule de la *C-Value* est :

$$\text{C-Value}(a) = \begin{cases} \log_2 \mid a \mid \cdot occ(a) & \text{si } a \text{ est imbriqué} \\ \log_2 \mid a \mid \cdot (occ(a) - \frac{1}{|T_a|} \sum\limits_{b \in T_a} occ(b)) & \text{sinon.} \end{cases}$$

Ainsi cette mesure prend en compte non seulement la taille et la fréquence d'un candidat terme, mais également le fait qu'un candidat terme apparaisse à l'intérieur d'autres candidats termes. Cette première étape de calcul sur l'ensemble des candidats termes permet de les classer par *c-value* décroissante.

— La *NC-Value* a pour but d'améliorer l'extraction de termes complexes de façon plus générale en intégrant plus d'information sur les contextes. Chacun des candidats termes apparait dans un contexte textuel, composé de mots « intéressants » (par exemple eu égard à leur catégorie morpho-syntaxique). À chacun de ces mots m apparaissant dans le voisinage de x termes parmi X est assigné un poids $W(m) = \frac{x}{X}$. Une fois l'ensemble des contextes pondérés, la *NC-Value* est calculée selon la formule :

$$\text{NC-Value(a)} = \alpha \cdot \text{C-Value}(a) + \beta \sum_{b \in C_a} occ_a(b) \cdot W(b)$$

avec :

— α et β des poids déterminés empiriquement (auxquels Frantzi *et al.* (1998) attribuent respectivement $0,8$ et $0,2$),
— C_a est l'ensemble des mots b qui sont des contextes de a,
— $occ_a(b)$ est le nombre de fois où b se trouve dans les contextes de a,
— $W(b)$ est le poids assigné à b lors de l'étape de pondération des contextes.

D'autre mesures associent plus directement les résultats linguistiques dans des mesures statistiques ou probabilistes. Partant de l'hypothèse que l'information mutuelle des syntagmes diffère de manière significative selon qu'ils sont compositionnels ou non, Lin (1999) a proposé une méthode pour identifier les syntagmes non-compositionnels en intégrant des informations sur la structure en dépendance à une mesure d'information mutuelle.
La *LSM* (*Limited Syntagmatic Modifiability*), proposée par Wermter & Hahn (2004), est une autre méthode hybride. Il s'agit d'une mesure d'association linguistiquement motivée ciblant une structure syntaxique dite « collocationnelle ». Cette mesure sélectionne l'attachement syntaxique particulier à une ou plusieurs tête(s) (choisie(s) dans la collocation envisagée) dont la probabilité est maximale. Wermter & Hahn (2004) l'ont spécifiquement conçue et testée sur l'extraction des collocations allemandes, proches des constructions à verbe support (en l'occurrence, des triplets préposition-nom-verbe).

Ces méthodes restent très dépendantes des outils d'analyses syntaxiques utilisés. Par ailleurs, leur spécificité tend à ne sélectionner qu'un sous-ensemble très exclusif de termes ou de collocations.

2.1.3 Classement et catégorisation des candidats

Cette étape relève une fois de plus de différentes heuristiques. La plus simple est un classement par ordre alphabétique, qui prévaut généralement lorsque aucune autre règle n'est définie.
Dagan & Church (1994) ont regroupé les termes supposés avoir une tête identique, c'est-à-dire sur leur listes de candidats termes en anglais, le dernier mot de chaque terme. Ces têtes sont elles-même classées en fonction de leur fréquence d'apparition dans les documents. Pour les termes ayant encore un ordre indéterminé, leur classement se fait par ordre alphabétique en envisageant les mots avec un sens de lecture inversé (de droite à gauche).
Sur le même principe, Bourigault & Jacquemin (1999) ont représenté l'ensemble des termes candidats sous forme de réseau. Les termes complexes sont décomposés en constituants de tête et d'extension grâce à des règles syntaxiques. Les termes ainsi regroupés sont également normalisés, c'est à dire que les différentes variantes d'un terme sont regroupées sous une forme dite canonique. Cette étape de classification peut consister à raciniser, lemmatiser, enlever les

mots grammaticaux, ou encore à établir des règles syntaxiques de simplification capables de détecter les modifieurs.

Il est également possible de rapprocher des relations de concepts entre les candidats termes en utilisant des informations sur des liens de synonymie, hyperonymie, antonymie entre items lexicaux ou *via* la morphologie (Daille, 2003 ; Claveau & L'Homme, 2005) par exemple. Des informations distributionnelles sur les contextes syntaxiques peuvent également être utilisées, comme l'ont noté Bourigault *et al.* (2004).

2.1.4 Validation des candidats termes

La validation des candidats termes est avant tout un travail manuel fait par des experts du domaine ou des terminologues. Ce processus est coûteux en temps. Pour cette raison, il arrive que la phase de filtrage tienne lieu de validation automatisée. Nous aborderons plus en détail dans le chapitre 7 (et plus particulièrement à la section 7.2) la problématique de l'évaluation automatique de candidats termes pouvant faciliter cette étape de validation.

2.2 Extraction terminologique et diversité linguistique

La majorité des techniques dont il a été question dans les études précédemment citées ont été développées et testées sur une ou deux langues à la fois, le plus souvent dans des langues européennes comme l'anglais, le français, l'allemand, parfois sur le russe, l'espagnol, le portugais ou le suédois. Certaines ont également été testées sur d'autres langues, notamment l'arabe (Boulaknadel *et al.*, 2008), le chinois, le coréen, le japonais (Nakagawa & Mori, 2002). Pour la plupart des autres langues, la recherche en extraction terminologique automatique est embryonnaire.

2.2.1 Extraction multilingue

Depuis quelques années, de plus en plus de méthodes d'extraction bilingues ou multilingues ont fait l'objet de recherches. Il s'agit d'extraire et d'établir des liens de traduction réciproques entre les termes présents dans des corpus de spécialité dans différentes langues.

En ce qui concerne l'extraction terminologique bilingue, elle s'appuie généralement sur l'utilisation de corpus parallèles ou comparables dans deux langues différentes pour lesquelles une extraction terminologique monolingue est possible. Si cette extraction est effectuée sur des corpus parallèles, les termes extraits pour chaque langue sont mis en correspondance à l'aide d'outils d'alignement, soit au niveau du mot (Fan *et al.*, 2009), soit au niveau des séquences comme en Traduction Automatique (Hjelm, 2007 ; Ideue *et al.*, 2011). Pour pallier la difficulté d'accès à des corpus parallèles de spécialité (notamment pour des couples de langues

très éloignées), certaines recherches ont proposé de tirer parti de corpus comparables. La mise en correspondance des traductions de termes peut être effectuée en calculant un profil de co-occurrence, supposé être similaire dans les langues sources et cible (Rapp, 1995) et en utilisant des dictionnaires multilingues (Fung & McKeown, 1997 ; Daille, 2012 ; Déjean *et al.*, 2002).

D'autre part, l'énorme quantité de données textuelles présentes sur Internet permet de trouver des termes (dans différentes langues) qui co-occurent dans leur contexte de spécialité. Partant de cela, Nazar *et al.* (2008) ont proposé une approche permettant de s'affranchir de corpus parallèles ou comparables en utilisant le web comme moyen de détection de traductions de termes.

Sur certains couples de langues, il est possible d'utiliser des heuristiques plus spécifiques et moins coûteuses. C'est ce qu'ont fait Nagata *et al.* (2001), Cao *et al.* (2007) et Bond *et al.* (2008) à partir de ressources monolingues en chinois ou japonais contenant des indices textuels (encodages des caractères et présence de parenthèses) pour construire des dictionnaires bilingues de la langue du texte source en anglais. Les méthodes d'extraction bilingues généralisables à d'autres couples de langues que celui initialement envisagé peuvent être appliquées dans le but de procéder à une extraction multilingue (Hjelm, 2007).

Daille & Blancafort (2013) ont proposé un premier élément de comparaison entre une technique d'extraction (traditionnelle) état-de-l'art utilisant un étiqueteur morphosyntaxique et une technique qui, à partir d'un petit échantillon de données étiquetées, utilise des méthodes d'apprentissage automatique pour inférer des étiquettes morpho-syntaxiques (avec l'outil développé par Clark (2003)) puis extraire des termes (en utilisant des champs aléatoires conditionnels (Lafferty *et al.*, 2001)). Cette dernière étape est envisagée comme une sous-tâche de l'analyse syntaxique de surface (*shallow parsing*) . Sur six langues (allemand, anglais, espagnol, français, letton, russe), il ressort que la méthode traditionnelle semble meilleure en français, allemand et letton, même si cet écart n'est pas maintenu d'un corpus à l'autre dans une même langue. Toutefois, l'approche utilisant les outils d'apprentissage automatique donne des résultats satisfaisants avec un minimum d'efforts. Cet article est très intéressant car il est, à notre connaissance, le premier à envisager la tache d'extraction terminologique automatique à la fois dans une perspective très multilingue et avec le souci des langues disposant de peu ou pas de ressources linguistiques.

2.2.2 Portabilité des processus entre les langues

Les extractions bilingues ou multilingues sont le plus souvent des processus d'extraction monolingues menés en parallèle, dont on relie les résultats via des liens de traductions déterminés de différentes manières. Un écueil courant dans la littérature relative à l'extraction terminologique automatique consiste à déclarer qu'une technique utilisant des outils linguis-

tiques fortement dépendants d'une langue puisse mener à une extraction indépendante de la langue. Les méthodes linguistiques ou hybrides sont monolingues, ou au mieux, faiblement multilingues, dès lors que les outils linguistiques utilisés en pré-traitement le sont également. Ces dernières doivent au moins redéfinir les étiquettes morpho-syntaxiques et les patrons des règles utilisées d'une langue à l'autre. Un exemple concret concerne la proposition de Valderrábanos *et al.* (2002), qui s'adresse à l'extraction de termes médicaux à visée d'indépendance de la langue. Bien que leur approche réduise les coûts de développement pour chaque langue envisagée (à savoir l'allemand, l'anglais, l'espagnol et le français) par rapport à une approche traditionnelle utilisant un étiqueteur morpho-syntaxique, le fait qu'il s'agisse de langues typologiquement proches d'une part, et du domaine médical [3] d'autre part, interroge sur la portabilité de leur technique à d'autres langues, non européennes, ainsi qu'à d'autres domaines.
Les méthodes qui s'appuient sur des outils statistiques ou probabilistes, parfois combinés dans des approches d'apprentissage automatique, semblent pouvoir offrir une plus grande latitude quant à des applications multilingues (Daille & Blancafort, 2013), mais ce n'est pas toujours le cas. Il arrive que des méthodes valables pour un ensemble de langues ne le soient pas pour d'autres. C'est ce que déplorent Grigonytė *et al.* (2011) et Pinnis *et al.* (2012), notamment en ce qui concerne l'extraction terminologique de langues à morphologie très riche comme le letton, le croate ou le lituanien. De la même manière, Daille & Blancafort (2013) indiquent que pour l'allemand, l'approche statistique possède les même défauts que l'approche traditionnelle état-de-l'art (utilisant un étiqueteur morpho-syntaxique et des règles définies manuellement) utilisée pour la comparaison.

3. Beaucoup de termes médicaux dans les langues européennes sont issus de racines latines et grecques, ce qui en fait encore un cas particulier d'extraction terminologique.

Chapitre 3

Vers l'independance de la langue

Sommaire

Il est généralement considéré qu'il existe plus de 7 000 langues, classifiées en environ 150 familles. Parmi ces dernières, 24 macro-langues [1] comptent à elles seules près de 50 millions de locuteurs natifs [2]. Dans ce vaste horizon, imaginer des traitement indépendants de la langue relève de l'utopie. En revanche, s'intéresser à des démarches qui peuvent maximiser le nombre de langues différentes traitables devient un but ambitieux, mais raisonnable. Par abus de langage, nous entendons donc concrètement par « traitement indépendant de la langue » tout traitement pouvant s'appliquer à une grande variété de langues. Bien que cette approximation s'avère hasardeuse, elle n'en est pas moins acceptable au regard de la logique typologique qui sous-tendra notre approche.

Nous avons vu à la section 2.2.2 que les différences principales entre les méthodes d'extraction terminologique issues des communautés de linguistes et les méthodes de recherche d'information issues des sciences de l'information, sans affiliation théorique, relevaient de l'importance donnée à la nature des termes extraits.

Dans un cas, ces derniers sont très codifiés, aussi bien sur le fond que sur la forme, et les critères de *termhood* et d'*unithood* sont centraux. Ces méthodes sont très peu multilingues et requièrent des efforts considérables pour une adaptation à une langue nouvelle.

Dans l'autre cas, il suffit à ces termes d'être pertinents pour une application donnée, quels que soient leurs signifiés ou leurs éventuels liens avec une thématique. Cette quasi indifférence sur la nature des objets manipulés rend ces méthodes potentiellement utilisables sur n'importe quelles données textuelles disposant « d'unités raisonnables ».

Dans l'idéal, concilier le meilleur des deux mondes devrait permettre à la fois d'extraire des termes pertinents au regard d'un domaine sur un grand nombre de langues, sans coût d'adaptation exorbitant. Or, si nous ne désirons pas nous appuyer sur des informations linguistiques fastidieuses à obtenir, nous ne souhaitons pas non plus traiter des données brutes sans au préalable comprendre quelles pourraient être les « unités raisonnables » et pourquoi. Comme cela a déjà été abordé dans les chapitres précédents, un des principaux problèmes auxquels nous sommes confrontés dans notre recherche pour le développement d'un système d'extraction terminologique multilingue concerne la sélection des unités minimales de traitement. Ces dernières ne sont pas équivalentes dans toutes les langues, que ce soit d'un point de vue typographique (section 3.1) ou morphologique (section 3.2). Pour cette raison, un aperçu typologique des cas de figure possibles s'impose avant de décider comment seront en réalité pré-traitées nos données textuelles.

Qui plus est, la perspective de réutiliser pour d'autres langues des modèles d'extraction terminologique qui ont fait leurs preuves dans une langue donnée nous poussera également

1. Une macro-langue est définie comme étant « un ensemble de langues fortement apparentées, que l'on suppose être dans certains contextes d'usages, une seule langue » (norme ISO 639-3).

2. Ces chiffres sont issus de la base Ethnologue (www.ethnologue.com), consultée fin 2013.

à examiner la typologie relative à l'ordre d'apparition des différents éléments dans la phrase (section 3.3).

Approcher ces questions via un angle typologique, comme le préconise Bender (2011), permet d'intégrer des connaissances linguistiques globales à même de factoriser certains traitements, sachant quelles sont les propriétés partagées des langues, ou au contraire, ce qui les éloigne. Un autre intérêt du passage en revue de certaines connaissances typologiques est que la sélection de l'ensemble de langues pour les tests et les évaluations en est facilité.

Qu'est-ce qu'une typologie ?

La typologie est un outil conceptuel universel qui trouve sa place dans n'importe quel domaine de la linguistique, sans rapport avec aucune affirmation théorique [3] ou aucun principe d'explication ; il s'agit uniquement d'une description (Moravcsik, 2007).
Ses objets principaux sont la catégorisation et la classification de notions translinguistiquement comparables, en association avec d'autres paramètres telles que les familles de langues, leur localisation géographique, etc. Une typologie qui s'intéresse à un ensemble de phénomènes apparentés envisage généralement la diversité et l'uniformité de ce qui est observé dans différents référentiels linguistiques (une langue, un ensemble de langues, ou « toutes les langues »).

Plusieurs démarches sont possibles dans le cadre d'investigations typologiques (Nichols, 2007) : Il peut s'agir uniquement de collecter des hypothèses, auquel cas la typologie est utilisée dans un cadre théorique neutre et menée sur un petit échantillon de langues ; il peut également s'agir de vérifier des hypothèses, auquel cas la typologie prend en compte une ou des théories linguistiques sur de plus gros échantillons de langues pour trouver des corrélations statistiques significatives entre ses différents paramètres.

Dans les deux cas, on statue sur des généralisations dont les modalités peuvent être absolues (par exemple « *Pour tous les mots anglais, s'ils commencent par trois consonnes, alors la première est un /s/* ») ou, plus souvent, statistiques ou probabilistes (par exemple « *Dans la plupart des langues, s'il existe une distinction de genre pour les pronoms pluriels, alors il existe aussi une distinction de genre pour les pronoms singuliers.* ») (Moravcsik, 2007).

À la différence des théories linguistiques, qui poursuivent des buts d'abstractions souvent difficiles à atteindre pour distinguer l'universel du possible, si des universaux sont identifiés dans le cadre d'une typologie linguistique, il s'agit d'un effet de bord des observations. Autrement dit, bien qu'il soit souvent question en typologie d'« universaux », il s'agit en réalité

3. Il arrive cependant régulièrement que les typologues établissent leurs généralisations en s'appuyant sur des théories linguistiques, obtenant alors des données artificielles. Polinsky & Kluender (2007) soulignent néanmoins que des données ainsi obtenues sont utiles au même titre que des données apparaissant naturellement.

de grandes tendances, d'universaux non encore contredits, ou rarement remis en cause. Toute description typologique est donc sujette à caution, quelle que soit la divergence entre la catégorie typologique d'une langue et les exceptions qui en ressortent. Comme nous le verrons dans les sections suivantes, il arrive souvent qu'une langue puisse avoir plusieurs valeurs pour un trait typologique donné : soit que cette variable change en fonction de ses contextes d'utilisation, soit que les données observées souffrent en réalité d'un artefact, soit encore que la question typologique ne soit pas pertinente.

Une fois énoncé le prémisse qu'une typologie ne fait généralement pas référence à des universaux mais plutôt à des grandes tendances à travers les langues, notre point de vue sur *l'indépendance de la langue* en est clarifié. Partant de ça, notre recherche sera orientée vers un traitement typologique de l'extraction terminologique.

Traits typologiques pertinents

> Typically, when we think of linguistic knowledge-based NLP systems, what comes to mind are complicated, intricate sets of language-specifc rules. While I would be the last to deny that such systems can be both linguistically interesting and the best approach to certain tasks (cf. Uszkoreit (2002)), my purpose here is to point out that there are other kinds of linguistic knowledge that can be fruitfully incorporated into NLP systems. In particular, the results of linguistic typology represent a rich source of knowledge that, by virtue of being already produced by typologists, can be relatively inexpensively incorporated into NLP systems.
>
> Bender (2011, p. 6)

La barrière de la langue ne peut pas être franchie en utilisant des correspondances isomorphes entre des termes. Les langues sont tout à la fois *influencées par* et *vecteur de* culture et d'identité, comme s'attellent à le démontrer plusieurs études relevant du relativisme linguistique (Whorf, 1940 ; Gumperz & Levinson, 1996 ; Gilbert *et al.*, 2006).
Les motivations pour s'intéresser à une extraction terminologique indépendante de la langue sont nombreuses. Les techniques d'extraction terminologique courantes sont difficiles à appliquer à de gros corpus textuels bruités comme les verbatim récoltés à l'issue d'enquêtes internationales auprès d'employés. En cause, la qualité et la diversité des données obtenues, et ce même lorsque ces données sont recueillies dans une langue disposant de nombreux outils de traitement automatique. Par exemple, un analyseur syntaxique état de l'art peut ob-

tenir de mauvais résultats sur des textes tout-venant (l'ambiguïté morphologique, un ordre des constituants flexible, la créativité, les fautes d'orthographe et de grammaire étant les principales raisons) (Tsarfaty *et al.*, 2010). De tels outils peuvent également dépendre d'autres tâches telles que la segmentation (par exemple pour le chinois) ou l'analyse morphologique (par exemple pour l'arabe).

Nous sommes donc intéressés par les traits élémentaires qui permettent de normaliser le traitement de textes en différentes langues tout en faisant un usage minimal d'outils de traitement automatique. Notre principal objectif consiste à traiter de larges corpus de textes relevant d'un domaine particulier, et écrits dans des langues disposant de systèmes d'écritures établis et pouvant être numériquement encodés. Il est donc important pour la suite de sélectionner un ensemble de traits typologiques pouvant faciliter une extraction terminologique qui se veut indépendante de la langue.

Les premiers critères typologiques auxquels nous nous intéresserons sont relatifs aux spécificités des systèmes d'écriture. Ces derniers constituent un important paramètre dès lors que l'on souhaite traiter du texte, comme déjà évoqué au chapitre 1. Cela fera l'objet de la section 3.1.
Les seconds critères typologiques, linguistiquement plus pertinents, mais toujours controversés, concerneront les disparités au niveau morphologique. Cette thématique, liée aux problèmes de segmentation et de dispersion des données, sera traitée à la section 3.2.
Enfin, nous nous intéresserons à la typologie de l'ordre des mots, qui sera abordée à la section 3.3.

Cette revue succincte de quelques traits typologiques permettra non seulement de regrouper des langues comparables, mais également de choisir sur quelles bases sélectionner les unités minimales qui serviront à notre analyse. En effet, la question du découpage en unités de traitement primaires paraît triviale pour les langues disposant d'une segmentation en *mots typographiques*. Elle est néanmoins épineuse pour les autres, comme nous l'avons vu à la section 1.2.1.

3.1 Systèmes d'écriture

Les systèmes d'écriture sont des conventions non idiosyncrasiques et arbitraires utilisées pour transcrire des énoncés. Ils peuvent être décodés par n'importe quel locuteur natif entraîné. Ils sont généralement considérés comme secondaires par les linguistes, leur préoccupation principale relevant plus des systèmes phonologiques. Cependant, le traitement automatique des langues repose sur des textes utilisant différents systèmes d'écriture. Dès lors que

l'on cherche à développer des traitements indépendants de la langue, on ne peut ignorer ce paramètre : il est la première différence visible entre les langues.

Depuis « la première synthèse moderne » par Gelb (1952) de l'histoire des écritures, la question des modalités d'une typologie des systèmes d'écriture a continué à se poser de façon subsidiaire pour certains linguistes généralistes (Catach, 1997). La nomenclature qui revient régulièrement en ce domaine est illustrée dans l'article de Baroni (2011), partiellement résumé dans le tableau 3.1.

TABLE 3.1 – Synthèse de la typologie des systèmes d'écriture proposée par Baroni (2011)

PLÉRÉMIQUE (unité sémantique [4])	Morpho-syllabaire	Des pictogrammes ou des idéogrammes tiennent lieu de mots ou de morphèmes	Ancien égyptien, chinois
CÉNÉMIQUE (unité phonique)	Syllabaire	Chaque élément graphique tient lieu de syllabe (normalement de type CV)	Japonais (kana), cherokee
	Abjad	Les consonnes sont représentées, pas les voyelles (diacritiques vocaliques possibles)	Arabe, hébreu
	Alphabet	Note dans l'idéal tous les phonèmes d'une langue, idéalement séparément	Grec, anglais, russe
	Abugida	Syllabaire où les éléments graphiques pour les consonnes et les voyelles sont distinguables	Sanskrit (devanāgarī), guèze
	« Featural »	Alphabet où les formes des signes graphiques sont liées à des traits phonémiques distinctifs	Hangeul coréen

Ce classement typologique scinde les systèmes d'écriture en deux grandes familles : les systèmes d'écriture à dominante *plérémiques*, dont une partie significative des éléments tiendraient lieu d'unités de sens (c'est-à-dire systèmes *idéographiques* et *logographiques*), et les systèmes plutôt *cénémiques*, dans lesquels l'expression phonique prévaudrait en règle générale (c'est-à-dire syllabographiques et phonographiques) (Baroni, 2011). Cette dichotomie simpliste indique, comme souvent en typologie, des tendances générales partagées par certains systèmes au sein desquels des ambiguïtés subsistent malgré tout : aucun système d'écriture n'appartient purement à un type, mais chaque système emprunte, dans des proportions diverses, à chaque type. Par exemple le chinois, dont le système d'écriture a longtemps été qualifié d'idéographique, a vu son statut remis en question dès lors que la majorité des idéogrammes ont été requalifiés de phonogrammes : l'écriture chinoise pour les langues sinitiques relève

3. Baroni (2011) l'admet en note de bas de page, cette définition de l'unité plérémique est bancale. Son pointeur bibliographique transfère le lecteur sur Daniels (2003), mais nous préférons diriger notre lecteur vers l'article plus explicite de Lurie (2006) traitant du « mythe idéographique ». Ce dernier relate les étapes des discussions relatives au fait que certains systèmes d'écriture dit idéographiques ne transcrivent pas prioritairement du sens mais avant tout du son (une syllabe).

plus d'un syllabaire avec beaucoup d'homophones (voir notamment Magistry (2013, p. 50) ou Lurie (2006)). Cette distinction persiste néanmoins dans les typologies des systèmes d'écriture malgré les débats.

Toutes les langues encore écrites aujourd'hui utilisent un des systèmes listés dans le tableau 3.1, mais en combinent parfois plusieurs. Par exemple, le japonais utilise les kanjis, les chiffres arabes, les hiragana, katakana et furigana (systèmes syllabographiques), et l'alphabet latin (système phonographique).

De la même façon, de nombreuses habitudes d'écriture ont émergé récemment avec l'utilisation d'outils technologiques dans les langues européennes telles que l'anglais le français, l'espagnol ou l'italien : l'alphabet est parfois supplémenté idéographiquement (par exemple avec le symbole « ♡ »), et l'orthographe s'« abdjadifie » à certaines occasions (par exemple « *tmb* » utilisé en lieu et place de « *también* » en espagnol) (Baroni, 2011, p. 147). Tous les systèmes d'écriture continuent d'être façonnés par l'histoire et les besoins des locuteurs qui les utilisent.

Depuis l'avènement des technologies de l'information, de nombreuses formes de graphies on du être transposées dans un format numérique. Il existe plusieurs façons parfois contradictoires d'encoder des ensembles d'éléments graphiques (appelés caractères). Néanmoins, l'industrie des technologies de l'information favorise des normes capables d'encoder universellement tous les caractères. À ce titre, un consensus de plus en plus répandu consiste à utiliser la norme Unicode, un jeu universel de caractères garantissant la stabilité des données encodant les jeux de caractères majeurs sur les plans nationaux, internationaux et industriels.

La norme Unicode offre non seulement la meilleure interopérabilité, mais est aussi un avantage en ce qui concerne le traitement automatique de textes :

- elle identifie des catégories de caractères (e.g. *Letter, Lowercase* ;*Punctuation, Dash* ; *Number, Digit* etc.) ;
- elle prend en compte le formatage dans différentes langues ;
- elle permet à des systèmes dédiés d'afficher correctement, en fonction de la langue considérée, les sens d'écriture (sinistroverse ou dextroverse) ;
- n'importe quel code Unicode représente toujours sans ambiguïté le même caractère ; à ce titre, des caractères très similaires issus de différents systèmes d'écriture disposent de codes différents.

L'ensemble des algorithmes développés à l'occasion de cette thèse ont été prévu pour traiter des textes utilisant la norme Unicode sous sa déclinaison la plus utilisée, à savoir l'UTF-8.

3.2 Questions Morphologiques

Les systèmes d'écriture constituent un indice pour comprendre le processus de formation des mots, mais ils ne sont que la partie émergée de l'iceberg. Pour poursuivre cette exploration, il est nécessaire de plonger un niveau d'abstraction plus bas afin d'étudier le niveau morphologique. Dans la théorie morphologique, il est coutume de dire que les mots (ici *mots morphologiques*) sont composés de plus petites unités porteuses de sens appelé *morphèmes*. Ce terme a été employé pour la première fois par Baudouin de Courtenay (1895), qui l'a défini comme étant « la partie d'un mot qui est dotée d'une autonomie psychologique et qui, pour les mêmes raisons, n'est plus divisible. En conséquence, il englobe des concepts tels que la racine (radix), tous les affixes possibles (suffixes, préfixes), les terminaisons représentant des relations syntaxiques, et d'autres » (traduction d'une citation en anglais proposée par Anderson (2013)).

Malgré sa longue existence et son rôle central dans le domaine de l'analyse linguistique, cette notion reste toujours vaguement définie dans la plupart de ses usages. À l'heure actuelle, la définition du morphème donnée par les manuels consiste à dire qu'il s'agit de la plus petite unité porteuse de sens, qui ne peut être divisé en d'autres unités porteuses de sens plus petites, sauf à porter radicalement atteinte à sa signification. Il réapparaît dans différents environnements verbaux avec un sens relativement stable (Syal & Jindal, 2007). Nous discuterons plus avant de cette définition idéale, qui est contredite par des données empiriques, à la section 3.2.3. Nous allons temporairement nous en contenter afin de pouvoir présenter la typologie morphologique traditionnelle.

Différentes classes d'unités morphologiques peuvent être distinguées sur la base de ce qu'elles dénotent (Ducrot & Schaeffer, 1995, p. 431) :

— soit des concepts liés à la réalité, et l'on parle alors de *morphèmes radicaux* ;
— soit des marques grammaticales, généralement appelés morphèmes *flexionnels* ou *dérivationnels*.

Les bases de la typologie morphologique traditionnelle furent formulées durant la première moitié du XIX^{e} siècle par les frères von Schlegel (1908, 1918) et par von Humboldt (1822, 1836) (Haspelmath, 2009). Aux langues sont attribuées différentes catégories morphologiques (isolante, agglutinante, fusionnelle et incorporante) qui, même si elles ont été souvent critiquées, sont aujourd'hui utilisées dans presque toutes les introductions à la linguistique. Dès 1921, Edward Sapir a souligné le côté problématique de cette typologie.

Dans son ouvrage le plus influent (Sapir, 1921), Sapir a contribué à définir plus précisément la nature des concepts exprimés et leur association en but d'améliorer la classification morphologique des langues. Bynon (2004) propose une synthèse de Sapir que nous reprenons dans la suite, pour élaborer sur les trois questions principales auxquelles s'est intéressé ce der-

nier : la nature des concepts (section 3.2.1), leurs techniques de combinaisons dans la chaîne syntaxique (section 3.2.2) et le degré de complexité interne des mots (section 3.2.3). Nous en présentons ici les bases car elle constitue le socle de la typologie morphologique traditionnelle et des problèmes qui en découlent.

3.2.1 La nature des concepts

La typologie mise en œuvre par Sapir (1921) pour établir la classification des concepts exprimés dans la langue examine tout d'abord leur caractère concret et leur faculté à exprimer des relations syntaxiques.

TABLE 3.2 – Synthèse des *types fondamentaux* de concepts donnés par Sapir (1921)

Contenu		*Concept*	*Définition*
Concret	(I)	Basique	Mots ou racines indépendantes (sens purement lexical)
	(II)	Dérivationnel	Modification interne des éléments radicaux ou affixation d'éléments non radicaux à des radicaux. Diffère du concept de base en définissant des idées qui sont sans rapport avec l'idée globale de la proposition, mais qui modifient la signification du radical (création d'un nouveau sens lexical).
Relationnel	(III)	Relationnel concret	Modification interne des éléments radicaux ou affixation d'éléments non radicaux à des radicaux, mais généralement de façon moins intime dans le cas des concepts dérivationnels. Diffère fondamentalement des concepts dérivationnels car les relations qu'ils indiquent ou impliquent transcendent le mot auxquels ils sont immédiatement rattachés (sens partiellement lexical et partiellement grammatical).
	(IV)	Purement relationnel	Affixation d'éléments non radicaux à des radicaux (auquel cas ce concept est indissociable d'un concept de type III) ou par modification interne, par des mots indépendants ou par une position ; sert à relier les éléments concrets de la proposition entre eux (indique uniquement une/des relation(s) grammaticale(s)).

Ces types de concepts, allant de I à IV, sont placés du plus concret au moins concret. D'après Sapir, les extrémités de cette échelle du concret (I et IV) se retrouvent nécessairement dans toutes les langues. Ce n'est pas le cas des concepts intermédiaires (classes II and III).
Cette constatation permet de ranger, sur ce plan, les langues en quatre groupes :

— Groupe A : Relationnel-Pur Simple Types I et IV seulement.
— Groupe B : Relationnel-Pur Complexe Types I et IV + Type II.
— Groupe C : Relationnel-Mixé Simple Types I et IV + Type III.
— Groupe D : Relationnel-Mixé Complexe Touts les types.

Les langues du groupe A ont tendance à transférer l'élaboration du sens d'une phrase sur la syntaxe seule, c'est-à-dire par la position des mots dans la phrase. Les langues du groupe B ont toujours des contraintes syntaxiques fortes mais autorisent la modification du sens de leurs éléments radicaux par les moyens d'affixe ou de changements internes. Les langues du groupe C utilisent la syntaxe uniquement lorsque aucun procédé morphologique n'est disponible pour modifier le sens des éléments radicaux. Enfin, les langues du groupe D construisent le sens en utilisant tout à la fois des contraintes syntaxiques et les processus morphologiques.
Une fois encore, les frontières entre les groupes ne sont pas bien définies lorsqu'il s'agit de classifier une langue en particulier. En effet, les langues peuvent emprunter, à des degrés divers, à plusieurs groupes. Néanmoins, ce plan de classement permet d'identifier les tendances utiles à une typologie linguistique.

3.2.2 Techniques de combinaison

Sapir (1921) a subdivisé chaque groupe présenté ci-dessus concernant « la méthode prévalente pour la modification de l'élément radical ».
Les langues de chaque groupe présenté plus haut peuvent afficher une préférence pour l'*agglutination*, la *fusion* ou le *symbolisme*. Pour le groupe A (et certaines langues appartenant plutôt au groupe B), dans lequel le radical n'est pas modifié, il existe également un autre sous type appelé *isolation*. Bien que par la suite nous présenterons des langues comme appartenant à un groupe en particulier, nous précisons tout de même que ces prototypes morphologiques ne sont pas mutuellement exclusifs. Il s'agit donc encore une fois d'un *continuum* typologique.

Isolation

Dans les langues isolantes, aussi appelé langues *analytiques*, les mots sont idéalement composés d'un unique morphème, couramment associé à un ton. Ces mots ne portent aucun marqueur de relation syntaxique, et la phrase transporte plus de sens que les mots qui la composent. Pour compenser les lacunes d'information qui sont autrement portées par les variations morphologiques, les langues isolantes favorisent l'utilisation de mots grammaticaux ainsi qu'un ordre fixe pour les mots (Eifring & Theil, 2004, chap. 4). On donne souvent le chinois comme exemple de langue se rapprochant du type isolant.

(3.1) 他 特 别 强
tā tè bié qiáng
'Il est particulièrement fort'

Dans la phrase 3.1 (rappel de l'exemple 1.1), il est intéressant de noter que le mot « 特别 » (« *particulièrement* »), composé des caractères « 特 » (« *spécial* ») et « 别 » (« *distinguer, autre* »), va à l'encontre de la correspondance idéale mot-morphème. Une analyse morphologique plus fine est possible. Par exemple, Feuillet (2006, chap. 1) insiste sur la différence entre des *racines isolées* et des *mots isolés* : les langues à racines isolées (par exemple le chinois) n'autorisent pas d'affixe alors que les langues à mots isolés (par exemple l'indonésien) utilisent des affixes pour composer ou indiquer des fonctions grammaticales. Toutes les langues isolantes peuvent exhiber des processus morphologiques. Autrement dit, il n'existe pas de langue purement isolante.

Agglutination

Le procédé d'agglutination fait partie des constructions dites *synthétiques*. Les constructions synthétiques sont des configurations dans lesquelles des relations syntaxiques peuvent être incarnées par des morphèmes ou des mots. Un mot est composé de deux morphèmes ou plus : une racine lexicale et des affixes.
Dans les langues agglutinantes, les morphèmes portant l'information grammaticale (les « éléments relationnels » de Sapir) sont affixés aux racines en suivant un ordre immuable et sans que chaque élément soit modifié. Tous les morphèmes sont clairement identifiables car il n'y a pas ou peu d'allomorphie, c'est à dire que chaque morphème ne possède qu'une réalisation quels que soient ses contextes d'utilisation.
Poussée à l'extrême, l'agglutination peut aboutir à des constructions *polysynthétiques*. Dans ce cas, les mots portent la quasi-totalité, si ce n'est la totalité, des informations syntaxiques et sémantiques de la phrase qu'ils composent.

(3.2) *tə-meyŋə-levtə-pəɣt-ərkən*
⟨SBJ.1.SG⟩-⟨gros⟩-⟨tête⟩-⟨douleur⟩-⟨IMPF⟩
'J'ai un violent mal de tête'

La phrase 3.2 (Skorik, 1977, p. 102) illustre sur le tchouktche (langue paléo-sibérienne parlée dans le nord-est de la Sibérie), le phénomène d'agglutination. Elle contient trois morphèmes lexicaux (⟨big⟩, ⟨head⟩, ⟨ache⟩) et deux morphèmes grammaticaux (⟨SBJ.1.SG⟩ and ⟨IMPF⟩).

Fusion

Dans les langues fusionnelles, les frontières de morphèmes sont troublées. Il n'y a pas de correspondance univoque entre un affixe une fonction grammaticale : un affixe peut remplir plusieurs fonctions grammaticales, et une fonction grammaticale peut être exprimée par plusieurs affixes (allomorphisme). À l'instar de l'agglutination, la fusion fait également partie des constructions synthétiques.

Les propriétés morpho-syntaxiques exprimées par ce paradigme morphologique peuvent être classés dans des *catégories flexionnelles* : nombre, temps, personne, degré, mode, cas, aspect, voix, participialité, etc. Les réalisations de certaines catégories flexionnelles peuvent coïncider (comme c'est le cas par exemple pour l'impératif et le présent subjonctif en anglais) (Lieber & Štekauer, 2005). Cependant, il est fréquent de rencontrer des allomorphies imprévisibles (cf. les exemples 3.3 et 3.4 en français). Ce cas de figure correspond, dans les termes de Sapir, à la *fusion symbolique*.

(3.3) *irai*
⟨aller⟩⟨1.SG.IND.FUT⟩

(3.4) *vais*
⟨aller⟩⟨1.SG.IND.PRS⟩

La notion de *morphe* a été introduite entre autres pour faciliter la description théorique de l'allomorphie. Un morphe correspond à n'importe quelle réalisation phonique d'un morphème qui soit porteuse de sens. En d'autres termes, un morphème peut être considéré comme un ensemble de morphes. Dans les exemples 3.3 et 3.4, ⌈i-⌉ and ⌈v-⌉ sont deux morphes correspondant au morphème ⟨go⟩. La distinction entre agglutination et flexion est un point d'achoppement récurrent en morphologie. Haspelmath (2009) dresse un bilan relativement tranché en faveur de l'illégitimité de la distinction entre agglutination et fusion.

Les principes très généraux ainsi énumérés souffrent néanmoins de certaines imperfections régissant des contre-exemples ou des cas particuliers (supplétion, troncation, morphèmes vides, clitiques, morphèmes liés...). Alors que la plupart des langues jusqu'ici abordées, sauf cas exceptionnel, exhibent une morphologie concaténative compatible avec les différentes techniques de combinaison énumérés ici, il existe des langues pour lesquelles la morphologie doit être envisagée en discontinuité. Autrement, dit, pour ces langues, le système morphologique ne souscrit pas aux règles énoncées jusqu'ici. Parmi elles, on trouve notamment des langues sémitiques.

Prenons pour exemple l'arabe. La majorité des mots arabes, à l'exception des particules, sont exprimés par trois (parfois quatre, et rarement plus ou moins) consonnes radicales représentant la racine sémantique. Cette racine (notée $\sqrt{}$) est modifiée à l'aide d'un patron de voyelles

(infixe discontinu, parfois également appelé « gabarit ») exprimant la catégorie grammaticale du mot, qui s'intercalent entre les consonnes radicales. Les exemples 3.5 et 3.6 montrent deux mots construits à partir de la racine arabe $\sqrt{NQB}$ (Rubio, 2005). Une fois la racine associé à des voyelles, il est alors possible d'y attacher d'autres affixes (Badawi *et al.*, 2013).

(3.5) *munaqqibun*
muC_1aC_2C_2iC_3–
chercheur

(3.6) *tanqībun*
taC_1C_2īC_3–
enquête

Pour de telles langues, la typologie morphologique classique perd de son sens. McCarthy (1981) a proposé de dédoubler l'analyse en deux strates différenciant les analyses morphologiques et phonologiques/prosodiques. Ainsi, les éléments morphologiques peuvent être associés à des formes phonologiques discontinues correspondant pourtant à un seul et unique segment. Sans rentrer dans les détails de sa grammaire morphologique, McCarthy (1981) a grandement influencé la reconnaissance et le traitement de la morphologie non concaténative. Toutefois, séparer le patron de consonnes de ces voyelles intercalées est une procédure coûteuse (Rubio, 2005).

Par ailleurs, ce mode de description n'est cohérent que dès lors qu'il peut être généralisable à un nombre d'observations suffisamment grands pour une langue. C'est ce que fait remarquer Anderson (1992, p. 62) lorsqu'il s'interroge sur le bien-fondé de la méthode de McCarthy (1981) dans une langue à morphologie concaténative, en prenant l'exemple du verbe *dive* (« plonger » en anglais). Ce dernier peut avoir de formes passées : *dived* et *dove*. Selon l'analyse morphologique utilisée pour détecter la racine de ce mot (concaténative ou non-concaténative), on peut obtenir des conclusions paradoxales, voire contre-intuitives , en terme de complexité interne.

3.2.3 Complexité interne

Nous venons de voir qu'une langue peut être typologiquement décrite par les caractéristiques suivantes :

- Le *continuum* de synthèse couvrant les types analytique, synthétique et polysynthétique.
- Les types conceptuels correspondants au groupes A, B, C, ou D de Sapir.
- Le *continuum* de combinaisons couvrant les méthodes d'isolation, d'agglutination et de fusion.

Il n'existe pas de langue qui soit purement d'un type ou d'un autre. Par exemple, le français est à la fois analytique et modérément synthétique (Sapir, 1921). Cela constitue la critique principale adressée à cette typologie. Appliquée à une langue particulière, ces traits semblent être trop désordonnés pour rendre fidèlement compte d'une tendance dominante ou tout simplement être exploitables.
Cette classification de Sapir a été étendue par Greenberg (1960), qui a proposé de quantifier le degré d'appartenance d'une langue à ces types à l'aide de différents index numériques (Tableau 3.3).

TABLE 3.3 – Indexes proposés par Greenberg (1960) et leur rapport à la typologie de Sapir. ↓ indique un indice faible, ↑ un indice élevé.

Index	Formule	Liens avec les types de Sapir
Synthèse	$\frac{\#morphèmes}{\#mots}$	↓ : isolation ; ↑ : polysynthèse
Agglutination	$\frac{\#constructions\ agglutinatinantes}{\#jonctions\ de\ morphèmes}$	↓ : fusion ; ↑ : agglutination
Derivation	$\frac{\#morphèmes\ drivationnels}{\#mots}$	↑ : correspond aux classes B et D de Sapir
Flexion	$\frac{\#morphèmes\ flexionnels}{\#mots}$	↑ : correspond aux classes C et D de Sapir
Composition	$\frac{\#morphèmes\ radicaux}{\#mots}$	Plus d'un morphème radical implique de la composition (sec. 1.2.2.1)

Toutes ces mesures sont basées sur la notion de *mot morphologique* et de morph(ème), et nécessitent de disposer d'une segmentation en morph(èmes). Elles sont applicables la plupart du temps dans le cadre de la morphologie concaténative, mais difficilement pour les langues non-concaténatives.

3.2.4 En résumé

D'après Haspelmath (2009), la tendance actuelle serait d'ignorer ou de critiquer la typologie morphologique du XIXe siècle. À divers égards, elle souffre en effet d'inexactitudes dommageables, spécialement lorsque l'on se penche sur des phénomènes spécifiques d'une langue pour peu que ces derniers aient été envisagés comme des phénomènes marginaux alors qu'en réalité il s'avèrent moins secondaires qu'ils n'y paraissent à l'usage.
Néanmoins, cette typologie morphologique, régulièrement taxée d'« incohérence et d'inuti-

lité » (Comrie (1989, p. 52), Spencer (1991, p. 37-9), Haspelmath (2009) entre autres), a paradoxalement réussi à conserver une place de choix dans la tradition linguistique. Cela est probablement lié au fait qu'il existe un biais récurrent dans la plupart des études typologiques, qui consiste à sélectionner globalement le même ensemble de langues, particulièrement étoffé en langues indo-européennes.

À notre sens, la véritable question soulevée par cette revue des différents types morphologiques relève de la valeur donnée à la notion de *morphème*. Cette dernière semble ne pas être une réalité linguistique mais une abstraction, une notion applicable à un ensemble d'éléments comparables, sur laquelle ont pu être élaborées des théories. En ce sens, envisagé pour le sous-ensemble de langues « standard », le morphème et les notions attenantes peuvent être considérés comme des universaux, c'est-à-dire des tendances fortes qui possèdent des contre-exemples et des cas d'usage boiteux. En revanche, si l'on considère un sous-ensemble de langues plus grand, plus universel, fonder une analyse linguistique sur la notion de morphème pose de nombreux problèmes. Rétrogradé dans son statut d'universel, le morphème pourrait n'être qu'un type morphologique, très répandu dans les langues les mieux étudiées et permettant d'unifier certains traitements. Reste à trouver quels sont les autres types morphologiques englobants, dans les sous univers de langues pour lesquelles le morphème est inadapté, comme en morphologie non-concaténative. Cette question va néanmoins bien au-delà de la problématique de cette thèse.

En ce qui nous concerne, l'approche que nous adopterons pour la sélection des unités minimales (détaillée dans le chapitre 5) sera malgré tout basée sur l'idée de morphèmes (plus spécifiquement sur la notion d'affixes) et la typologie morphologique traditionnelle. Elle répondra pragmatiquement aux questions suivantes :

— *Dans quels cas découper, et comment ?* Cette question sera notamment importante d'une part pour les langues *scriptio continua*, mais également pour les langues à morphologie riche, pour distinguer les affixes du reste.

— *Dans quels cas supprimer, et pourquoi ?* Cette question sera pertinente lorsque l'on soupçonnera qu'un trop grand nombre d'affixes flexionnels ou dérivationnels rendrait inaccessible les informations sémantiques contenues dans les racines.

Quelle que soit la place que l'on souhaite donner au morphème dans son système de valeurs linguistiques, on ne peut ignorer le fait qu'il existe des chevauchements entre morphologie et syntaxe dans la plupart des langues. La section suivante, dévolue à l'ordre des mots, est donc liée à la présente section.

3.3 Ordre des mots

Dans l'éventualité où nous voudrions utiliser un modèle d'extraction terminologique entraîné sur une langue sur d'autres langues, il convient d'aborder la question de l'ordre des mots dans la chaîne textuelle. Cette section s'appuie largement sur l'article de Dryer (2007).
En adéquation avec ses recherches sur la morphologie, Greenberg s'est logiquement penché sur l'ordre d'apparition des différents éléments dans la phrase. Son ouvrage séminal (Greenberg, 1963) a permis de poser les jalons d'une typologie de l'ordre des mots. Quel que soit le type morphologique d'une langue, il s'avère en effet que ses mots/morphèmes sont juxtaposés les uns aux autres pour former une proposition. La tradition linguistique veut que l'on classifie les langues selon l'ordre d'apparition canonique du *sujet* (S), de l'*objet* (O) et du *verbe* (V). L'ordonnancement dit « de base » de ces trois éléments dénote les traits typologiques fondamentaux (Dryer, *ibid.*).
L'ensemble des possibilités théoriquement offertes pour l'arrangement de ces trois éléments permet de dresser le schéma (provisoire) suivant :

— deux cas de figure pour lesquels le verbe se trouve en position initiale (VSO, VOS) ;
— deux cas de figure pour lesquels le verbe se trouve en position médiane (SVO, OVS) ;
— deux cas de figure pour lesquels le verbe se trouve en fin de proposition (SOV, OSV).

Toujours selon Dryer (*ibid.*), quelle que soit la position du verbe, il existe deux configurations qui semblent être rares et mal documentées en typologie : les langues à objet initial (OVS et OSV). Pour le reste, les ordres les plus fréquemment représentés dans les langues sont en grande majorité SOV, suivi de SVO et VSO. Un intérêt supplémentaire à cette typologie basée sur l'ordre du sujet, de l'objet et du verbe est qu'il semble être corrélé à l'ordre d'autres constituants de la phrase. Le tableau 3.4 présente certaines de ces caractéristiques.

TABLE 3.4 – Ordres de constituants corrélés à l'ordre du sujet, de l'objet et du verbe donné par Dryer (*ibid.*)

SOV	VSO / VOS	SVO
Adv V	V Adv	V Adv
NP Po	Pr NP	Pr NP
G N	N G	G N ou N G
St M Adj	Adj M St	Adj M St
Clause Subord	Subord Clause	Subord Clause

Le tableau 3.4 indique que les langues SOV semblent généralement placer les adverbes avant

les verbes, les syntagmes nominaux avant les postpositions, les syntagmes nominaux génitifs (G) avant les noms qu'ils modifient. Dans les constructions comparatives, le standard de comparaison (St), qui est un syntagme nominal auquel autre chose est comparé, est suivi du marqueur de comparaison (M), pouvant, selon la langue, être réalisé soit par un affixe, soit par un mot indépendant. Le tout est suivi de l'adjectif en jeu dans la comparaison. Enfin, dans le cas des subordonnées, les adverbes de subordination sont placés en fin de syntagme subordonné. Toutes ces valeurs sont inversées en ce qui concerne les langues VSO / VOS et SVO, sauf pour les constructions génitives dans les langues SVO, qui semblent ne privilégier aucun ordre.

Une fois encore, cette classification sur la base de l'ordre des éléments n'indique qu'une tendance qu'ont les langues à organiser leurs clauses en suivant un motif prédéfini, dans le cas canonique. Cette organisation idéale est régulièrement mise à l'épreuve par les observations. Tout d'abord, identifier l'ordre du sujet, de l'objet et du verbe n'est pas toujours trivial. Souvent parce que l'ordre d'une langue est trop flexible, mais parfois aussi parce que l'on a du mal à différencier le sujet de l'objet. Un exemple extrême est le tagalog, une langue des Philippines, qui a parfois été décrite comme ne disposant pas de catégorie sujet (Schachter (1996), Dryer (2007, p. 16)).

Il arrive souvent que plus d'un ordre soit possible dans une langue. Néanmoins, selon Dryer (2007, p. 12), lorsque les langues autorisent des ordres alternatifs, un des deux ordres est généralement beaucoup plus fréquent que l'autre. Dryer (*ibid.*) souligne par ailleurs que si plusieurs ordres sont possibles, il se peut qu'une telle différence soit un artefact du texte (que la fréquence ne soit pas un bon indicateur) ou au contraire, que le fait que les tous soient possibles constitue également un indice important.

Même dans le cas où les langues ne contraignent pas l'ordre des mots, on peut souvent observer de petites tendances, car même si l'ordre est flexible pour des éléments, il est moins pour d'autres. Dryer (1995) indique en effet que dans les langues ayant un ordre flexible, un ordre est décrit comme (pragmatiquement) *non marqué* (ou neutre). Il s'agit de l'ordre utilisé dans les clauses qui poursuivent le discours. Au contraire, les ordres (pragmatiquement) *marqués* impliquent « un changement dans la direction du flux d'information ». Néanmoins, il est difficile d'établir des généralités sur quel ordre, marqué ou non, est le plus fréquent comme l'avait fait Greenberg (1966), car cela peut varier d'une langue à l'autre (Dryer, 1995).

Dryer (*ibid.*) suspecte que ces langues flexibles, généralement corrélée à la polysynthèse, soient plus associées à des caractéristiques de type OV que VO. Néanmoins, être une langue flexible constitue en soi un élément de typologie, au même titre que les différentes configurations des langues disposant d'un ordre préférentiel.

3.4 En conclusion

Les unités élémentaires que nous pourrions utiliser comme base pour la suite des traitements que nous souhaitons appliquer sont sujettes, selon les langues, à un degré de variabilité considérable. En effet, même les notions fondamentales de mots et morphèmes sont inappropriées dès lors que l'on recherche l'universalité pour un traitement automatique.
Notre but ici n'est pas de discuter de la validité d'une théorie, de l'existence réelle de notions linguistiques sujettes à débat. Nous préférons plutôt questionner la possibilité que toutes ces notions ayant pu perdurer dans la tradition linguistique malgré l'*incertitude* qu'elles véhiculent soient en réalité de bons indices dont l'idée sous-jacente est exploitable dans le cadre d'un traitement global de l'information.

Les langues humaines ont une *complexité organisée* au sens de Weaver (1948). Ce dernier la distingue de la *complexité désorganisée*, pour laquelle les statistiques seules peuvent faire face, partant du constat que la complexité organisée « implique de gérer simultanément *un nombre considérable de facteurs qui sont interdépendants au sein d'une unité organique* ». Paradoxalement, même si ce système était complètement renseigné, le traiter nécessiterait des recherches combinatoires massives. Afin de modéliser un modèle complet, il faudrait de toute manière en diminuer la complexité en jouant sur l'incertitude que l'on attribuerait à ses paramètres. En d'autres termes, afin d'accroitre l'utilité et la crédibilité d'un modèle, les simplifications à apporter au modèle nous condamneraient à perdre de l'information (Klir & Wierman, 1999). Monde idéal ou pas, la multidimensionnalité engendre de l'incertitude.
Selon Klir & Wierman (1999, p.103), nous sommes régulièrement confrontés à trois types fondamentaux d'incertitude :

- *le flou*, relevant du manque de définition ou de distinction nette,
- *les dissensions*, c'est-à-dire la mésentente lorsqu'il s'agit de choisir entre différentes alternatives,
- *la non spécificité*, lorsque des alternatives ne sont pas précisées.

En ce qui nous concerne, nous considérons que la perspective du multilinguisme (ou, de l'utopique *indépendance de la langue*) fait exploser tous ces types d'incertitude sur les informations linguistiques. Comme le soulignent Klir & Wierman (1999), dissension et non-spécificité peuvent être regroupés sous la bannière de l'ambiguité. Or entre ces deux options, d'un point de vue théorique, nous préférons favoriser la non-spécificité. Ainsi, nous nous positionnerons autant que possible suivant les deux axiomes suivants, correspondants respectivement au flou et à la non-spécificité :

1. Tout est situable sur un continuum (ou vu le nombre de dimensions, dans un espace). Chaque instanciation d'un prototype contribue à nourrir sa description le long d'un

continuum entre intuition [5] et analyse , deux paramètres régis par la fréquence d'observation en contexte. La position d'une langue dans une dimension affecte ses coordonnées dans l'autre.

2. Tout est prototype, depuis les éléments primaires de la théorie linguistique jusqu'aux motifs complexes qui en sont dérivés. Les prototypes peuvent entretenir des relations de subsomption, et sont mieux adaptés aux aléas du multilinguisme. Pour ce qui nous intéresse, nous ne souhaitons pas regarder en détail chaque trait typologique, mais trouver comment lisser ces continuums, et réduire cet espace en déterminant quels aspects principaux gouvernent la construction du sens dans une langue en fonction de sa position dans l'espace typologique. Il s'agit de conserver une granularité qui soit à la fois suffisamment fine pour faire les distinctions essentielles, mais suffisamment générale afin de ne pas faire de sur-classification.

Afin poser les jalons d'un nouveau cadre d'extraction terminologique automatique multilingue permettant de tenir compte des propriétés concrètes des langues, nous ne prendrons aucune dimension typologique comme preuve exclusive du fonctionnement des langues car trop de modulos rentrent en compte. La multilingualité n'étant pas formalisable, une certaine neutralité est de mise. Cet agnosticisme théorique rejoint les principes de la terminologie sociocognitive de Temmerman (2000) : simplifier sans corrompre, grâce à la sous-spécification.

Ce chapitre avait pour but de recenser quelques notions translinguistiquement comparables, exploitables en phase de pré-traitement des corpus pour, dans l'idéal, minimiser l'influence de la langue lors de l'entraînement des modèles d'extraction terminologique. Nous avons tout d'abord abordé la question des systèmes d'écriture. En ce qui nous concerne, cette question n'est plus pertinente dès lors que l'ensemble des textes possède un encodage normalisé tel que l'UTF-8. Puis un bref tour d'horizon des caractéristiques principales de la morphologie des langues nous a permis de rappeler les différences entre les langues isolantes, flexionnelles (à morphologie concaténative ou pas), agglutinantes et polysynthétique. Cette partie sera notamment mise à profit en ce qui concerne les choix effectués pour le calcul des unités de traitement élémentaires dans la partie Expérimentations. Enfin, aborder le sujet de la typologie de l'ordre des mots ici nous permettra éventuellement de comprendre plus tard (section 8.2) pourquoi certains modèles d'extraction terminologique entraînés sur une langue ne sont pas transposables à d'autres langues.

5. Pour Laughlin (1997), l'intuition n'a rien d'illégitime ; il s'agit de la façon dont notre système nerveux crée notre monde d'expérience à partir d'un réseau éminemment complexe de cellules qui interagissent entre elles sous l'influence de notre culture et de nos expériences passées et de bien d'autres facteurs.

Deuxième partie

Extraction de terminologie multilingue

Introduction de la deuxième partie

Cette partie est consacrée à la description des traitements mis en œuvre afin d'aboutir à une extraction terminologique automatique maximalement indépendante de la langue. Les chapitres précédents auront permis de pointer quelques questions problématiques dès lors qu'une extraction terminologique sort d'un cadre linguistique donné pour s'étendre à un grand nombre de langues très différentes. Afin de nous confronter à des cas réels, et dans le cadre de l'évaluation finale, nous avons sélectionné pour ce travail un sous-ensemble de sept langues typologiquement variées pour lesquelles nous disposions de suffisamment de données. Ces dernières seront brièvement décrites, dans la section 4.1.

Une des préoccupation majeure dans le cadre de cette thèse relève de la sélection d'*unités élémentaires de traitement* comparable entre les langues. Distinguer les mots, les tokens ou encore les termes de façon automatique est une tâche qui doit être approximée à défaut d'être faite en conformité avec la tradition linguistique dans chaque langue. L'enjeu des pré-traitements présentés au chapitre 5 sera donc de trouver une définition de ces unité de traitement élémentaire qui soit aux mots, aux tokens, et éventuellement, aux termes, ce que les *unités de compréhension* de Temmerman (2000) sont au concept.

À ce titre, il n'est pas inutile de souligner qu'à différents stades de la chaîne de traitement globale décrite dans cette thèse, la nature des unités de traitement envisagées peut énormément varier.

Si l'on reprend la figure 0.1 (p. 5) schématisant la chaîne de traitement en jeu dans cette thèse, on constate que les éléments constitutifs des corpus en amont sont initialement des tokens, voir des caractères en ce qui concerne des langues dites plérémiques sans segmentation évi-

dente comme le chinois [6]. Les autres ressources impliquées dans la chaîne de traitement (terminologie de référence, graphe de traduction) peuvent quant à elles comporter aussi bien des tokens, de mots lexicaux (ou sémantiques), des lemmes que des expression-multi mots.

Schématiquement, on peut donc considérer que vis-à-vis des unités que l'on manipule, deux stratégies complémentaires seront adoptées :

1. Une approche ascendante (partant du texte brut) imposée, entre autres, par la nécessité de conserver une indépendance vis-à-vis de la langue dans le cadre d'un traitement automatique,
2. Une approche descendante (partant des multiples ressources lexicales que nous avons intégrés dans la chaîne de traitement) prescrite dans le cadre multilingue et pour garantir une information sémantique.

6. La question des langues polysynthétiques est suffisamment marginale en linguistique, et *a fortiori* dans le cadre d'enquêtes internes, pour que nous ne l'évoquions que brièvement par la suite.

Chapitre 4

Langues de test et corpus utilisés

Sommaire

Les données que nous utilisons comme base de l'extraction terminologique sont des verbatim issus d'enquêtes internes que des entreprises multinationales ont proposées à leurs salariés, et compilés par la société *Verbatim Analysis – VERA*. Ainsi, chaque verbatim est un texte court rédigé par un employé dans la langue de son choix. Les répondants à l'enquête sont interrogés sur des problématiques liées à l'entreprise, leur ressenti, et sont invités à donner des perspectives d'amélioration. En ce sens, on peut considérer que l'ensemble des verbatim récupérés dans le cadre de ces enquêtes correspond à un même domaine. En revanche, tous les employés n'abordent pas nécessairement les même thèmes. S'ils le font, ils peuvent ne pas partager la même opinion. De plus, chaque employé a un style d'écriture différent, relatif aussi bien à son niveau d'éducation qu'à son état émotionnel ou d'implication vis-à-vis de l'enquête considérée. Ainsi, l'ensemble des verbatim récupérés durant une enquête constituent un corpus quasi-comparable difficilement exploitable.

Comme indiqué à plusieurs reprises dans le chapitre précédent, la contrainte majeure que nous posons pour le système d'extraction terminologique présenté ici est l'*indépendance de la langue*, autrement dit une grande flexibilité d'application entre les langues. Les chapitres précédents ont préfiguré que cela posait un lot de problèmes non triviaux. La morphologie et l'ordre des mots, entre autres choses, peuvent influer sur la dispersion de l'information (*data sparseness*) dans des modèles n-grammes. La section 4.1 présente, pour les langues sélectionnées pour nos expérimentations, plusieurs particularités à considérer pour s'engager dans différents types de pré-traitements présentés au chapitre 5.

4.1 Langues

Afin d'enquêter sur ces paramètres, nous avons mené l'expérience dans sept langues (pouvant éventuellement comporter des variantes ou des dialectes) pour lesquelles nous disposions de suffisamment de données de référence : l'allemand, l'anglais, l'arabe, le chinois, le français, le polonais et le turc (tableau 4.1). Étant donné la quantité de verbatim dans ces différentes langues, il nous est impossible, dans l'hypothèse où leurs auteurs aurait déclaré une langue particulière mais utilisé un dialecte, de faire la différence entre les deux. C'est pourquoi pour certaines langues pouvant disposer de plusieurs dialectes, nous considérerons qu'il s'agit en majeure partie de variantes standard usuellement utilisées à l'écrit en cas de diglossie. C'est le cas notamment pour l'arabe standard moderne ; il s'agit d'une langue fortement institutionnalisée et donc maîtrisées par la majorité des locuteurs lettrés de dialectes vernaculaires.

Les sections suivantes présentent succinctement quelques traits typologiques liés à ces langues ainsi que leur caractéristiques d'encodage suivant la norme Unicode, et les sources

des corpus génériques utilisés pour ces langues. Sauf indication contraire, les caractéristiques typologiques et géographiques des langues sont issues des données mises à disposition sur les sites du *World Atlas of Language Structure* (WALS) [1] et *Ethnologue* [2].

4.1.1 Allemand

L'allemand est une langue indo-européenne germanique. Le WALS distingue jusqu'à 16 dialectes allemands, parlés pour la plupart en Allemagne, en Suisse, en Autriche, mais la plupart des locuteurs de ces dialectes s'expriment à l'écrit dans ce que l'on peut qualifier d'allemand standard.

L'allemand standard est une langue dont l'ordre du sujet de l'objet et du verbe est relativement libre (les deux ordres dominants étant SOV et SVO) grâce à une morphologie (casuelle) riche, à suffixation dominante. L'allemand possède 3 genres (féminin, masculin et neutre) et 4 cas. Ses marques grammaticales de personnes et de nombre sont syncrétiques. L'allemand, à l'instar des autres langues germaniques occidentales et des langues scandinaves, a une forte propension à construire des composés (voir section 1.2.2.1).

L'allemand utilise l'alphabet latin standard (26 lettres) ainsi que des diacritiques (*umlaut*), et éventuellement un caractère supplémentaire (le « ß »), variante contextuelle de la séquence « ss ». Les différentes lettres peuvent être en majuscule ou en minuscule. L'allemand, comme la plupart des langues européennes, s'écrit de gauche à droite, et utilise l'espace comme séparateur typographique.

En ce qui concerne l'encodage en norme Unicode, les marques diacritiques sont généralement encodées à la suite du caractère avec lequel elles se combinent, mais pas systématiquement. Par exemple, le caractère « Ä » peut être soit stocké comme un « A » suivi de « ¨ » (codes U+0041 U+0308), soit directement comme une lettre à part entière (code U+00C4). Pour peu que des normes diffèrent, on peut donc obtenir le même mot encodé différemment, ce qui fausserait le traitement automatisé de ce mot. C'est pourquoi il est nécessaire d'avoir défini des équivalences entre des chaînes de caractères à l'aide de processus de normalisation, par lesquelles des chaînes de caractères sont converties dans une forme normalisée qui permet une comparaison directe avec d'autres chaînes de caractères normalisées. Unicode propose quatre types de normalisation (NFC, NFD, NFKC et NFKD) , qui garantissent, si elles sont utilisées avec cohérence, un encodage unifié, non ambigu, stable, productif et permettant une composition dynamique (The Unicode Consortium, 2011, ch. 2 et 3).

Dans la suite des expérimentations, le corpus générique utilisé pour l'allemand sera consti-

1. http://wals.info/
2. www.ethnologue.com

tué d'une portion de phrases aléatoirement tirées du *Stuttgart DeWaC*[3] (v3). Ce corpus est lui même une version nettoyée et analysée syntaxiquement du *DeWaC* (Baroni & Kilgarriff, 2006 ; Baroni *et al.*, 2009). Ce dernier, ainsi que tous les corpus issus du projet *WaCky*[4] (*The Web-As-Corpus Kool Yinitiative*) sont des gros corpus construits à partir de données textuelles issues de l'Internet.

4.1.2 Anglais

À l'instar de l'allemand, l'anglais est une langue indo-européenne germanique. Le WALS ne recense qu'une variété d'anglais (décrit comme l'anglais standard), celui parlé en Irlande et au Royaume-Uni, alors que l'anglais est l'une des langues les plus parlées au monde : la base Ethnologue, qui liste les pays dans lequel l'anglais est parlé officiellement, recense en effet plus de 335 millions de locuteurs à travers le monde [5]. Tous les continents comptent :

— un grand nombre de locuteurs natifs, parlant des variantes nationales standard, ou des variétés ethniques, régionales ou sociales (Groupe L1), voire des pidgins et des créoles dérivés de l'anglais (Groupe Pidgin/Créole) ;
— une importante communauté de locuteurs pour lesquelles l'anglais est une langue seconde (par exemple Cameroun, Hong-Kong, Pakistan, Singapour, etc.) ou une langue de contact (anglais des affaires, recherche, etc.) (Groupe L2) ;

Szmrecsanyi & Kortmann (2009) recensent 46 variétés d'anglais, incluant des pidgins et des créoles. À quelques anomalies près, ils déterminent que des généralisations typologiques peuvent être décelées au sein des trois groupes (L1, L2 et Pidgin/Créole). Les différences les plus marquées ont lieu entre les groupes L1 et Pidgin/Créole. Leur conclusion est que les plus grandes divergences typologiques entre ces deux groupes ont lieu principalement au niveau de la complexité morphosyntaxique et du caractère analytique.

Nous nous intéresserons ici aux variétés d'anglais L1 et L2, en postulant qu'elles sont toutes des langues flexionnelles à suffixation dominante, dont les marques grammaticale de personnes et de nombre sont syncrétiques. L'anglais ne marque pas le genre, sauf pour les pronoms pour lesquels il existe trois genres (féminin, masculin et neutre). L'ordre privilégié des constituants est Sujet, Verbe, Objet.

En termes d'encodage, ce qui a été dit pour l'allemand est également vrai pour l'anglais, à la différence que l'anglais n'utilise pas, sauf rares exceptions, de signes diacritiques.

En ce qui concerne le corpus générique en anglais utilisé dans la suite des expériences, il est constitué de lignes aléatoirement tirées du *WaCkypedia_EN* (Baroni *et al.*, 2009), un corpus

3. http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sdewac.en.html
4. http://wacky.sslmit.unibo.it/doku.php?id=start
5. www.ethnologue.com/language/eng

proposé par le projet *WaCky* correspondant à une version de la Wikipédia anglaise récupérée en 2009, étiquetée et analysée en dépendances, dont nous n'avons gardé que le texte brut.

4.1.3 Arabe (Standard Moderne)

Le WALS distingue 21 variétés de l'arabe, parlées depuis le Maghreb, à l'Ouest, jusqu'à la péninsule arabique, en passant par l'Égypte, le Soudan (dialectes centraux) et les pays du Levant. Bien que toutes ces variétés fassent partie de la famille afro-asiatique et des langues sémitiques, certaines ont des différences typologiques. Par exemple, le dialecte égyptien privilégie l'ordre SVO [6] alors que le dialecte syrien peut être indifféremment SVO ou VSO [7].
Concernant l'arabe standard moderne (ASM), qui est la langue privilégiée à l'écrit (mais également dans les médias audiovisuels) dans l'ensemble des pays arabophones, le WALS le référence, entre autres, suivant les traits typologiques suivants : il s'agit d'une langue flexionnelle (suffixation dominante) à morphologie non-concaténative. La négation a tendance à être placée avant le sujet. L'arabe a ceci de particulier que ses différents registres, allant de nombreuses variations dialectales à l'arabe littéraire, en passant par l'arabe véhiculaire, correspondent à différents niveaux de langue, voire à des « patrimoines » culturels hétérogènes. Ainsi, ce sont plusieurs dynamiques linguistiques et extra linguistiques qui rendent difficile une description précise et systématique. En effet, d'une collection de textes à l'autre, on observe des inadéquations. Et cela ne vaut pas qu'entre les différents registres, mais parfois même au sein de chaque registre. Par exemple, l'arabe possède des marques casuelles, mais celles-ci ne sont pas respectées dans l'usage courant de l'arabe standard ou littéral (Baccouche & Mejri, 2007).

L'ASM utilise un abjad (voir section 3.1), constitué de 28 consonnes (pouvant avoir jusqu'à quatre formes distinctes en fonction de leur contexte d'apparition), de deux voyelles longues et d'une ligature. Les voyelles courtes et d'autres marques peuvent être facultativement combinées aux consonnes à l'aide de signes diacritiques (appelés *harakat*). L'écriture se fait de la droite vers la gauche.
En ce qui concerne l'encodage en norme Unicode (The Unicode Consortium, 2011, ch. 8), chaque lettre ne reçoit qu'un seul code [8]. L'interprétation de la forme de chaque caractère en contexte est laissée aux bons soins de l'éditeur de texte. La plupart des marques de ponctuation sont unifiées avec la ponctuation latine, à l'exception de quelques cas de ponctuation typiquement arabe (« ، », « ؛ » , etc.) [9]. Lorsque plusieurs *harakats* viennent modifier une

6. Voir http://wals.info/languoid/lect/wals_code_aeg, trait 81A.
7. Voir http://wals.info/languoid/lect/wals_code_asy, traits 81 A et B.
8. En réalité, chaque variante de chaque lettre arabe est également un caractère unicode. L'utilisation de ces variantes positionnelles est toutefois fortement déconseillée au profit de caractères graphémiques correspondants.
9. De façon générale, la norme Unicode propose systématiquement pour toutes les langues un encodage spécifique des caractères de ponctuation dès lors qu'ils diffèrent dans l'usage de la ponctuation latine dite standard, c'est-à-dire celle utilisée en anglais.

consonne, la norme Unicode ne spécifie initialement aucun ordre séquentiel. Cela peut être problématique étant donné que deux séquences de caractères Unicode différentes peuvent encoder de façon non ambiguë le même sens. Afin de remédier à cela, le standard Unicode prévoit des processus de normalisation (The Unicode Consortium, 2011, ch. 3) permettant d'attribuer un ordre canonique à ces séquences.

Le corpus générique utilisé par la suite est issu d'une sélection de phrases aléatoires du corpus d'arabe contemporain de l'université de Leeds (CCA) (Al-Sulaiti & Atwell, 2006). Ce dernier a été conçu avec le souci particulier de représenter aussi bien l'arabe moderne standard que des variétés régionales en utilisant des données textuelles issues principalement de magazines, de sites web et de journaux.

4.1.4 Chinois (Mandarin)

La Chine fait partie des plus grands pays au monde, et compte plus d'un milliard d'habitants, officiellement répartis en 56 groupes ethniques distincts [10]. Ces derniers peuvent pratiquer un ou plusieurs dialectes/langues. Néanmoins, la langue qui prédomine en Chine et dans les autres pays ayant pour langue officielle le chinois est, notamment à l'écrit, le chinois mandarin (ou *Putonghua*), la langue sino-tibétaine la plus pratiquée au monde.

Le chinois est une langue fortement analytique, ayant un ordre préférentiel Sujet, Verbe, Objet. La morphologie du chinois est très limitée, et principalement suffixale. Il n'existe pas de genre, ou de marqueurs casuels. Le chinois utilise beaucoup le procédé de composition.

Il s'agit d'une langue dite « à dominante plérémique » utilisant un système d'écriture morphosyllabaire (voir section 3.1 pour plus de détails), sans segmentation évidente. La transcription du chinois est faite à l'aide de caractères chinois [11] (sinogrammes ou *hànzì*), qui peuvent être traditionnels ou simplifiés. Par exemple, *hànyǔ*, littéralement « la langue des Han », s'écrit « 漢語 » avec des caractères traditionnels, et « 汉语 » avec les caractères simplifiés. L'utilisation préférentielle de l'un ou de l'autre dépend des zones géographiques dans lesquelles elles sont appliquées. Par exemple en République Populaire de Chine, Singapour et en Malaisie, l'écriture simplifiée est standard. En revanche, à Taiwan, Hong-Kong et Macao, c'est l'écriture traditionnelle qui prime (The Unicode Consortium, 2011, ch. 12). En pratique, une vaste majorité de caractères chinois sont identiques dans leur version simplifiée et traditionnelle. Toutefois, la conversion automatisée de la forme simplifiée à une forme traditionnelle peut poser problème à plusieurs titres. Tout d'abord, étant donné qu'un caractère simplifié

10. La liste des groupes ethniques est par exemple disponible sur le site du portail officiel des groupes ethniques : http://mz.china.com.cn/?action-viewnews-itemid-4643

11. Nous n'aborderons pas ici les différentes normes de transcriptions possibles en alphabet latin.

peut correspondre à plusieurs caractères traditionnels, cette conversion n'est pas nécessairement triviale. Passer de la forme traditionnelle à la forme simplifiée est une tâche plus sûre. Néanmoins, dans les deux cas, il existe des différences dans le lexique entre les communautés de locuteurs du mandarin, ce qui rend la simple conversion souvent insuffisante. Nous nous intéresserons dans ce travail qu'aux écrits utilisant des caractères simplifiés.

Bien que traditionnellement le chinois s'écrive de haut en bas et de gauche à droite, la majorité des écrits modernes, et particulièrement les écrits numériques, ont un sens d'écriture en ligne allant de gauche à droite, à l'instar les langues européennes. La norme Unicode répertorie plus de 75 000 caractères unifiés, regroupés dans différents blocs classés par fréquence (commun, rare, historique, variante unifiable, duplicata etc.) (The Unicode Consortium, 2011, ch. 12). Aucune méthode de tri unifiée n'est proposée par la norme Unicode, car cela dépend des paramètres régionaux (méthode phonétique, basé sur le radical et le nombre de traits, etc.). Les caractères sont classés grâce à une heuristique combinant différentes méthodes dans chaque bloc. Bien que la majorité des caractères les plus courants soient encodés avec succès, il existe toujours des caractères très rares qui ne disposent pas de code, et ce même si des caractères sont constamment ajoutés. Concernant la ponctuation, les remarques faites pour l'arabe s'appliquent également au chinois. Toutefois le chinois utilise rarement des caractères de ponctuation latins, mais un ensemble de caractères de ponctuation distincts à largeur fixe, respectant les usages typographiques du chinois.

Pour ce qui est du corpus générique, nous avons utilisé l'intégralité du *Lancaster Corpus of Mandarin Chinese* (LCMC) (McEnery & Xiao, 2004), soit 45 735 phrases.

4.1.5 Français

Le français est une langue indo-européenne romane parlée principalement en France, mais également en Belgique, en Suisse, au Luxembourg, dans certains pays d'Afrique (Sénégal, Cameroun, Maroc, etc.) et au Canada. Il est également à la base de certains créoles (Guadeloupe, Guyane, Martinique, Réunion).

Le français est une langue flexionnelle (suffixation dominante) possédant 2 genres (féminin et masculin) et aucun cas morphologique, du moins pour les catégories ouvertes. L'ordre privilégié est Sujet, Verbe, Objet. Ses marques grammaticales de personnes et de nombre sont syncrétiques. Concernant la négation en français, elle est optionellement double, parfois discontinue (séparée par un verbe conjugué et d'éventuels adverbes).

Le français utilise l'alphabet latin de 26 lettres, et quelques signes diacritiques (accents, tréma, cédille). L'encodage d'un texte en français suit donc les mêmes principes que pour l'allemand.

En ce qui concerne le corpus générique en français utilisé dans la suite des expériences, il est constitué d'une portion de phrases aléatoirement tirées du *FrWaC* (Baroni *et al.*, 2009),

un corpus proposé par le projet *WaCky* constitué à partir de textes issus de l'Internet français, étiqueté et lemmatisé, dont nous n'avons conservé que le texte brut.

4.1.6 Polonais

Le polonais est une langue indo-européenne slave parlée principalement en Pologne et plus sporadiquement dans les pays environnants.
Avec 7 cas (nominatif, génitif, datif, accusatif, instrumental, locatif et vocatif), le polonais est une langue flexionnelle à morphologie riche et à suffixation prédominante. Le nombre de genres en polonais est une question controversée qui a donné lieu à de nombreuses discussions (Drzazga, 2013, p. 20). Selon Corbett (1983), il est possible de décrire l'ensemble des genres du polonais en utilisant six catégories, dont la séparation est justifiée par des formes d'accord différentes : le féminin, le neutre, le masculin inanimé, le masculin animé, le masculin « dévirilisé » et le masculin personnel. L'ordre privilégié des constituants est Sujet, Verbe, Objet.
L'alphabet polonais comporte 32 lettres. Il est basé sur l'alphabet latin, et comporte des lettres supplémentaires pouvant être interprétées comme des lettres latines associées à des diacritiques. En termes d'encodage, le Polonais ne comporte pas de difficultés supplémentaires par rapport à l'allemand ou au français.

Le corpus générique utilisé est une sélection aléatoire de phrases issues du corpus national du polonais (NKJP) développé par Przepiórkowski *et al.* (2008).

4.1.7 Turc

Le turc, qui est la langue officielle notamment de la Turquie et de la république de Chypre est une langue altaïque oghouze [12] agglutinante (pas ou peu syncrétique), utilisant majoritairement des suffixes (Göksel & Kerslake, 2005). Ces derniers sont en règle générale soumis à une harmonie vocalique suivant des règles phonologiques bien définies.
Le turc est une langue Sujet, Objet, Verbe. On dénombre six cas en turc (nominatif, génitif, datif, accusatif, locatif, ablatif) et pas de genre. La négation est marquée par un morphème lié au verbe (SO[V-Neg]).
L'alphabet turc est basé sur l'alphabet latin étendu (dont le caractère spécial « ı ») et comporte 29 lettres, dont certaines sont des lettres latines assorties à des diacritiques. À ce titre, le turc ne présente aucun problème d'encodage particulier.

C'est un ensemble de phrases choisies aléatoirement dans le corpus turc METU (Say *et al.*, 2002), compilant des textes en 10 genres dans un turc post-1990, qui a été utilisé comme cor-

12. Les langues oghouze sont une branche des langues turques définie sur une base généalogique. Pour plus de détails à ce sujet, consulter Pakendorf *et al.* (2007).

pus générique.

4.2 Synthèse

Comme trop souvent dans des travaux linguistiques multilingues, ce sont les langues indo-européennes qui sont sur-représentées. Cela est dû au fait que les données récupérées dans le cadre des enquêtes internes sont majoritaires dans ces langues. Malgré ce biais, nous avons autant que possible sélectionné des langues aux caractères typologiques variés, et issues de sous-familles linguistiques différentes (germanique, latine, slave). Nous avons par ailleurs essayé de diversifier ce panel de langue avec le chinois, l'arabe et le turc.
L'ensemble des langues sur lesquelles nous allons tester notre système sont des langues à suffixation prédominante. Selon le WALS, cette caractéristique typologique et majoritaire dans les langues du monde : Sur les 697 langues pour lesquelles cette information est renseignée, 293 langues sont à suffixation prédominante, 110 langues n'exhibent aucune préférence, 94 langues utilisent peu d'affixes, et 85 langues sont dites « faiblement suffixantes ». Les 115 langues restantes sont des langues favorisant l'usage des préfixes.

Le tableau 4.1 présente quelques catégories typologiques pour ces langues, généralement regardées comme imprécises et peu pertinentes. Elle permettent toutefois de capturer, même grossièrement, des propriétés générales qui ont un impact sur notre travail. Notre préoccupation étant plus de comparer les types de langue en jeu afin de déterminer dans quelle mesure un modèle d'extraction terminologique peut être appliqué à des langues différentes, plutôt que de plonger dans des détails linguistiques plus fins, nous risquons là une simplification considérable.

Table 4.1 – Résumé de caractéristiques typologiques grossières pour les 7 langues de test

<table>
<tr><th rowspan="2">Langue</th><th rowspan="2">« Genus »</th><th colspan="3">Traits typologiques</th><th rowspan="2">Autre</th></tr>
<tr><th>syntaxiques</th><th colspan="2">morphologiques</th></tr>
<tr><td>Mandarin</td><td>Chinois</td><td rowspan="4">SVO</td><td>Isolation</td><td rowspan="2">analytique</td><td>Composition</td></tr>
<tr><td>Anglais</td><td>Germanique</td><td rowspan="5">Fusion</td><td></td></tr>
<tr><td>Français</td><td>Latine</td><td rowspan="2">analytique</td><td>moyennement synthétique</td></tr>
<tr><td>Polonais</td><td>Slave</td><td rowspan="4">synthétique</td><td>Ordre des mots relativement libre</td></tr>
<tr><td>Allemand</td><td>Germanique</td><td>SVO et SOV</td><td>Composition</td></tr>
<tr><td>Arabe</td><td>Sémitique</td><td>VSO</td><td>Introflexion</td></tr>
<tr><td>Turc</td><td>Oghouze</td><td>SOV</td><td>Agglutination</td><td></td></tr>
</table>

Nous avons volontairement testé notre système sur des langues typologiquement proches

telles que l'anglais et le français. Cela nous permettra de voir dans quelle mesure un même modèle est portable d'un corpus à un autre. Les corpus génériques utilisés sont supposés être équilibrés, c'est-à-dire qu'ils comprennent des textes ayant trait à des sujet et des genres différents.

Les corpus issus des différentes enquêtes internes pour le compte d'entreprises multinationales utilisés correspondent à l'ensemble des réponses aux questions ouvertes dont la langue déclarée est une de nos langue de test. Les différentes langues ne comportent pas le même nombre de verbatim. Le tableau 4.2 présente les tailles des différents corpus (spécialité et générique) pour les corpus « bruts », c'est à dire avant tout pré-traitement.
Les trois dernières colonnes présentent le pourcentage de différences en termes de lignes, tokens et caractères entre les corpus de spécialité et générique. Nous nous sommes efforcés de sélectionner, au sein d'une même langue, des portions (aléatoires) de corpus génériques dont le nombre de lignes, tokens et caractères est comparable à celui du corpus spécifique.

TABLE 4.2 – Taille des corpus avant pré-traitements textuels.

Langue (code ISO)		lignes	tokens	caractères	Diff. entre spec. et gen. (%) ligne	token	char.
ara	spec.	13689	339470	3551658	5,9	9,7	2,2
	gen.	14500	372521	3629184			
deu	spec.	313873	4966095	36603283	0	26,9	11,9
	gen.	314000	6302240	40960578			
eng	spec.	498294	10300976	59517341	0,3	16,5	9,8
	gen.	500000	12004978	65344760			
fra	spec.	147770	3266355	20939366	1,5	19,4	15,3
	gen.	150000	3899476	24139882			
pol	spec.	25201	404318	3061782	11,1	-5,1	-16,6
	gen.	28000	383537	2552243			
tur	spec.	38226	689153	6133174	2	17,1	-4,4
	gen.	39000	807221	5864742			
zho	spec.	56153	—	5499993	-18,5	—	-18,1
	gen.	45750	—	4505542			

Néanmoins, ces différences sont à prendre avec du recul car les corpus n'ont pas été segmentés en tokens ou en phrase. Un tableau similaire présentant des différences entre corpus générique et corpus de spécialité pour chaque langue à l'issue des pré-traitements sera présenté dans le chapitre 5. Bien que dans les étapes suivantes les caractéristiques numériques passent par des processus de normalisation et de discrétisation, nous avons fait le choix d'utiliser des tailles de corpus comparables au sein d'une même langue afin de minimiser les facteurs pouvant influer sur les différences de scores entre les langues lors de la phase d'évaluation. Pour cette raison, les tailles des corpus pour le développement et l'évaluation des modèles d'extraction terminologique seront inférieures pour l'allemand, l'anglais et le français (voir tableau 7.1

p. 119).

Chapitre 5

Pré-traitements textuels

Sommaire

Le but des pré-traitements textuels est de produire, à partir de textes bruts, des unités de traitement élémentaires (que nous appellerons UTE par la suite), dont les caractéristiques sont autant que possible robustes au changement de langue, computationellement viables et malgré tout sémantiquement transparentes (cf. section 1.2.1). Resitué dans le schéma de la figure 0.1 (p. 5) présentée dans l'introduction de ce travail, les pré-traitements se trouvent dans le bloc A, et relient les corpus des différentes langues au module d'extraction terminologique. Si l'on regarde ce même schéma dans son ensemble, on remarque que les unités élémentaires utilisées dans les ressources (corpus, terminologie de référence, graphe de traduction etc.) ne sont pas de même nature. Cela s'explique par l'approche ascendante à laquelle nous faisions référence dans l'introduction de cette partie, qui impose, dans le cadre fortement multilingue, de n'envisager que des unités ne demandant pas de traitement linguistique profond. Ainsi, toute définition relative à la syntaxe est *de facto* exclue. De même, étant donné le large spectre de différences morphologiques au sein des langues que nous souhaitons traiter, une analyse morphologique exhaustive semble inconcevable.

La phase de pré-traitement s'envisage en deux étapes. Dans un premier temps, il s'agira de découper les textes bruts aux endroits les plus propices selon les langues pour isoler des UTE , et de proposer une segmentation en phrases. Ces tâches de segmentation seront abordées à la section 5.1. Dans un second temps, nous allons nous intéresser au problème de dispersion des données dans les langues à morphologie riche. La solution mise en œuvre sera décrite à la section 5.2.

5.1 Segmentation en unités comparables

Savoir où appliquer un découpage pour l'obtention d'UTE comparables aux mots n'est pas une tâche aisée. Comme nous l'avons vu précédemment, la nature morphologique des « mots » d'une langue, voire même le statut phraséologique que l'on souhaite leur conférer, influe grandement sur la sélection possible des points de découpage dans la chaîne textuelle. Nous allons distinguer entre les langues disposant d'un séparateur typographique pertinent pour une segmentation à l'échelle du mot sémantique, et les autres. Bien que cette distinction soit réductrice, elle permettra de proposer une ébauche de solution qui pourra être raffinée dans le cadre de travaux ultérieurs.

5.1.1 Langues avec séparateur typographique pertinent

Pour les langues disposant d'un séparateur comme l'espace, le travail de segmentation, plus communément désigné sous le terme de *tokenisation*, consiste à normaliser les découpages suivant des règles strictes. Ces dernières peuvent être uniquement typographiques (par exemple « séparer du reste par une espace tous les caractères non-alphabétiques ») ou linguistiquement motivées (par exemple « accoler un clitique et son hôte dans un même token ») [1].

Nous avons choisi d'utiliser l'unité de tokenisation mise en œuvre par la chaîne de traitement SxPipe [2] (Sagot & Boullier, 2008). Ce choix a été motivé par plusieurs critères. Tout d'abord, en même temps qu'une segmentation en tokens, le texte est segmenté en phrases avec des règles complexes. L'algorithme de segmentation en phrases est en effet capable de faire la distinction entre un même caractère utilisé en tant que ponctuation finale ou pas. De plus, SxPipe utilise la normalisation NFC des chaines de caractères Unicode. Cette dernière favorise uniformément la forme compacte d'un caractère lorsque qu'elle existe. Pour reprendre l'exemple de « Ä », cette normalisation impose son encodage en un seul caractère « Ä » (U +00C4) plutôt qu'en deux caractères successifs « A »+« ¨ » (U+0041 U+0308). La segmentation en tokens proposée par SxPipe peut être universelle ou disposer de règles spécifiques à une langue qui améliorent la tokenisation. En ce qui nous concerne, et afin de rester fidèle à notre doctrine d'indépendance de la langue, nous avons utilisé pour toutes les langues concernées la version universelle du tokeniseur [3]. Le résultat est une tokenisation naïve. À la suite de cela, les corpus ont été découpés en *segments*, qui sont des phrases ou, lorsqu'elles sont trop courtes, des séquences de plusieurs phrases de qualité typographique (énumérations, listes...) et orthographique raisonnable.

Les textes tokenisés avec SxPipe comportent des informations méta-textuelles contenues dans des balises. Ces dernières ont été intégralement retirées, à l'exception des URL (car ces dernières utilisent une syntaxe normalisée, universelle) afin de ne conserver dans les corpus tokenisées que les éléments (parfois normalisés) qui se trouvaient dans les corpus bruts.

Par exemple, à la phrase du corpus brut :

Nous sommes revenus aux anciennes structures, les 'Centres/usines' d'il y a 30 ans.

SxPipe rend l'analyse :

Nous sommes revenus aux anciennes structures , les {'} " Centres / usines {'} " d'

1. En ce cas, il ne s'agit plus exactement de tokenisation mais de quelque chose qui approxime une segmentation en mots. Plus ce pré-traitement linguistique est spécifique à une langue, plus la distinction entre mots et tokens est explicite.

2. SxPipe est distribuée librement sous licence LGPL, par exemple sur le site du projet Alpage Linguistic WorkBench (http://lingwb.gforge.inria.fr).

3. Commande sxpipe-univ.

il y a {30} _NUM ans .

Finalement, la suppression des informations méta-textuelles en conservant les éléments typographiques normalisés (ici les guillemets), donnera la tokenisation finale utilisée par la suite :

Nous sommes revenus aux anciennes structures , les " Centres / usines " d' il y a 30 ans .

5.1.2 Langues sans séparateur typographique pertinent

Pour les langues ne disposant pas de marques de segmentation évidentes, il ne s'agit plus à proprement parler de tokenisation mais de *segmentation*. Parmi ces langues, il y a celles qui utilisent un système d'écriture *scriptio continua* (comme le chinois, le thai, le lao, etc.), et les langues isolantes utilisant l'espace comme séparateur sans pour autant marquer des frontières entre des unités qui sont des UTE raisonnables (par exemple le vietnamien [4]).

De nombreux algorithmes pour la segmentation des langues ont été décrits dans la littérature. Seules les approches non supervisées, et donc susceptibles d'être plus indépendantes de la langue, nous intéressent ici. Les systèmes non supervisés n'utilisent que des données brutes pour essayer d'induire une segmentation à partir de la distribution des caractères observés dans le corpus. On pourra se rapporter à Magistry (2013) pour un état de l'art complet sur les différentes techniques utilisées en segmentation non supervisée.

Le système de segmentation non supervisé de Magistry (2013) est basé sur la variation d'entropie de branchement normalisée (*Normalised Variation of Branching Entropy*, ou NVBE), couplée au principe de longueur de description minimale (*Minimum Description Length*, ou MDL). Bien qu'initialement développée pour la segmentation du chinois pour laquelle elle était, à l'heure de notre recherche, au niveau de l'état-de-l'art [5], Magistry (2013) indique que cette méthode pourrait obtenir des performances similaires sur d'autres langues sinitiques. Cette méthode a également été testée sur des langues ayant des systèmes d'écriture à dominante cénémique dont notamment le français, pour lequel les espaces ont été au préalable éliminés. Formellement, la différence majeure réside dans le fait que la taille de l'ensemble des symboles d'entrée est significativement plus petite (plusieurs milliers pour le chinois mandarin, quelques douzaines pour les systèmes d'écriture dit cénémiques). La figure 5.1 présente

4. À l'origine, le vietnamien utilisait plusieurs systèmes d'écritures différents (les *chữ Hán* et *chữ Nôm*, basés sur des idéogrammes chinois). À partir du XVII^{e} siècle, les missionnaires jésuites proposèrent une transcription du vietnamien avec des caractères latins (Healy, 2012). Cette version romanisée (*chữ quốc ngữ*) est l'écriture officiellement utilisée à l'heure actuelle, et celle à laquelle nous faisons référence. Si les espaces du vietnamien ne sont pas des frontières raisonnables de mots, c'est justement parce qu'il s'agit de transcriptions en alphabet latin d'unités proches de celles désignées par les caractères en chinois.

5. Magistry (2013, p. 130) fait état d'un f-score moyen de $0,80$, alors que le score de référence, détenu par Zhikov *et al.* (2010) obtenait sur le même ensemble de tests un f-score moyen de $0,78$. À l'heure où nous écrivons, la proposition de Magistry (2013) a pu être encore améliorée par Chen *et al.* (2014).

un exemple de segmentation obtenue pour le français par l'algorithme de Magistry (2013). Cette segmentation comporte de nombreuses erreurs. La sur-segmentation sur les frontières de morphèmes est probablement liée à la richesse de la morphologie (Magistry, 2013, p. 163). Pour autant, compte tenu du caractère entièrement non-supervisé de l'approche utilisée, le résultat reste honorable.

FIGURE 5.1 – Extrait d'un exemple de segmentation NVBE sur le français (Magistry, 2013, p. 177).

En même temps ,le contrôle del' exercice du métier se fait plus ri g our eux ,t ant sur letitre du métal précieux employé que sur l esqualités re qui ses del' orfèvre qui,pour accéder à lamaîtrise ,se voit dansl' obligation , désormais clairement spécifi ée, de fournirun che f d'oeuvre .

En l'absence de comparaison des scores de cet algorithme de segmentation appliqué à d'autres langues, nous ne pouvons affirmer catégoriquement qu'il s'agit d'une solution universelle. Nous avons malgré tout choisi de l'utiliser dans nos expériences pour le chinois, langue pour laquelle il a été initialement développé et évalué comme de niveau état-de-l'art. Nous soulignons le fait que sur certaines langues qui, comme le français, ont une morphologie relativement riche, cette solution ne serait probablement pas optimale.

En utilisant cette solution, à la phrase en chinois du corpus brut :

特别是工作了近三四年的人也无机会。

(lit. « En particulier, les gens qui travaillent depuis environ 3 ou 4 ans n'ont pas d'opportunité. »), l'algorithme de Magistry (2013) propose la segmentation (les espaces ont été remplacées par des points) :

特别是 • 工作 • 了 • 近 • 三 • 四 • 年 • 的 • 人 • 也 • 无 • 机会 • 。

Ce découpage regroupe les unités du texte ainsi :

particulièrement être ▪ travail ▪ ⟨PST⟩ ▪ approximativement ▪ 3 ▪ 4 ▪ an ▪ ⟨GEN⟩ ▪ homme ▪ aussi ▪ ⟨NEG⟩ ▪ opportunité .

Concernant la segmentation en phrases, cette dernière a été effectuée simplement à l'aide de la ponctuation finale, ou était déjà disponible, notamment en ce qui concerne le corpus générique.

Une fois la segmentation en tokens et en phrases effectuée, nous avons recalculé la taille des corpus utilisés pour nos tests (tableau 5.1). Les différences entre le nombre de tokens contenu dans les corpus segmentés génériques et de spécialité ne dépassent pas ±20%. La suite des

pré-traitements ne fera pas varier ces comptes.

TABLE 5.1 – Taille des corpus après pré-traitements textuels.

Langue (code ISO)		lignes	tokens	caractères	Diff. entre spec. et gen. (%) ligne	token	char.
ara	spec.	14381	345562	3562317	0,9	7,4	1,9
	gen.	14514	371119	3628526			
deu	spec.	370358	5667001	37361623	-5,8	12	9,1
	gen.	348883	6349289	40745951			
eng	spec.	558384	11255706	60675716	2,4	8,1	7,3
	gen.	571634	12171254	65130918			
fra	spec.	166215	3744155	21425743	14,8	17,8	15,5
	gen.	190873	4412071	24744023			
pol	spec.	29026	462160	3124833	5,7	-16,6	-18
	gen.	30666	385309	2562183			
tur	spec.	50597	764441	6202242	51,7	6,4	-5,4
	gen.	76764	813669	5869526			
zho	spec.	56153	1166760	6610601	-18,5	-8,4	-16,4
	gen.	45750	1068830	5528637			

5.2 Sous-spécification sémantique

Cette phase de pré-traitement, qui a lieu après la tokenisation, vise à trouver un compromis raisonnable entre (1) n'importe quel type de mot sémantique ayant sa place dans des ressources de traduction ou onto-lexicales et (2) une version latitudinaire du mot morphologique. Ce pré-traitement s'appliquera à des langues disposant de procédés morphologiques complexes afin de limiter les effets de dispersion des données.

Nous avons abordé précédemment (voir notamment la section 3.4) les avantages de sous-spécifier des unités en les amputant d'une partie de leur information, quitte à créer de l'incertitude. Par ailleurs, nous partons de l'approximation selon laquelle un token informé peut être assimilé à une unité sémantique ou à un composant d'unité sémantique porteur de sens. Il s'agira donc d'identifier de façon automatique les parties d'un token informé qui portent suffisamment d'information sémantique pertinente, pour pouvoir se débarrasser du reste.

La proposition faite par Anderson (1992, p. 71) ou Allwood *et al.* (2010) consiste à extraire des unités à mi-chemin entre mots et morphèmes : les radicaux (en anglais, *stems*). Ces derniers sont des « mots, moins les affixes flexionnels (productifs) ». À la différence des racines qui sont irréductibles, les radicaux sont des formes prises par la racine dans des réalisations diverses. Cette proposition constitue en effet un compromis raisonnable pour unifier, dans l'approche ascendante, les unités de traitement élémentaires qui nous intéressent, à la différence près que nous souhaitons également retirer les affixes dérivationnels ou tout autre élément issu d'un

découpage morphologique dès lors qu'ils influent de façon significative sur la dispersion des données. En ce qui nous concerne, nous préférons parler de *sous-spécification* plutôt que de *(quasi-)racinisation*, parce que les traitements qui seront appliqués à l'issue de la segmentation morphologique introduiront beaucoup plus d'incertitude sémantique. Nous nous autorisons cette suppression brutale uniquement parce qu'en aval des traitements, les tokens informés seront analysés en contexte et l'incertitude sera donc réduite.

On peut imaginer que des termes de la première colonne du tableau 5.2 puissent être symbolisés par l'unité sous-spécifiée présentée dans la deuxième colonne. Le symbole ⊧ annonce que ce qui suit est une unité sous-spécifiée, qui ne possède pas nécessairement une valeur sémantique claire (dont une proposition de traduction est annoncée par le symbole ≈). Ces dernières sont désignées par le terme d'UTE .

Table 5.2 – Exemples d'unités de traitement élémentaires (préfixées de ⊧) pouvant modéliser une forme sous-spécifiée de tokens informés

Langue	Token	Sous-spécification possible
arabe	بالامكان « possible »	⊧ امكان [6] ≈ possibilité
allemand	*beziehungsrealitäten* « formes de relations »	⊧ beziehrealität ≈ relier réalité
anglais	*ill-standardized* « mal standardisé »	⊧ standard ≈ standard
français	refidéliserait	⊧ fidéli
polonais	*zapytaniami* « requêtes »	⊧ pyta ≈ demande
turc	*fabrikalarda* « dans les usines »	⊧ fabrika ≈ usine

Les sections suivantes décrivent les approches testées pour obtenir ce type d'unités sous-spécifiées, dans les langues nous concernant. Ces approches reposent sur une analyse morphologique des tokens informés (section 5.2.1), et consisteront à élaguer de l'information morphologique (section 5.2.2).

5.2.1 Analyse morphologique non-supervisée

La question du bien fondé de la notion de morphème avait été abordée à la section 3.2.4. Nous avions toutefois annoncé qu'elle conserverait une place importante dans la suite des traitements. C'est justement parce que la plupart des solutions actuelles capables de répondre

6. Noter qu'en arabe, la forme des lettres est contextuelle. بالامكان correspond à بال + امكان.

à la question que nous nous posons de façon non supervisée s'appuient sur cette notion que nous avons fait le choix de garder des frontières théoriques aux morphèmes. Ce parti pris est donc moins un assujettissement à une école théorique qu'à une exigence pratique.

Hammarström & Borin (2011) proposent un état de l'art exhaustif des méthodes de segmentation morphologique non supervisée. Ces systèmes prennent souvent en compte certains procédés morphologiques et pas d'autres. Beaucoup d'entre eux ont été développés pour analyser uniquement des langues à morphologie concaténative et compositionnelle.
La granularité de l'analyse morphologique donnée en sortie peut varier d'une simple liste d'affixes à une liste de paradigmes associés à une liste de radicaux reliés au paradigme qu'ils utilisent. Hammarström & Borin (2011) distinguent quatre grandes approches fondamentales dans lesquelles classifier les algorithmes d'analyse morphologique non supervisée :

— Celles qui segmentent morphologiquement sur la base de la co-occurrences de chaînes de caractères partielles adjacentes (approches à frontières et fréquence). Les systèmes non-supervisés les plus répandus utilisant cette approche sont *Morfessor* (Creutz & Lagus, 2005) et *Linguistica* (Goldsmith, 2000).
— Celles qui regroupent les éléments ayant des ressemblances morphologiques sur la base de métriques (principalement les distances d'édition, parfois des caractéristiques sémantiques, des similarités distributionnelles ou autres). Une fois ces groupes obtenus, des patrons morphologiques récurrents sont identifiés dans chaque groupe (approches de groupement et d'abstraction). Bernhard (2010) a proposé des méthodes non-supervisées relevant de cette approche : *MorphoNet* et *MorphoClust*.
— Celles qui voient un mot comme étant un ensemble de traits (n-grames ou chaînes de caractères partielles selon la méthode utilisée). Les traits apparaissant sur de nombreux mots ont un pouvoir sélectif faible, alors que les traits rarement vus sont des indicateurs de mots ou de racines spécifiques. L'idée sous-jacente est comparable à celle du TF-IDF. La classification d'un mot inconnu revient à utiliser ces traits pour déterminer de quel(s) autre(s) mot(s) il peut être une variante morphologique (approches « traits et classes »).
— Celles qui, considérant qu'il existe une approximation correcte entre phonèmes et graphèmes, distinguent les éléments vocaliques et consonnantiques afin de découper le mot en « squelettes » (vocaliques et consonnantiques) sur lesquels appliquer l'approche à frontières et fréquence (approches par catégorie phonologique et séparation). Cette démarche vise spécifiquement les morphologies non concaténatives.

Il est possible de combiner certaines approches, notamment les approches à *frontières et fréquences* et celle *de groupements et d'abstraction*. L'approche à *frontière et fréquence* est particulièrement adaptée à la morphologie concaténative, alors que les approches *traits et classes*,

et *catégories phonologiques et séparation* ne traitent pas nécessairement des segments contigus. Concernant les langues polysynthetiques, la question de savoir s'il vaut mieux procéder à une segmentation non supervisée ou à une analyse morphologique n'a pas été approfondie ici, notamment car les langues polysynthetiques dépassent le cadre de cette thèse. Néanmoins, afin de nourrir cette réflexion, il serait utile de regarder quelles sont les unités élémentaires à extraire les mieux compatibles sémantiquement avec la tâche de traduction qui sera utilisée par la suite. Une ébauche de réponse est proposée par Nicholson *et al.* (2012).

Concernant les performances des différents systèmes, il est relativement difficile de les mesurer. Tout d'abord, parce que ces derniers sont, la plupart du temps, testés sur des données d'évaluation relativement limitées, dans un petit nombre de langues, souvent différentes. Ensuite parce que les types d'analyses voulues en sortie ne sont pas systématiquement comparables.

La campagne d'évaluation *MorphoChallenge* (Kurimo *et al.*, 2010a,b) a été conçue dans le but d'évaluer les analyseurs morphologiques basés sur des outils d'apprentissage automatique statistique. Elle permet de comparer les résultats de différents systèmes dans au plus 5 langues (anglais, finnois, allemand, turc et arabe) pour 3 tâches :

- comparaison à des segmentations en morphèmes de référence,
- évaluation au sein d'un tâche d'extraction d'information,
- évaluation dans une application de traduction automatique.

Selon l'année et la tâche envisagée, certaines langues ne disposent pas de données d'évaluation. C'est notamment le cas de l'arabe, moins souvent évalué que le reste des langues.

La majorité des systèmes disponibles ont été conçus pour traiter des langues à morphologies concaténatives. Pour les langues sémitiques à morphologie non-concaténatives, les algorithmes d'analyse morphologique peuvent être classés selon deux niveaux d'analyse (Xu *et al.*, 2002 ; Al-Sughaiyer & Al-Kharashi, 2004) :

- La recherche de radicaux ; dans ce cas, l'algorithme n'identifie que les affixes. Parmi ceux là, l'analyseur de Buckwalter (2002) (qui fait usage d'un moteur de règles spécifiques à l'arabe) ;
- La recherche de racines, auquel cas les radicaux identifiés sont réduits à des racines. Parmi ces derniers on trouve ALPNET (Beesley, 2001) (fonctionnant avec un transducteur fini) , le *Khoja Stemmer*[7] (Khoja & Garside, 1999) (qui combine moteur de règles et listes d'éléments spécifiques en arabe), ou encore le *ISRI Stemmer*[8] (Taghva *et al.*, 2005) (utilisant également un moteur de règles).

Parmi ces systèmes, tous ont été conçus spécifiquement pour l'arabe et aucun ne repose sur

7. Téléchargeable à l'adresse : http://zeus.cs.pacificu.edu/shereen/research.htm

8. Implémenté dans module NLTK ISRI. Voir http://nltk.googlecode.com/svn/trunk/doc/api/nltk.stem.isri-module.html

de l'apprentissage automatique. Il existe un nombre relativement limité de publications faisant état de recherches sur des analyseurs morphologiques non-supervisés destinés à traiter des langues à morphologie non-concaténative de façon générique. À notre connaissance, les seules approches proposées dans ce cadre précis sont celle de Xanthos (2008), qui joue sur la séparation des aspects phonologique et morphologique pour formuler des règles de combinaison, et celle de Khaliq & Carroll (2013), qui peut notamment traiter de l'arabe « naturel », c'est à dire sans voyelles courtes ou marqueurs diacritiques. Toutefois, aucune implémentation n'est disponible au téléchargement à l'heure où nous écrivons. Pour cette raison, nous avons décidé de traiter l'arabe avec le même analyseur morphologique utilisé pour les autres langues, et ainsi n'effectuer qu'une analyse au niveau des radicaux et non des racines, au sens de Xu *et al.* (2002). Bien que racines et gabarits aient peu de chances d'être dissociés lors de cette analyse, des suffixes et des préfixes seront malgré tout identifiés.

Pour la campagne d'évaluation de 2010, le système non-supervisé obtenant les meilleurs résultats globaux [9] était *Morfessor U+W*, une extension de la version non supervisée de l'algorithme *Morfessor Baseline* décrit par Virpioja & Lagus (2010). Ce dernier fait partie de la famille des approches à *frontières et fréquence*, et est donc développé pour des langues à morphologie concaténative complexe (comme le finnois ou le turc), mais fonctionne également pour toutes les langues comportant des mots composés ou des affixes non-fusionnels (Virpioja *et al.*, 2013). En plus d'une segmentation, *Morfessor* propose un étiquetage des éléments segmentés en fonction de leur type (suffixe, préfixe, radical). Des exemples d'analyses rendues par *Morfessor* sont présentées dans le tableau 5.3. Cela nous permettra de faciliter la sélection des éléments morphologiques à supprimer pour obtenir nos UTE .

La version la plus récente de l'outil *Morfessor* (v. 2.0.1) (Kohonen *et al.*, 2010 ; Virpioja *et al.*, 2013) apporte quelques corrections et améliorations à la précédente implémentation. Nous avons donc utilisé *Morfessor* v. 2.0.1 pour la décomposition morphologique des langues à morphologie concaténative de notre échantillon. Le programme possède des paramètres de base permettant de l'utiliser tel quel. Il est toutefois possible de faire varier deux paramètres :

— La valeur pour l'amorce (*random seed*) de la fonction aléatoire utilisée pour la production du modèle de segmentation de référence initial (non-deterministe). Nous avons conservé la valeur par défaut, qui est zéro.

— Le seuil de perplexité b, qui dépend de la taille des données. Les recommandations sont d'augmenter la valeur par défaut de $(b = 10)$ pour un plus grand nombre de données. Incidemment, plus b est grand, plus la précision de la segmentation augmente au détriment du rappel. À l'inverse, un b plus petit aura tendance à favoriser la sur-segmentation. En ce qui nous concerne, le choix du paramètre b dépendra plus

9. Évaluation pour les langues à morphologie concaténative de *Morpho Challenge*

des types de langues que l'on souhaite traiter que de la taille des corpus. Différents tests sur les langues concernées (avec $b = 10$, $b = 100$, $b = 300$ et $b = 400$) nous ont permis de déterminer empiriquement que, pour les langues dites « analytiques » (voir tableau 4.1), plus ce seuil était petit, meilleur étaient le découpage obtenu pour ce que l'on souhaite en conserver. En revanche, pour les langues synthétiques, favoriser un b petit, et donc une sur-segmentation dégrade les performances de Morfessor en ce qui concerne l'identification des racines. Les tests nous ont permis de déterminer empiriquement que la valeur $b = 100$ produisait un découpage raisonnable pour ces langues. Cette valeur sera donc le paramètre de référence pour les langues synthétiques. Toutefois, nous appliquons une exception à cette règle pour l'arabe. Le fait que sa morphologie non-concaténative ne soit pas traitée par Morfessor rend sa segmentation plus difficileme. Ce désavantage est donc (modérément) neutralisé en retenant, pour cette langue, $b = 10$.

5.2.2 Degré de sous-spécification

Dès lors que nous possédons une segmentation morphologique disposant d'étiquettes pour l'ensemble des mots typologiques obtenus lors de la phase de segmentation, jusqu'à quel point est-il raisonnable d'enlever de l'information à ces unités pour la suite des traitements ? Le tableau 5.3 reprend les exemples du tableau 5.2, pour illustrer l'analyse morphologique rendue par Morfessor.

Table 5.3 – Exemples d'analyses rendues par Morfessor pour les tokens présentés dans le tableau 5.2

Langue (code ISO)	Token	Analyse de Morfessor
ara	بالامكان	ب/PRE+ال/PRE+امكان/STM
deu	*beziehungsrealitäten*	be/PRE+zieh/STM+ung/SUF+s/SUF+ real/STM+ität/STM+en/SUF
eng	*ill-standardized*	ill-/PRE+standard/STM+ized/SUF
fra	refidéliserait	re/PRE+fidéli/STM+s/SUF+ e/SUF+r/SUF+ait/SUF
pol	*zapytaniami*	za/PRE+pyta/STM+nia/SUF+mi/SUF
tur	*fabrikalarda*	fabrika/STM+lar/SUF+da/SUF

Deux stratégies basiques s'offrent à nous : une modérée, et une franche.

La *sous-spécification franche* consiste à supprimer tout ce qui n'est pas étiqueté par l'analyseur morphologique comme une racine. Le tableau 5.4 présente des exemples d'unités sous-

spécifiées et de leurs tokens correspondant, obtenus avec la stratégie franche.

TABLE 5.4 – Exemples de sous-spécification franche pour les tokens présentés dans le tableau 5.2

Langue (code ISO)	Token	Résultat d'une sous-spécification franche
ara	بالامكان « possible »	⊧ امكان ≈ possibilité
deu	*beziehungsrealitäten* « formes de relations »	⊧ ziehrealität ≈ dessiner_réalité
eng	*ill-standardized* « mal standardisé »	⊧ standard ≈ standard
fra	refidéliserait	⊧ fidéli
pol	*zapytaniami* « requêtes »	⊧ pyta ≈ demande
tur	*fabrikalarda* « dans les usines »	⊧ fabrika ≈ usine

La *sous-spécification modérée* consistera à n'enlever qu'une partie de ce qui n'a pas été identifié comme étant une racine par l'analyseur morphologique. Cette partie doit être, si possible, la plus productive d'un point de vue morphologique. Une fois encore, nous allons utiliser une information typologique pour prendre une décision. Ce sera ici la catégorisation des différentes langues par rapport au trait typologique « préfixation ou suffixation en morphologie flexionnelle » (Dryer, 2013) défini dans WALS [10]. Il a déjà été question de cette caractéristique typologique à la section 4.2. Elle classifie les langues en six classes :

(a) Deux concernant une suffixation prédominante (faiblement et fortement),

(b) Deux concernant une préfixation prédominante (faiblement et fortement),

(c) Une classe statuant aucune préférence entre préfixes et suffixes,

(d) Une classe regroupant les langues qui utilisent peu d'affixes.

Sachant que les langues utilisant peu d'affixes ne sont pas concernées par cette étape de prétraitement morphologique, ce classement nous permet, dans la plupart des cas, de déterminer si l'on préfère supprimer les suffixes (cas (a)) ou les préfixes (cas (b)). Étant donné que les langues sur lesquelles nous allons tester cette méthode sont toutes catégorisées comme *à suffixation fortement prédominante*, le même processus de pseudo-racinisation peut leur être appliqué : il consiste à supprimer l'ensemble des éléments étiquetés comme suffixes, quel que

10. Trait 26A. Précisions sur la page http://wals.info/feature/26A#2/22.6/152.8

soit leur emplacement dans la chaîne initiale. Pour les langues n'affichant aucune préférence, il n'est pas certain qu'une sous-spécification modérée soit possible. Le résultat de cette stratégie pour les exemples du tableau 5.3 sont présentés dans le tableau 5.5.

Table 5.5 – Exemples de sous-spécification modérée pour les tokens présentés dans le tableau 5.2

Langue (code ISO)	Token	Résultat d'une sous-spécification modérée
ara	بالامكان « possible »	⊧ بالامكان ≈ possible
deu	*beziehungsrealitäten* « formes de relations »	⊧ beziehrealität ≈ relier_réalité
eng	*ill-standardized* « mal standardisé »	⊧ ill-standard ≈ mauvais-standard
fra	refidéliserait	⊧ refidéli
pol	*zapytaniami* « requêtes »	⊧ zapyta ≈ demander
tur	*fabrikalarda* « dans les usines »	⊧ fabrika ≈ usine

Il serait possible d'imaginer d'autres règles plus complexes de sous-spécification, qui pourraient par exemple prendre en compte le nombre de préfixes et de suffixes, et n'en supprimer qu'une partie. Nous nous limiterons à tester les effets des sous-spécifications modérée et franche. La figure 5.6 présente les rapports entre le nombre de formes [11] pour les tokens, les UTE modérées (issues de la stratégie modérée) et les UTE franches (obtenues avec la stratégie franche).

Comme attendu, les langues ayant une morphologie concaténative riche [12] sont particulièrement affectées par ce pré-traitement.

5.3 Remarques sur les mots génériques (*stopwords*)

Un mot générique est un mot qui apparaît très fréquemment dans un document, mais dont le sens est négligeable, ou trop ambigu, dès lors qu'il est isolé.

Ce genre de mot est généralement recensé dans une liste, ce qui permet d'évincer ces mots de la suite des traitements. Par exemple, Petrović *et al.* (2009) font état de nettes améliorations des performances de leurs patrons d'extension de mesures d'associations dans le cas où ces

11. Une forme regroupe toutes les instances d'une unité typologique dans un corpus.
12. A l'exception de l'allemand, dont les mots composés n'ont pas été scindés mais simplifiés.

TABLE 5.6 – Influence des approches de sous-spécification modérée et franche sur le nombre de formes dans les corpus. Les pourcentages s'envisagent par rapport aux comptes indiqués dans la colonne « base (tokens informés) ».

Langue (ISO)	UTE distinctes				
	base (tokens informés)	modérées		franches	
ara	78 073	54 933	(70,4%)	25 925	(33,2%)
deu	444 258	321 856	(72,4%)	270 901	(61%)
eng	492 489	354 099	(71,9%)	263 475	(53,5%)
fra	140 744	88 844	(63,1%)	71 555	(50,8%)
pol	89 326	58 894	(65,9%)	53 361	(59,7%)
tur	147 871	30 196	(20,4%)	25 534	(17,3%)

derniers étaient systématiquement ignorés lors des calculs. Leur identification est également primordiale dans le domaine de l'Extraction d'Information, où les mots génériques peuvent inclure des items lexicaux qui ne sont pas des mots génériques mais génèrent malgré tout trop de bruit. À ce titre, nous préférerons, malgré l'anglicisme, utiliser le terme *stopword* pour désigner un mot que l'on ne souhaite pas inclure dans les traitements en aval.

Cette liste peut donc varier d'un corpus à l'autre et d'une application à l'autre (listes spécifiques) même s'il existe des listes de référence dans la plupart des langues (listes génériques). Pour déterminer quels mots sont à ajouter dans une liste de *stopwords*, plusieurs méthodes sont possibles.Tout d'abord, il est possible d'effectuer la recherche de *stopwords* dans des corpus déjà pré-traités (c'est-à-dire lemmatisés, racinisés, etc.). Pour cette raison, nous parlerons d' UTE plutôt que de mots ou de tokens, afin d'englober l'ensemble des unités possiblement crées, indépendamment de la langue, à l'issue du pré-traitement.

La méthode de référence consiste à récupérer les n UTE les plus fréquentes, c'est-à-dire l'infime partie des UTE qui constituent une part significative d'un corpus. En plus de cela, il est possible, pour des langues utilisant un système d'écriture à dominante cénémique, de ne prendre, parmi ces n UTE, que ceux qui ne comportent qu'un petit nombre de lettres (surtout si ces UTE sont des mots). Ces derniers sont appelés « stopwords légers » (Sadeghi & Vegas, 2014), et correspondent en général aux mots grammaticaux. Wilbur & Sirotkin (1992) définissent les *stopwords* comme étant des mots ayant la même probabilité d'apparaître

dans n'importe quel document (de spécialité ou générique). Ils proposent ainsi d'identifier les *stopwords* à l'aide de tests statistiques. Lo *et al.* (2005) utilisent par exemple l'IDF, couplée au calcul de divergence de Kullback-Leibler (Kullback & Leibler, 1951) entre distributions des mots dans plusieurs corpus.

En relation avec la problématique des langues ne disposant pas de segmentation évidente, Zou *et al.* (2006) ont proposé une méthodologie pour le chinois (également utilisée sur l'arabe par Alajmi *et al.* (2012)), reposant à la fois sur des informations statistiques (probabilité et distribution) et sur la théorie de l'information (en particulier, sur l'entropie de Shanon (1948)). Les candidats issus des deux méthodes sont agrégés dans une liste finale grâce à un système de vote pondéré (la méthode de Borda (1781)). Appliquée au chinois (Zou *et al.*, 2006) et au persan (Sadeghi & Vegas, 2014), cette technique a permis de générer des listes de qualité acceptable.

Il est à noter toutefois que la plupart des méthodes citées plus haut fournissent des listes de *stopwords* dont certains éléments sont encore trop dépendants des corpus. Afin de palier ce biais, plusieurs publications ont proposé de faire l'union de plusieurs listes de *stopwords* (issues de différents corpus ou obtenus avec différentes méthodes) et constaté que cela avait un effet positif sur leur évaluation (Lo *et al.*, 2005 ; Sadeghi & Vegas, 2014).
Il aurait pu être utile d'identifier également les *stopwords* dans nos corpus afin d'en faire une caractéristique supplémentaire pour l'entraînement des modèles, ou pour, à l'instar de Petrović *et al.* (2009), les exclure de nos calculs de mesures d'association. Toutefois, nous avons noté deux inconvénients à l'utilisation de listes de *stopwords* pour notre application :

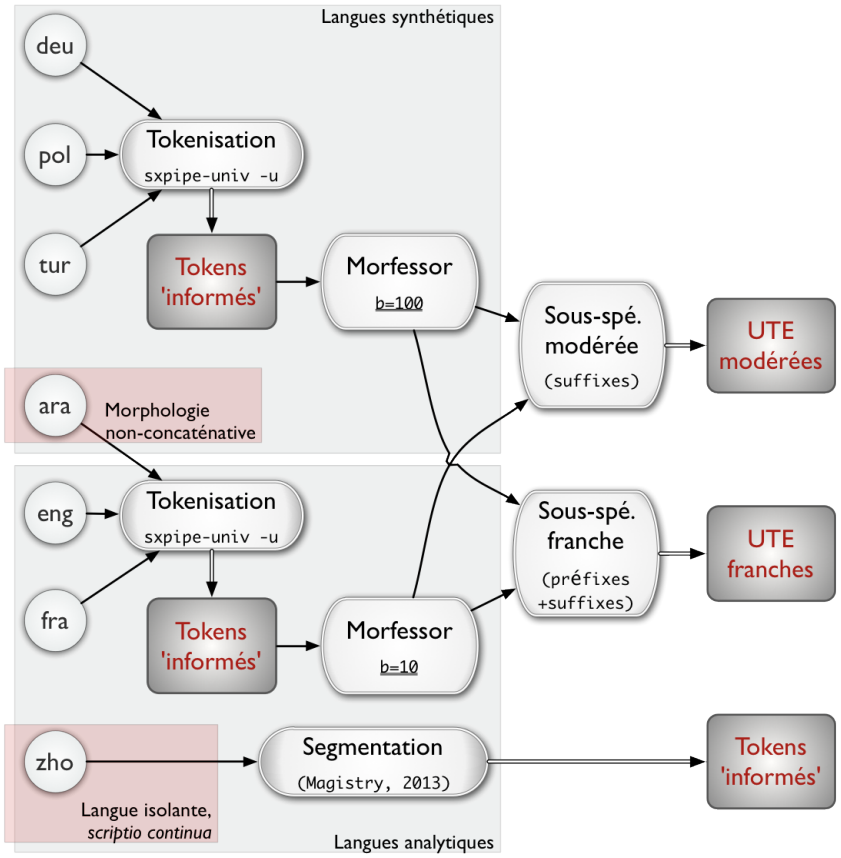
— Tout d'abord, même avec une liste subjectivement irréprochable, il arrive que certaines UTE fréquentes pouvant être qualifiées de *stopwords* soient également des items lexicaux importants que l'on ne souhaite pas supprimer. Par exemple, en anglais, le mot *it* sera probablement identifié comme *stopword*, alors qu'un autre de ses usage, spécifique et de plus en plus fréquent, désigne par un acronyme les technologies de l'information (*IT Systems*, *IT Trainings*, etc.). Ainsi, même une liste bien construite peut avoir des effets de bord indésirables. Ce genre de problème est d'autant plus gênant lorsque dans un corpus, les UTE sont fortement sous-spécifiées. Certaines UTE peuvent alors se confondre avec des mots grammaticaux dont elles prennent la forme orthographique.
— Théoriquement, étant donné les caractéristiques qui seront utilisées pour l'entraînement des modèles d'extraction (voir section 6.1.2), le système peut déterminer par lui même quels sont les *stopwords* de façon implicite, tout simplement en les excluant du modèle au besoin.

C'est principalement pour ces raisons que nous n'avons pas jugé nécessaire de développer des listes de *stopwords* à fournir au système d'extraction terminologique. En revanche, nous aurons besoin d'avoir ce genre de liste pour la phase d'évaluation. Selon Sadeghi & Vegas (2014) une liste d'environ 10 ou 30 *stopwords* choisis selon leur fréquence pourrait couvrir entre 20 et 30% des tokens de documents, aussi bien en anglais qu'en persan, et très probablement dans la grande majorité des langues car il s'agit là d'un corollaire de la distribution de Zipf-Medelbrot. C'est cette solution qui sera privilégiée pour les traitements décrits à la section 7.2.

5.4 Récapitulatif

La figure 5.2 résume en un schéma les choix qui ont été fait pour les pré-traitements textuels en ce qui concerne les sept langues sur lesquelles nous allons tester notre système.

FIGURE 5.2 – Schéma récapitulatif des pré-traitements pour sept langues.



Concernant la segmentation, les langues indo-européenes, l'arabe et le turc ont été naïvement

tokenisées et segmentées en phrases avec l'outil SxPipe, dont les étiquettes de sortie spécifiques ont été éliminées. Pour le chinois, une tokenisation naïve consiste à séparer l'ensemble des caractères par des espaces. Il faut donc appliquer un algorithme de segmentation afin d'obtenir des tokens comparables à ceux des autres langues de notre échantillon. Nous avons utilisé l'algorithme de segmentation non-supervisé développé par Magistry (2013) et Magistry & Sagot (2012).

Pour les langues à morphologie non triviale, une étape de découverte d'unités sous-spécifiées suit la tokenisation. Suite à une segmentation morphologique avec l'outil Morfessor, deux stratégies ont été envisagées :

— une stratégie modérée, qui consiste à déterminer à l'aide d'une typologie, si les affixes les plus productifs sont plutôt suffixaux ou préfixaux, et à éliminer les uns ou les autres en fonction ;
— une stratégie franche, qui supprime l'ensemble des affixes identifiées par l'analyseur morphologique.

L'arabe étant une langue à morphologie non-concaténarive, il est à noter que l'étape de sous-spécification basée sur un analyseur ne gérant pas ce genre de morphologie ne rendra probablement pas des éléments optimalement sous-spécifiés. Toutefois, le fait que l'arabe utilise des affixes nous permettra malgré tout de réduire la dispersion de données. De la même manière, en allemand, le fait que les mots composés aient été simplifiés plutôt que segmentés diminue la portée de la sous-spécification des unités complexes.

À la suite des pré-traitements textuels, on peut obtenir différents types d'unités de traitement sur lesquels seront basés les calculs de caractéristiques pour l'entraînement des modèles de l'extraction terminologique :

— des tokens informés,
— des UTE modérées,
— des UTE franches.

L'exemple récapitulatif suivant présente les effets des différents pré-traitements textuels sur une phrase en français :

— Phrase originale : Le management ne doit pas être focalisé uniquement sur le client mais aussi sur le bien être du salarié tant dans le déroulement de sa vie professionnel (en tenant compte de leurs avis, opinion lors des réunions,...) que dans sa vie privée.
— Après tokenisation : Le management ne doit pas être focalisé uniquement sur le client mais aussi sur le bien être du salarié tant dans le déroulement de sa vie professionnel (en tenant compte de leurs avis , opinion lors des réunions , ...) que dans sa vie privée .

— Après sous-spécification modérée : le manage ne doit pas être focal unique sur le client mais aussi sur le bien être du salarié tant dans le déroule de sa vie profession (en tenant compte de leur avis, opinion lors des réunion , ...) que dans la vie privé .

— Après sous-spécification franche : le manage ne doit pas être focal ique sur le client mais aussi sur le bien être du salarié tant dans le roule de sa vie fession (en tenant compte de leur avis , opinion lors des union , ...) que dans la vie privé .

Le chapitre 6 comparera les performances de différents modèles entraînés à partir des corpus en fonction des unités de traitement utilisés. Des correspondances univoques étant conservées entre tokens informés et les différents types d'unités de traitement élémentaire, de même que leur ordre linéaire d'apparition dans les corpus, il sera aisé de re-convertir les différentes unités sous spécifiées sous leur forme originale.

Chapitre 6

Extraction de termes

Sommaire

Nous avons vu section 2.1.2 les trois grands types d'approches pour l'identification de candidats termes : linguistiques, statistiques et hybrides. Étant donné la diversité et la qualité de nos données, le côté « discipline de terrain » lié à tous les paramètres imposés par le fort multilinguisme de notre travail, nous avons fait le choix de ne pas faire usage des approches linguistiques, mais de circonscrire notre recherche à des méthodes d'extraction terminologique statistiques et probabilistes, et plus particulièrement aux méthodes d'apprentissage automatique. Ces dernières permettent tout à la fois d'envisager de nombreux paramètres lors de l'apprentissage d'un modèle et de s'adapter facilement à de nouvelles données, et ce sur de gros volumes textuels.

Notre but ici n'est pas de comparer différents algorithmes pour trouver celui qui donne des résultats optimaux, les raisons de se disperser étant déjà trop nombreuses. Par ailleurs, les précédentes recherches ont déjà témoigné de la grande variabilité des résultats pour un même algorithme ou une même mesure sur des données différentes.

C'est pourquoi nous avons choisi un algorithme d'apprentissage automatique particulier, celui des Champs Markoviens Conditionnels (Lafferty *et al.*, 2001) déjà beaucoup utilisé en traitement automatique des langues et réputé donner de bons résultats pour de multiples tâches dans ce domaine, comme le démontrent de nombreuses publications (McCallum & Li, 2003 ; Sha & Pereira, 2003 ; Tsuruoka *et al.*, 2009 ; Tellier *et al.*, 2010 ; Constant *et al.*, 2011 ; Torii *et al.*, 2011).

6.1 Apprentissage avec des Champs Markoviens Conditionnels

Les Champs Markoviens Conditionnels, ci-après CRF (correspondant à l'acronyme anglais *Conditional Random Fields*) ont été proposés par Lafferty *et al.* (2001) pour construire, de façon supervisée, des modèles probabilistes capables de segmenter et d'étiqueter des données séquentielles.

Comme pour tous les algorithmes supervisés, l'utilisation de CRF fonctionne en deux temps. La première étape consiste à apprendre un modèle de classification en se reposant sur une large quantité de données d'entraînement étiquetées. La seconde étape consiste à classifier des données inconnues, mais de même nature que les données d'entraînement, en se reposant sur le modèle appris lors de la première étape.

6.1.1 Principes

Les Champs Markoviens Conditionnels reposent sur les variables d'entrée (X), les variables à prédire en sortie (Y) et la modélisation du lien entre les deux. Pour ce faire, un modèle doit définir quelle représentation donner aux X et Y en approximant une distribution

conditionnelle s'approchant le plus possible de la distribution conditionnelle réelle des données $p_{relle}(Y|X)$. L'espace hypothèses Φ est l'ensemble des fonctions indexées par des paramètres θ dont on suppose qu'ils sont intéressants pour définir le modèle. À cet effet, il faut déterminer quels sont les traits les plus pertinents afin d'éviter le sur-apprentissage. S'intéresser à des hypothèses trop spécialisées peut mener à de mauvaises performances sur des données inconnues [1].

On souhaite prédire correctement une séquence de variables aléatoires $Y = y_0, y_1, ..., y_T$ (correspondant aux étiquettes) étant donnée une séquence d'observations $X = x_0, x_1, ...x_T$ donnée en entrée. Chaque observation x_i représente en réalité un mot (token ou autre unité de traitement élémentaire) sous la forme d'un vecteur de caractéristiques préalablement calculées (voir section 6.1.2). La figure 6.1 représente une séquence de quatre vecteurs de caractéristiques et leurs étiquettes au format tabulé. Il s'agit d'un exemple simplifié, d'autres traits seront utilisés pour les expérimentations.

FIGURE 6.1 – Exemple jouet de quatre vecteurs d'observations à trois caractéristiques, assortis à une étiquette (voir section 6.2.2) présentés sous la forme d'un texte tabulé. Une observation par ligne. Les fréquences sont discrétisées (voir section 6.1.3.2)

$x = \{$ token,	ponct.,	fréq. spécialité,	fréq. générique $\}$	$y =$ étiquette
hausse	non	2	3	B
de	non	10	10	I
salaire	non	3	1	L
.	finale	9	8	O

Les CRF modélisent la distribution des probabilités conditionnelles $p(y|x)$ afin de déterminer quelle est la meilleure étiquette $y^* = \arg\max_y p(y|x)$ pour chaque observation (Sutton & McCallum, 2010). On donne en entrée à l'algorithme un ensemble de k caractéristiques, prenant des valeurs discrètes déterminées à partir des données d'entraînement. Ces dernières contribueront à déterminer les caractéristiques de la distribution des données d'entraînement que l'on souhaite conserver dans le modèle. Pour ce faire, l'algorithme pondère des fonctions de caractéristiques, que l'on décide souvent de limiter à l'un des deux types suivants (Moens, 2006) :

— Locales. Par exemple :

$$f_{loc}(y_i, X, i) = \begin{cases} 1 & \text{si l'observation à la position } i \text{ n'est pas de la ponctuation} \\ 0 & \text{sinon.} \end{cases}$$

1. Ce sur-apprentissage peut notamment être provoqué par un trop grand nombre (par exemple plusieurs centaines de milliers) de caractéristiques données en entrée du CRF.

— Transitionnelles (bigrammes). Par exemple :

$$f_{trans}(y_{i-1}, y_i, X, i) = \begin{cases} 1 & \text{si } y_{i-1} = \mathsf{B} \text{ et } y_i = \mathsf{I} \\ 0 & \text{sinon.} \end{cases}$$

La probabilité d'une séquence d'étiquettes $Y = y_0, y_1, ..., y_T$ sachant une séquence d'observations $X = x_0, x_1, ...x_T$ est définie comme étant le produit normalisé de fonctions potentielles de la forme :

$$\exp\left(\sum_j \lambda_j f_{trans}(y_{i-1}, y_i, X, i) + \sum_k \mu_k f_{loc}(y_i, X, i)\right)$$

où λ_j et μ_k sont les paramètres $\theta = (\lambda_1, \lambda_1, ...; \mu_1, \mu_2, ...)$ estimés pendant l'entraînement. C'est d'ailleurs parce que ces facteurs pondèrent les fonctions caractéristiques que ces dernières peuvent simplement prendre des valeurs binaires.

Il existe plusieurs approches pour trouver l'ajustement idéal des paramètres θ, comparables à celles des problèmes de régression logistique multiclasse [2] (Lafferty *et al.*, 2001 ; Lavergne *et al.*, 2010). En unifiant les notations $f_{loc}(y_i, X, i)$ et $f_{trans}(y_{i-1}, y_i, X, i)$ sous la désignation générique $f(y_{i-1}, y_i, X, i)$, et en regroupant λ et μ sous la désignation générique θ, la distribution des probabilités conditionnelles définie par un CRF est :

$$p(Y|X, \theta) = \frac{1}{Z} \exp\left(\sum_{j=1}^{K} \theta_j \sum_{i=1}^{n} f_j(y_{i-1}, y_i, X, i)\right)$$

où :

— K est le nombre de fonctions caractéristiques,

— Z est un facteur de normalisation dépendant des observations :

$$Z = \sum_{y \in Y} \prod_{i=1}^{n} \exp\left(\sum_{j=1}^{K} \theta_j f_j(y_{i-1}, y_i, X, i)\right)$$

Comme le fait remarquer Grouin (2013), de nombreuses implémentations de CRF sont librement disponibles et utilisables : Mallet (McCallum, 2002), CRF++ (Kudo *et al.*, 2004), MIST (Aberdeen *et al.*, 2010) ou encore Wapiti (Lavergne *et al.*, 2010).

Parmi ces dernières, nous avons choisi, à l'instar de Grouin (2013), d'utiliser Wapiti (Lavergne *et al.*, 2010) pour des raisons similaires, relevant de la simplicité d'utilisation, de la rapidité d'exécution, mais aussi du large choix pour le paramétrage de la méthode d'optimisation implémentée et d'autres paramètres.

2. Parmi elles : IIS (Improved Iterative Scaling), L-BFGS (Limited-memory BFGS (Broyden–Fletcher–Goldfarb–Shanno)), OWL-QN (Orthant-Wise quasi-Newton), SGD (Stochastic Gradient Descent), BCD (Block Coordinate Descent), RPROP (Resilient Propagation), etc.

6.1.2 Calcul des caractéristiques

Nous avons considéré empiriquement un ensemble de 24 caractéristiques indépendantes de la langue calculables sur des n-grammes, $n \leq 3$:

— une caractéristique relative à la ponctuation pouvant prendre huit valeurs, correspondant aux sept catégories Unicode assignées à la ponctuation [3], et une valeur indiquant un token non ponctué. Cette caractéristique est applicable aux unigrammes uniquement ;

— trois caractéristiques, applicable aux 1 à n-grammes ($n = 3$), relatives aux fréquences dans les corpus généralistes et/ou de spécialité : la fréquence dans le corpus de spécialité (abréviée Sfreq dans les tableaux de résultats), un trait associant par concaténation les fréquences du corpus de spécialité et du corpus généraliste (abrévié Dfreq) ainsi que le résultat d'un t-test de Welch (1947) pour la comparaison des fréquences du corpus de spécialité et du corpus généraliste (abrévié ttest [4]) ;

— 19 mesures d'association applicables aux n-grammes de toutes tailles, sauf unigrammes. Il existe différents types d'approches pour mesurer la force du lien qui unit une paire d'éléments à partir des informations de fréquences issues d'un texte score d'association (cf. section 2.1.2.2). Pecina & Schlesinger (2006) ont présenté un résumé présentant une vue d'ensemble de 82 mesures d'association généralement appliquées à des bigrammes pour l'extraction de collocations. Celles que nous avons implémentées sont présentées en table 6.1. Parmi ces dernières, une est probabiliste (probabilité conditionnelle), deux sont des tests d'indépendance statistique (t-test et z-score), trois sont plus ou moins apparentées à des mesures d'information mutuelle (information mutuelle, dépendance mutuelle et saillance). Les 13 autres sont des mesures diverses, parfois heuristiques. Evert (2005, 2007) indique que plusieurs paramètres rendent complexe l'interprétation des scores en distinguant entre plusieurs types génériques [5] des mesures d'association :

  - *Unilatérales* ou *bilatérales* : les mesures unilatérales (*one-sided*) dissocient les associations positives (apparaissant plus souvent que s'il y avait indépendance) et négatives (apparaissant moins souvent). Pour ces dernières, un score élevé indique qu'il y a une forte association positive ; en revanche, un score faible ne peut être interprété que comme l'absence de preuve d'une association positive. Par exemple, les mesures ZS, TT, USUB, FAG ou ODR (entre autres) du tableau 6.1 sont uni-

3. Détails sur le site officiel du Consortium Unicode (http://www.unicode.org/reports/tr44/tr44-4.html cf. Table 10. General_Category Values.

4. Ce trait est à ne pas confondre avec le trait abrévié TT correspondant à un t-test utilisé comme mesure d'association, comme défini dans le tableau 6.1.

5. D'après Evert (2007), certaines mesures d'association heuristiques sont difficiles à classer selon ces critères.

latérales. À l'opposé, les mesures bilatérales (*two-sided*) assignent un score élevé à tous les éléments fortement associés, que ce soit de façon positive ou négative. Des scores faibles sont un indice en faveur de l'indépendance entre les éléments. Ces mesures sont généralement dérivées de tests statistiques également bilatéraux.

- Fondées sur *l'ampleur de l'effet* ou sur *l'importance* : les premières utilisent les estimations directes, ne prenant pas en considération les variations de taille entre les échantillons. En conséquence, les résultats ne sont pas fiables pour mesurer l'association des paires ayant une fréquence faible, et sont au contraire biaisés en faveur des paires ayant une haute fréquence. Cela est le cas par exemple pour les mesures comme ODR, SIM ou JAC. Les secondes mesurent la quantité théorique de preuves allant à l'encontre de l'indépendance entre les éléments d'une paire (hypothèse nulle) ; il est à noter que l'hypothèse nulle n'étant pas réaliste (les mots n'étant pas combinés par hasard), des scores d'association élevés peuvent être obtenus y compris pour une déviation minime par rapport à l'hypothèse nulle. Les mesures de ce genre sont par exemple ZS et TT.

Quelle que soit leur place dans cette classification, ces mesures d'association sont initialement définies pour des bigrammes. Nous avons donc utilisé un patron d'extension proposé par Tadić & Šojat (2003), cité par Petrović *et al.* (2009) (voir section 2.1.2.2, paragraphe relatif aux mesures liées à l'*unithood*). Celui-ci calcule récursivement la mesure d'association g sur un n-gramme en calculant la moyenne entre le résultat de cette mesure sur les $(n-1)$ premiers mots d'une part et sur les $(n-1)$ derniers mots d'autre part :

$$G(g, w_1 \cdots w_n) = \frac{g(w_1, w_2 \cdots w_n) + g(w_1 \cdots w_{n-1}, w_n)}{2}.$$

Petrović *et al.* (2009) ont montré que cette mesure obtenait, selon leur méthodologie, des résultats raisonnables sur les quadrigrammes en anglais. Elle présente par ailleurs les avantages d'être à la fois facilement généralisable à toutes les mesures d'association et de posséder un coût computationnel moindre par rapport aux autres patrons d'extension proposés.

Par ailleurs, étant donné que certains termes auront tendance à apparaître aux frontières de phrases (plus particulièrement en position initiale), nous avons ajouté des marqueurs de début et de fin de phrase dans les corpus. Ces derniers ne sont jamais comptabilisés dans les données quantitatives données sur les corpus dans ce travail, mais sont utilisés dans tous les calculs de caractéristiques.

6.1.3 Pré-traitement des caractéristiques numériques

Les pré-traitements numériques consistent à modifier les données afin de les rendre plus simples à exploiter par la suite. Il existe différents pré-traitements, pouvant servir à normaliser, lisser, filtrer ou encore classer des données. Comme le fait remarquer Niemann (1990, p. 26), il est en général très difficile de juger du succès des pré-traitements selon des critères objectifs. Une méthode efficace, mais coûteuse, pourrait consister à évaluer les performances des algorithmes en aval avec ou sans pré-traitements pour déterminer quelle option mène aux meilleurs résultats. Les coûts computationnels liés à ce type d'évaluation seraient néanmoins élevés. C'est pourquoi nous nous contenterons de juger subjectivement les différentes options de pré-traitement qui s'offrent à nous à l'aide de courbes de densité.

Les courbes de densité présentées en annexe à la figure A.1 (p. 244), calculées à partir des corpus de spécialité tokenisés, témoignent de l'hétérogénéité des données numériques à traiter, aussi bien quant au domaine de définition des différents scores qu'à l'allure de leurs courbes de densité. Six de ces courbes, représentant six mesures d'association présentées dans le tableau 6.1 (ZS, ODR, FAG, MD, CP et USUB) sont envisagées au sein d'une même langue, ici l'arabe. Elles ont été choisies pour leur diversité : toutes ont des distributions logarithmiques, la plupart sont asymétriques, certaines sont multimodales. Les deux dernières courbes représentent deux mesures d'association (CP et USUB) pour l'allemand. Elles permettent de comparer les distributions de ces deux mesures d'association entre l'allemand et l'arabe. Outre le fait que la langue diffère, les tailles de corpus diffèrent significativement.Il convient donc de transformer ces scores afin de les rendre comparables et utilisables en aval de la chaîne de traitement.

6.1.3.1 Normalisation

La normalisation des données consiste à ramener l'ensemble des valeurs numériques d'une variable donnée dans un même domaine, généralement petit (par exemple l'intervalle $[0, 1]$). Ainsi, des différences énormes entre les maximums et les minimums d'une mesure d'association (par exemple ODR, allant de 0 à $2,5.10^6$ pour l'arabe) sont ramenées à des intervalles de valeurs plus modestes. Les fonctions de normalisation doivent répondre aussi bien à des critères de robustesse (insensibilité aux valeurs extrêmes) qu'à des critères d'efficacité (proximité entre l'estimation obtenue et l'estimation optimale de la distribution des données). Or cet équilibre est difficile à trouver, d'autant plus lorsque les données à normaliser ne suivent pas une distribution gaussienne (Jain *et al.*, 2005).

Il existe de nombreuses techniques de normalisation. Les plus connues sont le *z-score*, la normalisation *min-max*, ou la normalisation par *mise à l'échelle décimale* (*decimal scaling*) (Han & Kamber, 2006 ; Saranya & Manikandan, 2013). Il existe d'autres variations comme, entre

autres, la normalisation par fonction double sigmoïde, une fonction combinant médiane et écart absolu médian ou la normalisation tanh (Jain *et al.*, 2005 ; Naït-ali & Fournier, 2012). Nous nous focaliserons sur les trois méthodes les plus populaires.

L'application du z-score pour la normalisation est basée sur la moyenne et l'écart type de la variable. Cette méthode est sensible aux valeurs extrêmes, surtout si elles sont très nombreuses. La distribution des données initiale n'est conservée que si elle est gaussienne. Dans le cas contraire, la moyenne et l'écart type ne sont pas des estimateurs optimaux.
Sachant la moyenne arithmétique μ et l'écart type σ d'une mesure d'association pour des n-grammes de taille donnée, la normalisation par z-score est effectuée par la fonction :

$$f_{z-score}(x) = \frac{x - \mu}{\sigma}$$

La figure A.1 reprend les courbes de densité de 6 mesures d'association (ZS, CP, FAG, MD, ODR, USUB) présentées en annexe A. La figure A.2 présente les courbes de densité des mesures d'association présentées figure A.1 après normalisation par z-score. On constate que les scores normalisés ont des domaines de définition variés et parfois peu comparables. Qui plus est, la distribution n'est pas conservée.

La normalisation min-max consiste à appliquer une transformation linéaire aux données originales, dont les extrémums sont min et max, et dont on souhaite désormais ramener les valeurs sur l'intervalle $[b_{inf}, b_{sup}]$, avec une fonction de type :

$$f_{min-max}(x) = \frac{x - min}{max - min}(b_{sup} - b_{inf}) + b_{inf}$$

Elle est appropriée dans les cas où les limites minimum et maximum des valeurs à normaliser sont connues et peuvent être conservées. Cette normalisation préserve, en théorie, la distribution originale des scores, à l'exception d'un facteur d'échelle (Jain *et al.*, 2005). Dans la figure A.3, nous avons utilisé les extrémums locaux pour chaque mesure d'association, chaque longueur de n-gramme pour une langue donnée. On constate que la normalisation ainsi obtenue manque de consistance. Or, pour bien faire, quels extremums conserver ? Faut-il les rechercher, pour une mesure d'association, au sein de chaque longueur de n-gramme ou pour toutes les longueurs ? Faut-il prendre les extremums au sein d'une même langue ou à travers les langues ? Que faire si la taille des corpus varie ? L'impact de ces facteurs sur la robustesse de la normalisation par min-max rend son utilisation impossible sur nos données.

La normalisation par mise à l'échelle décimale (*Decimal Scaling*) peut être appliquée notamment lorsque l'ensemble des scores à normaliser est situable sur une échelle logarithmique

(Jain *et al.*, 2005), ce qui est le cas de nos données. Cette méthode de normalisation conserve les unités de base en leur appliquant une transformation logarithmique. Sachant le maximum abs_max des valeurs absolues des extremums d'un score d'association pour une longueur de n-gramme donnée dans une langue, pour un corpus, la normalisation consiste à appliquer la fonction $f_{decimal_Scaling}$ à toutes les valeurs de la série :

$$f_{decimal\_scaling}(x) = \frac{x}{10^{\log_{10}(abs\_max)}}$$

Bien que cette approche soit réputée manquer de robustesse, c'est elle qui normalise le plus fidèlement les mesures d'association calculées. L'inconvénient de cette technique réside dans le domaine de définition des valeurs normalisées, qui peuvent aller jusqu'aux environs de 0 et -1 et ne pas occuper tout cet intervalle. C'est néanmoins la méthode que nous avons sélectionnée pour la normalisation de nos données.

6.1.3.2 Discrétisation

Les Champs Markoviens Conditionnels nécessitent en entrée des caractéristiques discrètes. Il faut donc, pour chaque variable continue correspondant à une caractéristique, trouver un ensemble de points de découpages cohérents qui minimise la perte d'information tout en réduisant au maximum le nombre de découpages (Kotsiantis & Kanellopoulos, 2006). S'il est bien mené, ce découpage d'une variable continue en un nombre fini de partitions permet d'accélérer l'apprentissage automatique, mais également d'éviter le phénomène de surapprentissage en réduisant l'espace d'hypothèses et de produire ainsi de meilleurs modèles.
Il existe un grand nombre d'approches pour la discrétisation de données : supervisées ou non supervisées, univariées ou multivariées, paramétriques ou non-paramétriques, hiérarchique ou non-hiérarchique, locales ou globales, « avides » (*eager*) ou « paresseuses » (*lazy*), dynamiques ou statiques (Dougherty *et al.*, 1995 ; Kotsiantis & Kanellopoulos, 2006 ; Yang, 2003).
Les techniques de discrétisation supervisées nécessitent de disposer des étiquettes *gold* utilisées pour la classification afin de sélectionner les meilleurs points de découpages. Or, une fois nos modèles entraînés avec leurs caractéristiques discrétisées, il faudra appliquer ce même pré-traitement aux caractéristiques calculées pour des données pour lesquelles nous ne disposons d'aucune étiquette *gold*, afin de pouvoir leur appliquer un modèle. C'est pourquoi nous n'avons pas retenu ces approches. Yang (2003) propose une revue de différentes techniques, partiellement reproduite dans la tableau 6.2 pour ne conserver que les méthodes non-supervisées. Il n'y en a pas une qui soit immuablement supérieure aux autres car l'efficacité de la discrétisation varie significativement en fonction de la distribution de la variable considérée, notamment si cette dernière est fortement asymétrique ou contient des pics (Ismail, 2003).

TABLE 6.2 – Méthodes de discrétisation non-supervisées (Yang, 2003, p. 90). (Abréviations : Uni.=univarié, Mul.=multivarié, P=paramétrique, Np=non-paramétrique, H=hiérarchique, Nh=non-hiérarchique, Glo.=global, Loc.=local, avi.=avide, Par=paresseuse, D=disjoint, Nd=non-disjoint)

Méthode	Uni.	Mul.	P	Np	H	Nh	Glo.	Loc.	Avi.	Par.	D.	Nd.
EW [6]	✓		✓			✓	✓		✓		✓	
EF [7]	✓		✓			✓	✓		✓		✓	
k-Means [8]	✓			✓	✓		✓		✓		✓	
Dyn. qualit. [9]	✓			✓	✓			✓		✓		✓
Rel. unsup. [10]		✓	✓		✓		✓		✓		✓	
Multi. [11]		✓		✓	✓		✓		✓		✓	

Afin d'élaguer les méthodes qui ne conviendraient pas à nos expériences, nous allons passer en revue les modalités des différentes approches détaillées par Yang (2003).

— Le choix entre une approche multivariée ou univariée dépend de l'envergure des données que l'on souhaite considérer. Les approches univariées ne prennent en compte qu'une variable à la fois. À l'inverse, les démarches multivariées vont considérer un ensemble de variables afin de pouvoir y détecter des motifs récurrents et discrétiser en fonction. Étant donné que l'ensemble des caractéristiques qui seront utilisées pour entraîner le modèle pourra varier d'une expérience à l'autre, nous préférons ne pas utiliser d'approches multivariées.

— Les approches paramétriques nécessitent de renseigner des paramètres tels que le nombre maximal d'intervalles. Les approches non-paramétriques déterminent elles-mêmes leurs paramètres.

— Concernant les approches hiérarchiques, elles vont sélectionner incrémentalement (avec des procédures de découpage et/ou de regroupement) les intervalles à découper en formant une hiérarchie implicite. Les approches non-hiérarchiques ne passent en revue les données qu'une fois pour y appliquer le processus de découpe.

— La différence entre approches globales et approches locales réside dans le traitement d'une même caractéristique dans différents contextes d'entraînement. Les méthodes globales vont lui assigner le même ensemble d'intervalles, alors que les approches lo-

3. EW (pour *Equal Width*) : discrétisation à intervalles d'amplitude égales.
4. EF (pour *Equal Frequency*) : discrétisation à intervalles de fréquences égales.
5. k-Means : discrétisation par clustering des k-moyennes (Torgo & Gama, 1997), cité par Yang (2003, p. 65).
6. Dyn. qualit. (pour *dynamic qualitative*) López *et al.* (2000), cité par Yang (2003, p. 82).
7. Rel. unsup. (pour *relative unsupervised*) Lud & Widmer (2000) , cité par Yang (2003, p. 84).
8. Multi (pour *multivariate discretisation*) Bay (2000), cité par Yang (2003, p. 85).

cales vont discrétiser sans considération pour l'espace de données global. En ce qui nous concerne, nous ne favoriserons pas une approche sur une autre suivant leur classification dans les groupes paramétrique/non-paramétrique, hiérarchique/non-hiérarchique et global/local.

— En revanche, étant donné la dimensionnalité des données à discrétiser pour chaque langue, nous préférons favoriser les approches « avides », qui s'appliquent comme un pré-traitement, plutôt que des approches « paresseuses », qui ont lieu à la volée lors de la phase de classification.

— Enfin, ce qui sépare les approches disjointes des méthodes non-disjointes est la possibilité, comme leur nom l'indique, d'obtenir des intervalles disjoints. Dans le cadre de méthodes non-disjointes, les intervalles peuvent se chevaucher. Les méthodes disjointes paraissent mieux adaptées à l'apprentissage avec des CRF.

Étant données nos conditions préalables, ne restent du tableau 6.2 que les méthodes EW, EF et k-means. Nous allons déterminer, parmi elles, quelle méthode est la mieux adaptée à nos données.

— La plus simple est la *discrétisation à intervalles d'amplitude égales* (*Equal Width* ou EW). Cette dernière consiste à diviser l'étendue de la variable à discrétiser en k classes contenant le même nombre de valeurs. Cette opération a une complexité algorithmique de $\mathcal{O}(n \log_2 n)$ pour une distribution comportant n valeurs Fang *et al.* (2013). Cette méthode est particulièrement adaptée dans le cas où les observations sont distribuées uniformément, ce qui n'est pas le cas dans nos données.

— La seconde méthode est la *discrétisation à effectifs égaux* (*Equal Frequency* ou EF), qui divise la variable continue en k intervalles où, sachant m instances, chaque intervalle contient $\frac{m}{k}$ valeurs adjacentes, éventuellement dupliquées. Les blocs issus des découpages peuvent donc avoir des tailles différentes. Sa complexité algorithmique est la même que celle de l'approche à intervalles d'amplitude égales (Fang *et al.*, 2013).

— Enfin, la méthode de discrétisation par clustering des k-moyennes (*K-Means Clustering Discretisation* ou k-means) consiste à déterminer les intervalles à donner aux variables discrètes en utilisant l'algorithme des k-moyennes (Hartigan & Wong, 1979). Ce dernier cherche à minimiser le carré de la distance euclidienne entre les valeurs à regrouper dans un cluster et leur centroïde correspondant (dont les valeurs sont itérativement affinées). La complexité, dans le pire des cas, de cette approche est supérieure aux deux précédentes : $\mathcal{O}(n^{k+1} \log_2 n)$ pour k clusters. Il est possible de décider soi-même du nombre d'intervalles k voulus pour la discrétisation. Dans ce cas, la complexité varie en fonction ; par exemple, pour $k = 10$, la complexité de cette approche

sera $\mathcal{O}(n^{10+1} \log_2 n)$.

En ce qui concerne la sélection du nombre d'intervalles k (nécessaire dans les approches paramétriques), déterminer leur nombre optimal nécessiterait de mener des tests pour plusieurs valeurs de k dans chaque langue. Qui plus est, si les résultats des tests diffèrent, un modèle n'est plus portable d'une langue à une autre. Ahmad *et al.* (2012) indiquent que peu d'intervalles, même s'ils représentent moins bien les valeurs qu'ils contiennent, donnent de meilleurs résultats pour résoudre des problèmes de classification. Traditionnellement, lorsqu'on ne peut pas déterminer cette valeur avec des tests, la valeur $k = 10$ est utilisée. La figure 6.2 présente le tracé des complexités algorithmiques pour la plage de données n (entre le plus petit et le plus gros corpus, en terme de tokens, présenté figure 5.1), et en ce qui concerne l'algorithme *k-means*, différents k.

Le coût computationnel de la méthode k-means étant significativement supérieur à celui des méthodes par EW et EF, d'autant plus lorsque k augmente, cela risque de poser problème sur de gros volumes de données. Par ailleurs, on remarque que la différence de complexité entre les méthodes EW et EF et la méthode k-means ($k = 5$) est du même ordre de grandeur que l'écart entre les complexités de k-means avec $k = 5$ et $k = 10$. Diminuer la taille de k présente donc un double avantage, au moins en ce qui concerne l'algorithme k-means.

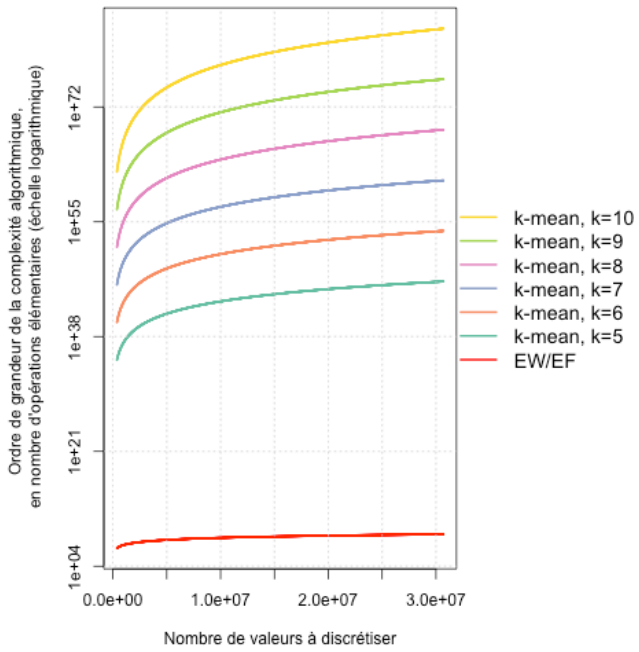
Afin de déterminer laquelle de ces approches est la meilleure étant donné la variété de nos données, nous avons utilisé la fonction discretize du paquet R *arules*[12] pour visualiser les découpages rendus par ces trois méthodes sur les mesures d'association ZS, ODR, FAG, MD, CP et USUB (en arabe) normalisées avec la méthode de mise à l'échelle décimale. Pour nos test, nous avons choisi d'utiliser le paramètre $k = 5$. Les figures A.5, A.6 et A.7 fournies dans l'annexe A (p.243) présentent respectivement les points de découpages proposés par les méthodes EW, EF et k-means. On constate que globalement, la méthode EW discrétise mal nos données, d'autant plus lorsque ces dernières sont fortement asymétriques et contiennent des pics. La discrétisation EF propose des partitions tenant mieux compte de l'asymétrie. Toutefois, le découpage des pics reste relativement peu équilibré. La meilleure méthode, au vu des figures A.7, est la méthode k-means, qui semble retrouver le plus clairement des « paliers de valeurs ». C'est donc celle que nous avons retenue.

6.2 Annotation des corpus

Afin que l'algorithme reposant sur les champs aléatoires conditionnels puisse entraîner un modèle, il faut lui donner en entrée un corpus annoté. Pour nos tests initiaux, les corpus génériques et de spécialités sont les corpus originaux, sans autre transformation qu'une seg-

12. http ://cran.r-project.org/web/packages/arules/index.html

FIGURE 6.2 – Comparaison des complexités algorithmiques des différentes approches de discrétisation en fonction du nombre de valeurs à discrétiser, et du paramètre k (l'échelle des ordonnées est logarithmique).



mentation en tokens informés (voir section 5.1).

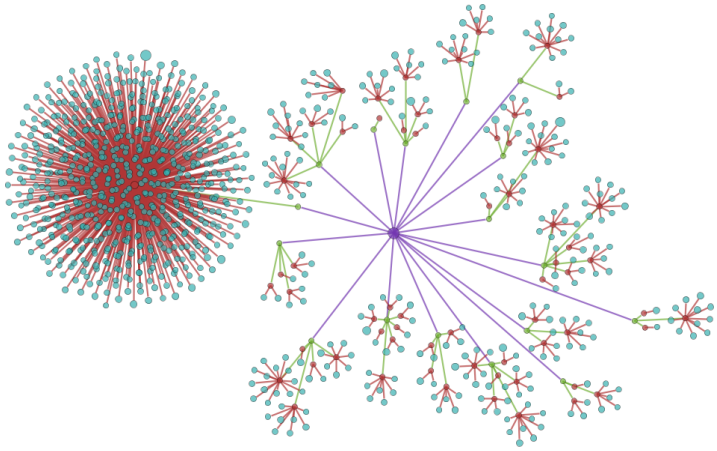
C'est à partir de cette version des corpus, ainsi que d'une terminologie multilingue *gold* (présentée à la section 6.2.1) que les annotations seront calculées. Par la suite, des tests seront menés avec les corpus génériques et de spécialités dont les tokens informés ont été substitués à leurs équivalents sous spécifiés (UTE modérées, UTE franches). L'ordre linéaire étant conservé, les annotations calculées pour les tests initiaux peuvent être simplement transposées à ces corpus.

6.2.1 Terminologie multilingue de référence (Ressources Humaines)

Nous disposions d'une terminologie multilingue structurée relative au traitement d'enquête dans le domaine de Ressources Humaines, dont le squelette est représenté sur la figure 6.3. Cette terminologie, développée manuellement par un expert du domaine au sein de l'entreprise *Verbatim Analysis – VERA*, est hiérarchiquement organisée en trois niveaux : 17 *thèmes*, une soixantaine de *super-classes* et plus de 800 *classes*. La partie gauche de cette hiérarchie semble bien équilibrée, alors que dans la partie droite, 646 *classes* sont regroupées sous une même *super-classe*. Cette dernière regroupe des classes génériques jugées pertinentes pour

le domaine, mais dont l'expert du domaine ne souhaite pas détailler la structure. Ce déséquilibre n'aura aucune influence sur la suite de nos traitements.

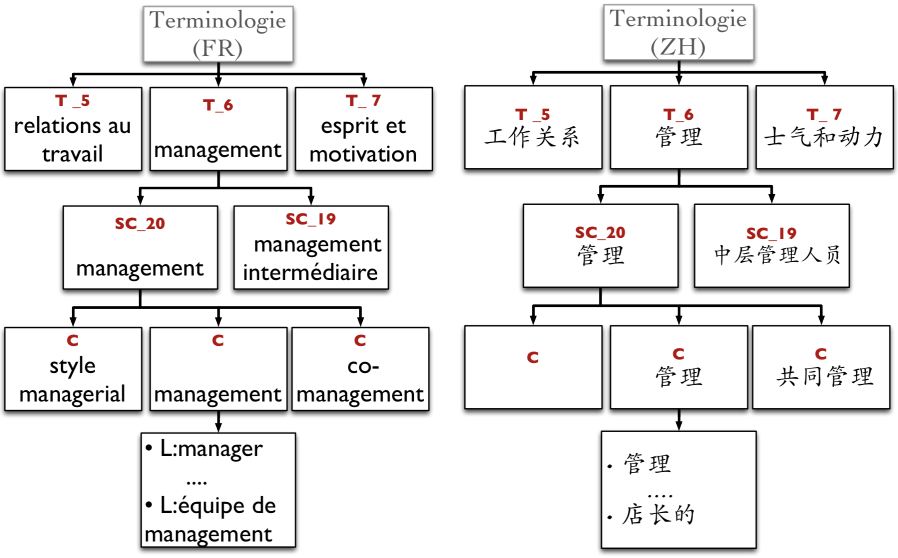
FIGURE 6.3 – Visualisation sous forme de graphe du squelette [13] de la terminologie multilingue structurée utilisée comme référence dans les expériences. Le nœud central (violet), correspond à la racine, les nœuds verts aux thèmes, les nœuds rouges aux super-classes, les nœuds bleus aux classes.



En principe, pour une langue donnée, les niveaux supérieurs de la hiérarchie (*thèmes* et *super-classes*) ne sont pas instanciés par des termes ; ils sont identifiés par un identifiant unique et possèdent des instances de « titre » dans différentes langues. Seul le niveau des *classes*, le plus spécifique, peut comporter des instances de termes dans n'importe quelle langue. Les titres d'un niveau hiérarchique supérieur (au plus un par langue) ont été choisis par l'expert du domaine pour leur représentativité parmi les termes instanciés dans au moins une des classes subsumées. La figure 6.4 présente un fragment de la terminologie multilingue de référence, en français (a) et en chinois simplifié (b). Il est à noter qu'une classe existant pour une langue peut ne pas être instanciée dans d'autres langues. C'est le cas ici avec la classe « style managerial », qui ne dispose d'aucune traduction en chinois dans la terminologie de référence.

11. Ce graphe ne comporte aucune instance de terme. Il représente uniquement les classes thématiques et leurs liens hiérarchiques. Plus une classe est générique, plus elle est proche de la racine de l'arborescence.

FIGURE 6.4 – Représentation partielle d'une portion de la terminologie multilingue présentant les titres de trois *thèmes*, deux *super-classes*, trois *classes* et deux instances. Les titres des classes correspondent à des traductions, ce qui n'est pas nécessairement le cas pour les instances de termes. La notation « L : » représentant des lemmes, sera expliquée plus bas.



(a) En français (b) En chinois simplifié

Cette terminologie multilingue compte actuellement dans sa hiérarchie des termes 36 langues et variétés de langues. Le tableau 6.3 présente le nombre de termes [14] par langues pour l'ensemble de la ressource de référence dans la dernière version sur laquelle nous avons travaillé (version de 2014). En postulant que la version française de cette terminologie soit complète, la troisième colonne du tableau 6.3 (« Avancement (%) ») propose une estimation du pourcentage d'avancement pour toutes les langues.

Elle a été construite par validation et complétion manuelles à partir, entre autres, des résultats d'un système simple d'extraction de candidats qui repose sur des métriques classiques (*t-test* sur la fréquence de n-grammes par rapport à un corpus de référence et mesures d'association). Les termes d'une langue ont été autant que possible traduits à l'aide d'outils de traduction automatique dans les autres langues, par des experts en Ressources Humaines ne parlant pas nécessairement les langues pour lesquelles ils ont développé ces terminologies. Cette ressource est donc incomplète dans certaines langues, et peut également comporter des entrées incorrectes. La figure 6.3 indique, en ce qui concerne nos langues de travail, que la ressource termi-

14. Ce nombre fait référence aux instances « non développées », c'est à dire qu'un lemme ou une expression régulière compte pour un seul terme.

11. Bosniaque-croate-monténégrin-serbe, parfois appelé serbo-croate.

nologique possède une excellente couverture pour l'anglais et le français. L'allemand est également très bien représenté. Sur ces trois langues en particulier, les experts ayant développé la ressource ont contrôlé la qualité des termes avec précision. Les termes dans les autres langues résultent en majorité de ce premier développement. La couverture de cette terminologie est également satisfaisante en ce qui concerne le polonais et le turc, mais beaucoup moins pour le chinois simplifié ou l'arabe. Pour ces derniers, le taux de couverture descend en dessous de 50% par rapport à la ressource en français. Plus particulièrement, les auteurs de la terminologie en arabe indiquent qu'il existe une marge d'amélioration considérable. Toutefois, son utilisation régulière durant plusieurs années pour une application industrielle à grande échelle indique que sa couverture est de qualité suffisante pour des applications réelles.

Cette terminologie ayant pour but d'automatiser et de systématiser autant que possible les tâches liées au traitement des verbatim, les termes recensés sont en réalité des mots, des lemmes (pour certaines langues européennes [16]), des expressions régulières, et des séquences de mots, lemmes et expressions régulières pertinents pour l'identification de concepts liés aux ressources humaines les plus abordés par les répondants aux enquêtes. Par exemple, dans la figure 6.4 (a), les mots précédés d'un « L : » sont des lemmes, dont les formes fléchies peuvent être générées en aval des traitements. Ains le terme « L :équipe de management » désigne indifféremment « équipe de management » ou « équipes de management ». Un exemple d'expression régulière, en allemand, est le terme « *wachstum* » (« *croissance* »). Ce dernier permet d'inclure de façon compacte des termes comme « *Vermögenswachstum* » (« la croissance des actifs ») ou « *Wachstumsraten* » (« taux de croissance ») à la terminologie.

Dès lors qu'un item présent dans la terminologie multilingue pour une langue donnée correspond à un ou plusieurs token(s) informé(s) dans le corpus de spécialité de cette même langue, ces tokens se voient assigner des étiquettes permettant de les repérer dans le corpus. La section suivante présente le jeu d'étiquettes choisi.

6.2.2 Jeu d'étiquette

Il existe de nombreuses manières d'indiquer la présence d'unités d'intérêt dans un corpus en fonction de leur nature et du lien qu'elles entretiennent (Sang & Veenstra, 1999).
Les jeux d'étiquettes les plus connus utilisent des lettres, comme dans l'exemple du tableau 6.4. Le plus simple consiste à indiquer de façon binaire si oui ou non l'item courant est remarquable. La convention veut que ces deux étiquettes soient désignées par les lettres *I* (pour

16. Dans ce cas, le système utilise des lexiques flexionnels conformes à l'architecture Alexina pour identifier toutes les formes fléchies : le *Lefff* pour le français et *EnLex* pour l'anglais (Sagot, 2010), *DeLex* pour l'allemand (Sagot, 2014) et *PolLex* pour le polonais (Sagot, 2009)) et le *Leffe* pour l'espagnol (Molinero *et al.*, 2009).

Inside) et O (pour *Outside*). Le principal inconvénient de ce jeu d'étiquette pour notre application réside dans le fait qu'il ne permet pas de distinguer deux termes successifs d'un terme complexe. Par exemple dans la figure 6.4, le trigramme « gestion de carrière » pourrait être interprété en trois termes (« gestion », « de » et « carrière »).

Un jeu d'étiquette plus fréquent, appelé IOB, permet d'identifier en plus des items étiquetés I et O ceux, étiquetés B (pour *Begin*), qui sont la borne initiale d'une unité d'interêt. De la même façon, une étiquette (E pour *End* ou L pour *Last*) peut être utilisée pour signaler l'item final d'une unité d'interêt.

L'étiquetage $BILOU$, plus expressif que le jeu d'étiquette *IOB*, rajoute l'étiquette U (pour *Unit*) afin de prendre en compte la spécificité des items isolés (Ratinov & Roth, 2009).

C'est cette distinction explicite entre termes simples (étiquetés U) et termes complexes (étiquetés par une séquence $BI * L$) qui nous a incité à utiliser ce jeu d'étiquettes plutôt qu'un autre.

TABLE 6.1 – Mesures d'associations utilisées comme caractéristiques pour l'entraînement des modèles. Extrait de Pecina & Schlesinger (2006)

Code	Mesure d'Association	Formule
CP	Probabilité conditionnelle	$P(y\|x)$
TT	T-test	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)\left(1-(f(xy)/N)\right)}}$
ZS	Z-score	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)\left(1-(\hat{f}(xy)/N)\right)}}$
PMI	Information mutuelle	$\log \frac{P(xy)}{P(x*)P(*y)}$
MD	dépendance mutuelle	$\log \frac{P(xy)^2}{P(x*)P(*y)}$
SAL	Salience	MD. $\log f(xy)$
RCT	R-cost	$\log\left(1+\frac{a}{a+b}\right).\log\left(1+\frac{a}{a+c}\right)$
USUB	Unigram subtuples	$\log\frac{ad}{bc}-3,29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
FAG	Fager	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2}\max(b,c)$
MTD	Mountford	$\frac{2a}{2bc+ab+ac}$
PRS	Pearson	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
FSS	Fourth Sokal-Sneath	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c}\right)$
KLO	Klosgen	$\sqrt{P(xy)}.\max\left[P(y\|x)-P(*y),P(x\|y)-P(x*)\right]$
DRK	Driver-Kroeber	$\frac{a}{\sqrt{(a+b)(a+c)}}$
ODR	Odds Ratio	$\frac{ad}{bc}$
JAC	Jaccard	$\frac{a}{a+b+c}$
FKZY	First Kulczynski	$\frac{a}{b+c}$
BB	Braun-Blanquet	$\frac{a}{\max(a+b,a+c)}$
SIM	Simpson	$\frac{a}{\min(a+b,a+c)}$

$f(x*)$	$a=f(xy)\quad b=f(x\bar{y})$	Table de contingence. N est le nombre total de bigrammes.
$f(\bar{x}*)$	$c=f(\bar{x}y)\quad d=f(\bar{x}\bar{y})$	$*$ désigne n'importe quel mot. $\bar{w}$ désigne n'importe quel mot autre que w.
N	$f(*y)\qquad f(*\bar{y})$	Les tests d'indépendance statistique utilisent les *fréquences attendues* $\hat{f}(xy)=\frac{f(x*)f(*y)}{N}$

TABLE 6.3 – Langues représentées dans la ressource terminologique de référence, avec leur nombre d'instances et le pourcentage de ce que cela représente par rapport à la langue la mieux documentée.

Langue	Nombre de termes	Avancement (%)
français	7589	100
anglais	6923	91,2
allemand	6014	79,2
polonais	5901	77,8
turc	5501	72,5
roumain	5134	67,7
russe	5132	67,6
espagnol	5107	67,3
italien	5095	67,1
portugais	4956	65,3
flamand	4828	63,6
suédois	4339	57,2
danois	3994	52,6
tchèque	3787	49,9
chinois simplifié	3771	49,7
grec	3427	45,2
japonais	3350	44,1
indonésien	3040	40,1
arabe	2963	39
slovaque	2858	37,7
chinois traditionnel	2482	32,7
hongrois	1643	21,6
hindi	1635	21,5
BCMS [15]	1546	20,4
croate	1546	20,4
bosnien	1544	20,3
bulgare	1493	19,7
ukrainien	1457	19,2
malais	1032	13,6
vietnamien	832	11
ourdou	787	10,4
norvégien	734	9,7
marathi	485	6,4
slovène	398	5,2
finnois	378	5
coréen	138	1,8

TABLE 6.4 – Séquences d'étiquettes assignées par différents jeux d'étiquettes pour une phrase d'exemple en français.

	La	gestion	de	carrière	et	la	mobilité	.
IO	O	I	I	I	O	O	I	O
IOB	O	B	I	I	O	O	B	O
BILOU	O	B	I	L	O	O	U	O

Chapitre 7

Protocole d'évaluation

Sommaire

Avant de pouvoir entraîner des modèles CRF à même d'extraire des termes potentiellement corrects dans une langue, il est important de sélectionner, parmi tous les modèles possibles, les plus performants sachant un ensemble de traits. Il faut donc disposer d'une méthodologie d'évaluation intégrable dans le protocole expérimental global, qui permette de comparer les performances des différents modèles. Or évaluer un système d'extraction terminologique est une tâche complexe, pour laquelle il n'existe pas de consensus : les tâches d'extraction terminologique sont aussi diverses que les moyens de les évaluer. Comme pour n'importe quelle autre tâche de traitement automatique des langues, il existe plusieurs stratégies (Vivaldi & Rodríguez, 2007 ; El Ayari, 2009 ; Grouin, 2013) :

— Globale ou transparente :
  - Une *évaluation globale* va, sachant une entrée, une sortie et des données de référence, établir un score d'évaluation, indifféremment de la manière dont les sorties sont obtenues. Cette stratégie est dite de type « boîte noire ».
  - Une *évaluation transparente* va, au contraire détailler tout ou partie des éléments de la chaîne de traitement pour l'évaluer. Cette méthode permet d'obtenir de vraies informations quant aux défauts et aux qualités du système.

— Directe ou indirecte :
  - Les méthodes d'*évaluation directes* vont attester de la qualité des composants en évaluant des propriétés intrinsèques, indépendante de l'usage qui leur est dévolu, en calculant par exemple des métriques comme la *précision* ou le *rappel* (voir section 7.2).
  - Au contraire, les méthodes d'*évaluation indirectes* vont consister à estimer la qualité d'un composant en regardant l'influence de sa sortie sur la performance d'autres systèmes (stratégie d'évaluation dite « *task-based* », ou « orientée-tâche »).

— Humaine ou automatique :
  - L'*évaluation humaine* consiste à examiner manuellement la qualité des résultats. Il s'agit d'un processus coûteux en temps, et il est fréquent que les avis divergent entre plusieurs évaluateurs humains sur ce qui est acceptable et ce qui ne l'est pas.
  - L'*évaluation automatique* consiste à comparer la sortie du système à une référence manuellement construite au préalable et dont l'acceptabilité fait consensus.

Étant donné le nombre important de tests à réaliser, la réalisation d'évaluations manuelles est exclue. De plus, les méthodes permettant de faciliter la comparaison à d'autres systèmes sont celles empruntant à la fois aux stratégies *globale*, *directe* et *automatique*. C'est donc dans ces optiques que notre évaluation sera construite.

Nous présenterons dans la section 7.1 le protocole expérimental mis en place pour procé-

der à la sélection des traits produisant les meilleurs modèles : la mise en place d'une validation croisée pour la production des modèles (section 7.1.1.1), les métriques utilisées pour le calcul des scores servant à l'évaluation de chaque modèle (section 7.2) et le protocole de sélection des meilleurs modèles (section 7.3).
Le chapitre 8 exposera les résultats ainsi obtenus suivant ce protocole expérimental pour chaque langue, et pour chaque type d'unités de traitement possible (token informé, UTE modérée et UTE franche).

7.1 Protocole expérimental

Étant donné le nombre de langues pour lesquelles il faut évaluer des modèles et le nombre de modèles possibles par langue sachant qu'un modèle peut combiner un ou plusieurs des 24 traits, nous avons pris le parti d'utiliser des corpus de taille modérée. Comme le tableau 7.1 l'indique, les corpus de l'arabe, du polonais, du turc et du chinois ont été conservés dans leur intégralité, fournissant entre 350 000 et 1, 1 million de tokens. En revanche, les corpus de l'allemand, de l'anglais et du français ont été re-dimensionnés à des tailles comparables, en sélectionnant aléatoirement 60000 phrases (ou segments [1] pour les corpus de spécialité) issues des corpus dont nous disposons réellement pour la langue, soit environ entre 1, 1 et 1, 5 million de tokens.

TABLE 7.1 – Taille des corpus pour l'évaluation des modèles.

Langue (code ISO)		lignes	tokens	caractères	Diff. entre spec. et gen. (%) ligne	token	caractère
ara	spec.	14381	345562	3562317	0,9	7,4	1,9
	gen.	14514	371119	3628526			
deu	spec.	60000	1065356	6847469	0	13,5	15,3
	gen.	60000	1209119	7892056			
eng	spec.	60000	1328028	7411048	0	4,9	4,1
	gen.	60000	1393305	7714641			
fra	spec.	60000	1488002	8707248	0	1	-0,6
	gen.	60000	1503021	8652290			
pol	spec.	29026	462160	3124833	5,7	-16,6	-18
	gen.	30666	385309	2562183			
tur	spec.	50597	764441	6202242	51,7	6,4	-5,4
	gen.	76764	813669	5869526			
zho	spec.	56153	1166760	6610601	-18,5	-8,4	-16,4
	gen.	45750	1068830	5528637			

1. Nous avions défini à la page 81 un segment comme une phrase ou, lorsque celle-ci est trop courte, comme une séquence de phrases.

Concernant le nombre d'exemples [2], le tableau 7.2 indique quelles classes sont minoritaires ou majoritaires, et dans quelles proportions elles apparaissent dans les corpus pour chaque langue.

TABLE 7.2 – Proportion des étiquettes pour les corpus d'entraînement pour l'évaluation.

Langues (ISO)	Proportion des classes (%)				
	B	*I*	*L*	*U*	*O*
ara	1,01	0,05	1,01	15,58	82,36
deu	0,65	0,07	0,65	13,17	85,46
eng	1,73	0,29	1,73	18,04	78,21
fra	1,96	0,66	1,96	17,4	78,01
pol	0,94	0,09	0,94	18,35	79,67
tur	1,04	0,05	1,04	16,01	81,86
zho	3,82	0,62	3,82	18,53	72,2
Moyenne	1,59	0,26	1,59	16,73	79,82

Dans un premier temps (section 7.1.1), nous allons traiter la question du découpage des différents corpus en échantillons d'entraînement, de développement et test, en abordant la question du déséquilibre entre les proportions des différentes étiquettes dans les corpus d'entraînement.
Les prédictions des différents modèles seront ensuite évaluées à l'aide d'un score terminologique, présenté à la section 7.2. Ces deux premières étapes permettront de comparer les performances de différents modèles. Finalement, la section 7.3 présentera le protocole de sélection des meilleurs modèles.

7.1.1 Organisation des corpus pour l'évaluation

La nature de l'algorithme d'apprentissage, en l'occurrence ici les CRF, impose de scinder les données de référence en trois parties pour l'entraînement d'un modèle et l'estimation de ses performances :

— une partie *apprentissage* (ci-après *train*), qui va entraîner un modèle,
— une partie *développement* (ci-après *dev.*), qui va permettre aux CRF d'ajuster les paramètres du modèle entraîné sur la partie *apprentissage*,

2. Un exemple correspond à une unité de traitement (token informé, UTE modérée ou franche selon l'expérience) à laquelle est associée un ensemble de traits et une étiquette de classe (B, I, L, U ou O).

— une partie *test* dont les données n'ont jamais été vues par l'algorithme d'apprentissage. Le modèle entraîné sur les parties *train*+*dev.* est appliqué à la portion de *test*. Les étiquettes alors obtenues sont comparées aux données de références pour l'obtention d'un score d'évaluation.

Le tableau 7.1 (p. 119) présentait les tailles des différents corpus. Certains corpus, comme celui de l'arabe, sont trop petits pour pouvoir à la fois entraîner un modèle et estimer correctement ses performances en une seule fois. La section 7.1.1.1 indique comment, pour surmonter ce problème de manque de données de référence, il est possible d'effectuer des ré-échantillonnages sur ces dernières, notamment par validation croisée.

La section suivante (7.1.1.2) évoquera la question de l'incidence du déséquilibre entre le nombre d'exemples positifs (étiquetés B, I, L ou U) et négatifs (étiquetés O) dans nos données (c.f. tableau 7.2).

7.1.1.1 Ré-échantillonage

Lorsque qu'il y a suffisamment de données de références, un découpage aléatoire de ces données en trois fragments mutuellement exclusifs (*train*, *dev.* et *test*) éventuellement de tailles équivalentes permet d'obtenir une estimation fiable de la qualité du modèle. Cette méthode est appelée « *holdout* » (Maimon & Rokach, 2005, p. 293). Il est possible de répéter la méthode *holdout* plusieurs fois, avec des découpages différents afin de calculer un score moyen, plus fiable. Or cette option est proscrite dans notre cas car nous ne sommes pas certains de disposer de corpus suffisamment volumineux.

Parmi les méthodes disponibles pour remédier à cette limitation, la plus populaire (Rao *et al.*, 2008) est la *validation croisée*. Il s'agit d'un cas particulier de la méthode *holdout* répétée, pour lequel les portions de *test* ne se recouvrent jamais d'une itération à l'autre. Les données sont découpées en k parties mutuellement exclusives. Un modèle est entraîné sur $k-2$ portions de *train* et une portion de *dev.*, et testé sur une portion de *test*. Cette opération est réalisée k fois, avec chaque fois des parties *dev.* et *test* différents. Plus la valeur de k est grande, meilleure est l'estimation du score [3]. Kohavi *et al.* (1995) indiquent qu'en dessous de $k = 10$, les estimations de scores tendent à souffrir plus d'un biais pessimiste. Avec $k = 10$, les estimations sont raisonnablement bonnes, et à partir de $k = 20$, ces dernières ne semblent plus souffrir aucun biais.

En ce qui nous concerne, pour des questions de temps de traitement, nous avons choisi $k = 10$ pour l'évaluation.

3. Un cas particulier également populaire de validation croisée, la *leave-one-out cross-validation* (abrégée LOOCV), consiste à réduire l'échantillon *test* à un seul exemple à classifier, et à utiliser le reste des exemples pour l'apprentissage du modèle. L'opération est donc répétée autant de fois qu'il y a d'exemples.

De la même façon, la normalisation et la discrétisation [4] des valeurs effectuées lors de la phase de pré-traitement numérique émancipe ces traits de certains paramètres liés aux corpus (comme leur taille) ce qui rend global le calcul des traits. Nous avons fait le choix d'utiliser les traits calculés sur l'ensemble du corpus d'évaluation pour chaque sous-corpus de l'évaluation croisée. Nous nous attendons à ce que ce choix n'ait pas une influence considérable sur les scores d'évaluation.

7.1.1.2 Rééquilibrage des corpus

Le tableau 7.2 (p. 120) présente le nombre d'instances par étiquette dans chacun des corpus de spécialité. La distribution des classes n'est pas uniforme, et certaines étiquettes sont même extrêmement rares en comparaison de la classe majoritaire : les exemples étiquetés I apparaissent en moyenne 0,26% du temps alors que les exemples négatifs (étiquetés O) représentent en moyenne près de 80% des exemples d'entraînement.
Il s'agit d'un problème courant lorsqu'on est confronté à des données réelles. L'ennui réside dans le fait que l'algorithme d'apprentissage automatique est conçu pour maximiser le nombre global d'étiquettes bien devinées, nombre sur lequel les classes disposant de peu d'exemples ont une influence quasi nulle. Par exemple, dans un cas de classification binaire pour lequel les exemples positifs représentent moins de 1% du nombre total d'exemples, il suffira à l'algorithme de ne pas les prendre en compte pour espérer obtenir un taux d'exemples correctement classifiés d'au moins 99%.

Visa & Ralescu (2005) et Mollineda & Sotoca (2007) entre autres ont proposé un état des lieux des solutions déjà envisagées dans la littérature pour tempérer cet inconvénient. Certaines pallient le problème directement au niveau de l'algorithme, en pondérant les classes de façon favorable ou non. D'autres scindent le problème, en le réduisant à plusieurs classifications binaires opposant les exemples majoritaires aux exemples minoritaires d'une seule classe à la fois et en appliquant un rééquilibrage sur les données restantes. Le *rééquilibrage* (« *resampling* » en anglais) est la famille de méthodes la plus couramment utilisée. Elle consiste à modifier le ratio exemples positifs (rares) / exemples négatifs (majoritaires). On peut procéder en ignorant certains exemples négatifs. En ce cas, une partie de l'information, qui peut être utile pour l'entraînement du modèle, est perdue. Cette option est appelée « *down-sampling* » ou « *under-sampling* ». L'option inverse, appelée « *up-sampling* » ou « *over-sampling* », consiste à dupliquer des exemples minoritaires. Cela augmente la taille des données d'entraînement sans gain d'information supplémentaire, et peut avoir pour effet de bord un sur-

4. La discrétisation divise un certain univers en un nombre de segments. On peut donc voir la normalisation comme un cas particulier de discrétisation dans un univers normalisé, d'autant plus que l'on perd en précision en arrondissant.

apprentissage [5]. Il est également envisageable d'associer *up-sampling* et *down-sampling*. Il est à noter toutefois qu'une distribution totalement équilibrée n'est pas une garantie d'amélioration (Provost, 2000), et que l'influence du déséquilibre dépend aussi de l'algorithme d'apprentissage utilisé. Concernant les CRF, il n'existe aucune étude spécifiquement dédiée à étudier l'influence du déséquilibre des données sur les performances des prédictions ; Toutefois, il y a des fortes présomptions, notamment d'après les constatations de Wang *et al.* (2011), que les CRF y soient sensibles.

Pour ce qui concerne nos données d'apprentissage, notre stratégie de *down-sampling* a consisté à ne conserver que les exemples négatifs (étiquette O) se trouvant dans une fenêtre de 4 tokens autour d'un exemple positif dans la limite des frontières de phrases. Le choix de la taille de la fenêtre a notamment été dicté par la taille maximale des n-grams envisagés pour le calcul des traits ($n = 3$). La figure 7.3 présente le résultat de cette heuristique de *down-sampling*, en terme de proportion des étiquettes dans le corpus d'entraînement rééquilibré.

TABLE 7.3 – Proportion des étiquettes pour les corpus d'entraînement pour l'évaluation, ré-équilibré par *down-sampling*.

	Proportion des classes (%)				
Langues (ISO)	B	I	L	U	O
ara	1,33	0,07	1,33	20,55	76,73
deu	0,93	0,1	0,93	18,88	79,15
eng	2,06	0,34	2,06	21,5	74,03
fra	2,29	0,77	2,29	20,32	74,33
pol	1,2	0,11	1,2	23,27	74,23
tur	1,39	0,07	1,39	21,33	75,83
zho	4,41	0,72	4,41	21,33	75,83
Moyenne	1,94	0,3	1,94	21,03	74,77

Les données d'entraînement sont toujours très déséquilibrées, notamment en ce qui concerne les étiquettes B, I et L. Cependant, les étiquettes U dépassent quasiment toutes le cinquième du nombre d'exemples. Par ailleurs, d'un corpus à l'autre, les proportions sont plus comparables. Les expériences menées avec les données ainsi ré-équilibrées produisent des scores intéressants (voir notamment le chapitre suivant et résultats détaillés présentés dans l'annexe B),

5. Des variantes plus élaborées, comme la méthode SMOTE (Chawla *et al.*, 2002), permettent de générer artificiellement des instances de classes minoritaires à partir d'exemples proches, et générer ainsi de l'information utile.

c'est pourquoi nous avons jugé ce ré-équilibrage suffisant pour cette recherche.

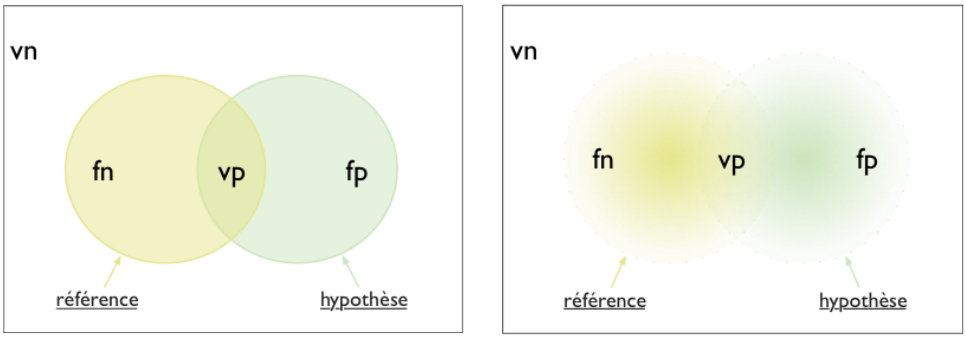
7.2 Métriques d'évaluation

7.2.1 Problématique

Comme nous l'avons vu à la section précédente, l'obtention d'un score d'évaluation est envisageable en comparant les étiquettes proposées par un modèle (entraîné sur les parties *train+dev.*) appliqué à la portion de *test* aux étiquettes de références pour cette même portion.

Cette comparaison peut être binaire (un terme est correct ou pas), et il s'agira juste dans ce cas de déterminer ce qui a été bien deviné, le bruit récupéré (lorsque les candidats termes ne sont pas corrects) et les silences (lorsque des termes corrects ne sont pas détectés par le système). Dans ce cas, les notions traditionnelles de *précision* et *rappel* peuvent s'appliquer, ainsi que d'autres mesures d'évaluation comme l'*exactitude*, la *spécificité*, etc. (Manning & Schütze, 1999). La figure 7.1 (a) représente, dans le cas binaire, les possibilités de comparer, pour une classe, les données de référence et les données prédites.

FIGURE 7.1 – Diagrammes ((a) adapté de (Manning & Schütze, 1999, p.268)) motivant les mesures de *précision* et *rappel*. Appliqué à l'évaluation de candidats termes, tout ce qui a été correctement deviné, les vrais positifs (vp) se situent dans l'intersection des ensembles de référence et d'hypothèse. Les éléments exclusivement dans l'ensemble de référence, les faux négatifs (fn), n'ont pas été devinés par le système. À l'inverse, les éléments présents exclusivement dans l'ensemble d'hypothèse sont des faux positifs (fp), devinés à tort. Enfin, tous les éléments ne faisant pas partie de l'union des deux ensembles sont des vrais négatifs (vn).



(a) Cas binaire (a) Gradient

Les mesures de précision (P) et rappel (R) binaires sont calculées suivant les formules :

$$P = \frac{vp}{vp + fp} \qquad R = \frac{vp}{vp + fn}$$

Une comparaison binaire souffre toutefois d'un biais négatif concernant l'évaluation de candidats termes par rapport à une référence. Tout d'abord, dans les typologies consacrées aux variations de termes (notamment Daille (2005), mais aussi Jacquemin (1999), et Savary (2000) parmi d'autres) [6], sont recensées les variations orthographiques (*proactif / pro-actif*), morphologiques (*uniformiser/uniformisation*), morpho-syntaxiques (*faire des économies/économiser*) et syntaxiques (*développement personnel et développement professionnel/développement personnel et professionnel*). Au total, Daille (2005) indique que ces variations peuvent représenter jusqu'à 35% des éléments d'une terminologie. Par ailleurs, Vivaldi & Rodríguez (2007) indiquent que pour leur expérience, des experts de domaine constituant une terminologie de référence se sont accordés uniquement sur 37% des termes choisis [7]. Ces divergences peuvent par exemple survenir lorsque les experts et terminologues ne s'accordent pas sur la granularité à donner à la terminologie. Parce qu'il est impossible de recenser l'ensemble des variations de termes pour un domaine spécialisé mais non technique, il est courant que certains termes équivalents n'apparaissent pas dans la terminologie de référence (*travailler correctement / bien bosser*). Il est donc impossible d'avoir une donnée terminologique de référence parfaite à tous points de vue.

Ce « gradient d'adhésion », qui apparaît aussi bien pour la terminologie de référence que pour la terminologie candidate, est schématisé dans la figure 7.1 (b). Pour l'ensemble de référence (resp. d'hypothèse), les termes faisant consensus (resp. ayant un score de prédiction élevé) se trouvent au centre et les termes acceptables (resp. ayant un score de prédiction à la limite du refus) se trouvent en périphérie. Ces contours imprécis rendent une évaluation automatique des terminologies difficile.

Nazarenko *et al.* (2009) ont dressé un bilan des tentatives qui ont été faites pour évaluer et comparer des systèmes d'extraction terminologique monolingues, ou tout du moins des tâches s'en approchant. Parmi elles, l'initiative japonaise NTCIR [8] incluant l'évaluation binaire d'une tâche de reconnaissance de termes (TERMREC) (Kando *et al.*, 1999), jamais reconduite. Plus récemment, la Campagne d'Évaluation des Systèmes d'Acquisition des Ressources Termino-

6. Dans ces typologies, c'est une approche quasi-wüsterienne du terme, dont les paradigmes syntaxiques et sémantiques sont très contrôlés, qui est entendue. Toutefois admettre la validité de ces variations relaxe le cadre formel des terminologies ainsi produites, en fonction des applications.

7. Vivaldi & Rodríguez (2007) font état de coefficients d'accord inter-annotateurs *kappa* entre leurs trois annotateurs très faibles : $-0,05$, $-0,12$ et $0,31$. Ils comparent ces chiffres au seuil optimal, selon Carletta (1996), de $0,8$.

8. *NII Testbeds and Community for Information access Research*, de l'Institut National d'Informatique du Japon

logiques (CESART) proposée par El Hadi *et al.* (2006) sur des corpus de domaines en français (médical, éducation), a permis d'obtenir, pour sa tâche d'extraction terminologique, des scores sur différents critères : cinq valeurs d'acceptation, allant de la correspondance parfaite entre un candidat terme et un terme de référence à une correspondance partielle à différents degrés. Cela a été fait en soumettant les résultats des systèmes en compétition à une évaluation manuelle menée par des experts, simplifiée par la mise à disposition d'une liste de termes de référence. Toutefois, ce mode d'évaluation est coûteux en temps, et demande par ailleurs la maîtrise parfaite de la langue concernée et du domaine pour évaluer les nuances. D'autres tentatives ont été faites, sur des tâches et avec des méthodes diverses. Nous renvoyons à Nazarenko *et al.* (2009) pour un état de l'art détaillé.

Plus récemment, Mondary *et al.* (2012) (inspirés par les propositions de Nazarenko & Zargayouna (2009)) ont lancé une nouvelle initiative, à notre connaissance la plus aboutie car la plus objective et la mieux reproductible, dans le cadre du programme *Quæro*. Elle a la triple intention de mesurer à la fois, pour chaque système en compétition, la mise à l'échelle sur de grosses masses de textes (et du même coup, la stabilité des métriques utilisées), les progrès effectués d'une version d'un système à une autre et l'influence du type de corpus sur les résultats. L'évaluation ne concerne que la tâche d'extraction terminologique monolingue, et aucune activité annexe comme le classement ou le regroupement de variantes. La comparaison se fait par rapport à une terminologie de référence, grâce à des métriques de *précision* et de *rappel* graduelles, adaptées à l'extraction terminologique. Parce que ces dernières semblent faire le lien entre évaluation binaire et évaluation manuelle en modélisant un gradient d'acceptation (fig. 7.1 (b)), c'est le cadre évaluatif dont nous avons souhaité nous rapprocher. Le détail des métriques pour le calcul des scores sera présenté dans la section 7.2.

7.2.2 Précision et rappel terminologiques de Nazarenko *et al.* (2009)

7.2.2.1 Description de la mesure

Nazarenko *et al.* (2009) (ou Nazarenko & Zargayouna (2009) pour une version résumée, en langue anglaise) préconisent que *Précision* et *Rappel* terminologiques considèrent, sans connaissance linguistique, les variations sur deux niveaux :

— au niveau des unités de traitement [9], pour prendre en compte les variations orthographiques et morphologiques. Cela donne lieu au calcul d'une distance d'édition sur les chaînes de caractères d_s (en l'occurrence, une distance de Levenshtein normalisée [10]) ;

9. Ici, d'éventuelles UTE ont été substituées aux tokens informés auxquelles elles correspondent initialement afin de permettre une comparaison avec la terminologie de référence.

10. La distance de Levenshtein est une mesure de similarité basée sur le nombre d'opérations (suppressions, insertions ou substitutions) entre une chaîne de caractère source s et un chaîne de caractères cible c. Plus la distance de Levenshtein est grande, plus s et c sont différentes. Cette distance de Levenstein peut être normalisée

— au niveau du terme, afin d'éventuellement détecter des variations morpho-syntaxiques ou syntaxiques ; Sur le même principe que la distance d_s, une distance sur les termes complexes d_c est calculée, à la différence que l'alphabet considéré ne concerne plus les caractères mais les unités de traitement considérées (dans notre cas, des tokens informés).

La distance globale entre deux termes $d_t(t_1, t_2)$, qui servira à modéliser le gradient de pertinence entre deux termes, est la moyenne arithmétique des distances $d_s(t_1, t_2)$ et $d_c(t_1, t_2)$.

Par la suite, *Précision terminologique* (PT) et *Rappel terminologique* (RT) comparent l'ensemble des termes candidats $\mathcal{C}$ à l'ensemble des termes $\mathcal{G}$ [11] de la terminologie de référence en respectant les impératifs suivants :

— Si $\mathcal{C} = \mathcal{G}$ (prédiction parfaite), alors $PT = RT = 1$;

— Si $\mathcal{C} \cap \mathcal{G} = \emptyset$ (aucun terme correctement deviné), alors $PT = RT = 0$;

PT et RT ne sont pas calculés directement par rapport à l'ensemble des candidats termes $\mathcal{C}$ mais sur une partition $\mathbb{P}(\mathcal{C})$ de $\mathcal{C}$, définie relativement à $\mathcal{G}$. Chaque élément p de $\mathbb{P}(\mathcal{C})$, qui est donc un sous ensemble de $\mathcal{C}$ contient :

— soit un ensemble de termes candidats qui sont proches d'un même terme de $\mathcal{G}$, sans que leur distance ne dépasse un seuil τ, car aucun des autres éléments de $\mathcal{G}$ n'est à une distance inférieure à τ ;

— soit un seul candidat terme qui ne peut être apparié avec aucun terme de référence.

Autrement dit, $\forall p \in \mathbb{P}(\mathcal{C})$ [12] :

$$p = \begin{cases} \{t_1, \ldots, t_n\} & \text{si } \exists t_{\mathcal{G}} \in \mathcal{G} : \\ & \quad \forall i \in [1, n],\ \forall t'_{\mathcal{G}} \in \mathcal{G} : \\ & \quad\quad t(t_i, t'_{\mathcal{G}}) \geq d_t(t_i, t_{\mathcal{G}}) \text{ et } d_t(t_i, t_{\mathcal{G}}) \leq \tau \\ \{t\} & \text{si } \nexists t_{\mathcal{G}} \in \mathcal{G} \text{ tel que } d_t(t, t_{\mathcal{G}}) \leq \tau \end{cases}$$

Le seuil τ est la limite globale qui interdit de comparer deux termes trop différents [13]. Chaque partition rassemble ainsi les termes candidats se rapprochant d'un même terme de référence $t_{\mathcal{G}}$. Nazarenko *et al.* (2009) définissent ensuite la pertinence d'un élément p de $\mathbb{P}(\mathcal{C})$ comme la pertinence maximale de tous les termes de p par rapport à $t_{\mathcal{G}}$. La fonction de *pertinence*, appli-

(ramenée à un espace de définition $[0, 1]$) afin de permettre une comparaison en divisant pas la longueur de la plus courte ou de la plus longue des chaînes comparées.

11. Nous prenons $\mathcal{G}$ pour *Gold* pour dénoter la terminologie de référence.

12. On notera que la définition donnée par Nazarenko *et al.* (2009) est insuffisamment précise. La version donnée dans Nazarenko & Zargayouna (2009) corrige cette imprécision. La version initiale ne donne en effet, dans le cas à plusieurs éléments, que la condition selon laquelle $d_t(t_i, t_{\mathcal{G}})$ doit être inférieur à τ. Mais la seconde condition y est sous-entendue : les éléments de p ne doivent pas être plus proches d'un autre $t'_{\mathcal{G}}$ de $\mathcal{G}$ que de $t_{\mathcal{G}}$.

13. Nazarenko *et al.* (2009) proposent de choisir ce seuil automatiquement en fonction du point d'inflexion des courbes de précision terminologique. Dans leur expérimentation pour l'anglais, ce seuil se situait entre $0,4$ et $0,5$.

quée à chaque élément de la partition de $\mathcal{C}$, vérifiant $|\mathcal{C} \cap \mathcal{G}| \leq Pert(\mathcal{C}, \mathcal{G}) \leq \min(|\mathcal{C}|, |\mathcal{G}|)$, est pour un terme candidat $t_\mathcal{C}$:

$$Pert_\mathcal{G}(t_\mathcal{C}) = \begin{cases} 1 - \min_{t_\mathcal{G} \in \mathcal{G}} d_t(t_\mathcal{C}, t_\mathcal{G}) & \text{si } \min_{t_\mathcal{G} \in \mathcal{G}} d_t(t_\mathcal{C}, t_\mathcal{G}) \leq \tau \\ 0 & \text{sinon.} \end{cases}$$

Elle est alors, pour unélément p de la partition, $Pert_\mathcal{G}(p) = max_{t_\mathcal{C} \in p} Pert_\mathcal{G}(t_\mathcal{C})$. Les mesures de rappel et précision terminologique sont définies comme suit :

$$TP = \frac{\sum_{p \in \mathbb{P}(\mathcal{C})} Pert_\mathcal{G}(p)}{|\mathbb{P}(\mathcal{C})|} \qquad TR = \frac{\sum_{p \in \mathbb{P}(\mathcal{C})} Pert_\mathcal{G}(p)}{|\mathcal{G}|}$$

Il est à noter que cette méthode ne prends pas (ou peu) en compte des variations sémantiques, liées à la synonymie. Par ailleurs, cette mesure n'a été éprouvée, à notre connaissance, que sur l'anglais (Nazarenko & Zargayouna, 2009 ; Nazarenko *et al.*, 2009 ; Mondary *et al.*, 2012). Bien que les auteurs la présentent comme une mesure ne dépendant pas de la langue, elle nécessite, comme nous le verrons au sujet de différents paramètres, des ajustements pour des langues typologiquement éloignées. C'est pourquoi nous avons fait le choix de nous inspirer largement du calcul des scores suivant les propositions de Nazarenko *et al.* (2009), en y ajoutant quelques adaptations supplémentaires, mieux applicables à une évaluation multilingue, en utilisant une formalisation sensiblement différente [14].

Nous avons choisi de formaliser le problème sous forme de graphe. La méthode de Nazarenko *et al.* (2009) serait implémentée comme suit si les deux terminologies à comparer étaient intégrées dans un graphe : chaque nœud du graphe représente un terme, associé à un type REFERENCE ou CANDIDAT. Pour chaque couple de nœud (REFERENCE, CANDIDAT) sont calculés les scores d_s et d_c, ainsi que le score d_t. Concernant le score d'édition, nous choisissons d'utiliser ici, ainsi que dans le reste de nos traitements et descriptions, un score d'édition normalisé basé sur une distance d'édition de Levenshtein [15] qui, à la différence de la distance d'édition du même nom, assigne 1 aux chaines de caractères identiques et 0 à celles n'ayant rien en commun. Les seuils à ne pas dépasser deviennent alors des limites inférieures, et non plus des limites supérieures. Si ce score d_t est supérieur à un certain seuil τ, le nœud CANDIDAT est raccordé au nœud REFERENCE correspondant via des arcs pondérés avec le résultat du score d_t.

14. Mondary *et al.* (2012) proposent une implémentation de leur méthode avec l'outil *Termometer*, librement disponible à l'adresse http://sourceforge.net/projects/termometerxd/

15. Plus précisément, une distance de Levenshtein qui pénalise plus les opérations de substitution (coût de 2) que les autres opérations (coût de 1). La normalisation se fait par rapport à la longueur de l'alignement le plus court entre les deux chaînes de caractère à comparer. Il s'agit de la fonction python ratio de la librairie Levenshtein.

Puis, afin d'identifier les différents sous-ensemble de la partition $\mathbb{P}(\mathcal{C})$ et leur pertinence, chaque nœud c_i CANDIDAT est passé en revue :

- si c_i est isolé, il constitue un sous-ensemble à lui seul. Sa pertinence est nulle ;
- sinon, on identifie l'arc sortant de c_i ayant le plus gros poids [16] menant à un nœud typé REFERENCE g_j. On scrute également les voisins CANDIDATS du nœud g_j pour récupérer l'arc le mieux pondéré reliant g_j à ses voisins typés CANDIDATS. Le poids de cet arc correspond à la pertinence de tous les termes candidats voisins de ce nœud g_j, et donc de c_i.

7.2.2.2 Inconvénients de cette méthode pour nos données

Concernant l'application de cadre évaluatif à un ensemble de langues variées, et pour une extraction de termes non-standards, nous voyons quelques inconvénients mineurs à l'algorithme proposé par Nazarenko *et al.* (2009).

Tout d'abord, concernant l'influence du type de langue sur le résultat des scores d_s et d_c, la mesure de Nazarenko *et al.* (2009) fonctionne correctement sur des langues ayant un système d'écriture alphabétique. Cela est dû au fait que les différents alphabets contiennent un nombre réduit de symboles, et les unités de traitement (tokens/mots) ont en général une longueur suffisante qui permet l'application d'une mesure d'édition. À ce titre, le chinois est une langue problématique pour l'évaluation terminologique. Chaque token informé est composé, dans notre expérience, d'en moyenne $2,5$ caractères pour le chinois. Ce nombre monte à 6 pour l'arabe, $7,2$ pour l'anglais, 8 pour le français, $8,8$ pour le polonais, $9,8$ pour le turc et $10,8$ pour l'allemand. Une distance d'édition telle qu'un score de Levenshtein n'aura pas une granularité suffisante en chinois, et ce même sur des transcriptions romanisées comme le Pinyin (Wang *et al.*, 2013). En parallèle de ce paramètre concernant le nombre de symboles (réduit pour un système d'écriture alphabétique, bien plus grand pour les systèmes d'écriture morpho-syllabaires comme le chinois), il faut également considérer le paramètre morphologique. Le score d'édition normalisé est efficace pour la comparaison de chaînes de caractères, et par extension pour la comparaison de mots dans les langues exhibant une variabilité morphologique relativement restreinte : il suffit que les variantes d'un terme puissent être identifiées de façon satisfaisante.

Or, pour certaines langues à morphologie riche, ce principe de comparaison n'est pas nécessairement adapté. Prenons par exemple le turc, qui concatène ses affixes (avec harmonie vocalique). Il arrivera souvent qu'une longue séquence d'affixes produise un ratio de Levensthein

16. Nous rappelons une fois de plus que là où Nazarenko *et al.* (2009) cherchaient un score minimal, nous recherchons un poids maximal car notre fonction d'édition assigne ses scores sur une échelle inversée.

plus élevé entre deux mots totalement différents qu'entre deux variantes d'un mot, comme cela est illustré dans le tableau 7.4.

TABLE 7.4 – Ratios de Levensthein normalisés pour les termes turcs « *kültür* » (culture), « *kültürümüzden* » (notre culture) et « *gözü* » (yeux), « *gözümüzden* » (nos yeux). Plus le ratio de Levenshtein est élevé, plus les chaines de caractères sont similaires.

	kültür	kültürümüzden	gözü	gözümüzden
kültür		0.64	0.43	0.38
kültürümüzden			0.26	0.67
gözü				0.63
gözümüzden				

En outre, le fait de calculer et conserver l'ensemble des distances d'édition nécessaires pour comparer tous les termes et les tokens entre eux fait exploser le nombre de chemins à envisager dans notre graphe pour la suite des traitements. Pour cette raison, nous voyons un intérêt à fixer également des seuils (τ_{min}, τ_{Lev}), indépendants de la langue, qui empêcheront la création d'arcs entre deux nœuds n'étant pas considérés comme comparables et à réduire ainsi les coûts computationnels de l'évaluation.

En second lieu, l'approche de Nazarenko *et al.* (2009) prend remarquablement bien en compte le fait que plusieurs termes candidats puissent correspondre à un terme de référence grâce aux pertinences des parties. En revanche, le fait qu'un terme candidat soit lié à plusieurs termes de référence n'est pas pris en compte. Dans nos données de test, nous avons remarqué à plusieurs reprises que l'acceptabilité de certains candidats termes complexes est sous-estimée. C'est le cas par exemple pour le terme « alléger les procédures ». Ce dernier n'apparait pas tel quel dans la terminologie de référence. En revanche, les termes « trop de procédures », « procédure », « procédures », « allégés » , « allégerai » , « allégeant » et « alléger » font partie des termes de référence.

Enfin, nous pensons qu'il peut être raisonnable de limiter l'influence de certains *stopwords*, aussi bien pour accélérer la vitesse de traitement que pour ne pas augmenter artificiellement les scores terminologiques entre termes candidats et termes de référence en contenant [17]. Pour ces raisons, nous avons utilisé un algorithme d'évaluation sensiblement différent, que nous décrivons dans la section suivante.

17. Notre terminologie de référence contient un certain nombre de *stopwords*. Par exemple en français : « sentiment *de* groupe », « recherche *et* développement », « garder *les* salariés », « *les plus* jeunes », etc.

7.2.3 Notre proposition de précision et rappel terminologiques

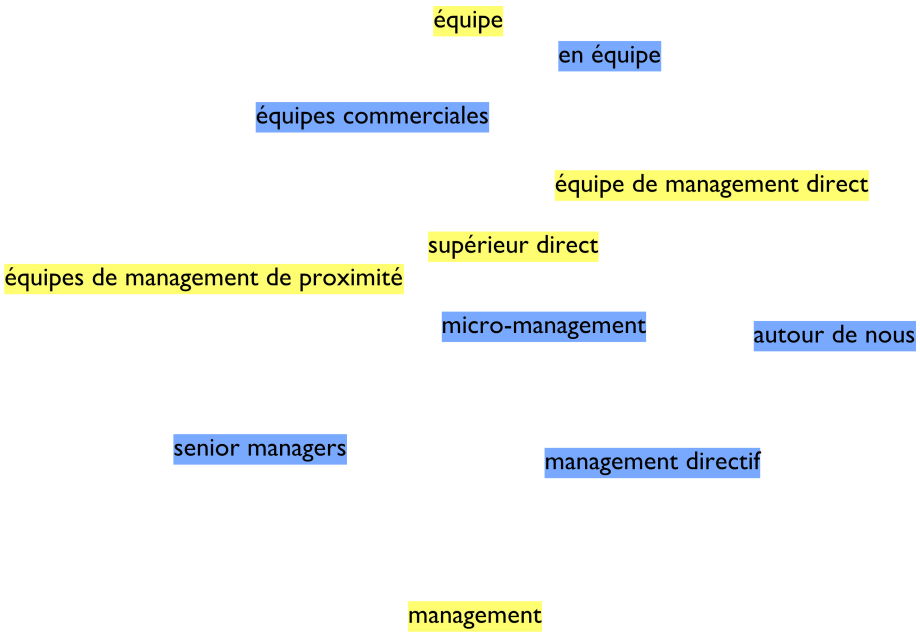
Comme nous l'avons énoncé plus haut, ce qui justifie l'utilisation de mesures d'édition dans le calcul de scores terminologiques relève de l'identification des variantes de termes à moindre frais et de façon satisfaisante. À ce titre, l'utilisation d'une mesure d'édition pour le chinois n'est pas une solution optimale. De plus, le coût computationnel d'une mesure de type Levenshtein est $\mathcal{O}(mn)$, avec m (resp. n) la longueur de la première (resp. deuxième) chaine de caractère donnée pour la comparaison. Pour les langues dont les tokens sont longs, comme l'allemand ou le turc, le nombre d'opérations élémentaire pour chaque comparaison est donc plus grand que pour les autres langues. Qui plus est, dans le cas du turc, l'hypothèse selon laquelle cette technique permettrait d'identifier les variantes d'un terme, même de façon approximative, n'est pas probante. Nous proposons donc une étape supplémentaire dans l'identification des variantes de termes, qui consiste à comparer plutôt des UTE modérées, dont nous disposons déjà, que des tokens. Pour rappel, les UTE modérées ont été obtenues suite à une segmentation morphologique avec l'outil Morfessor, avec l'élimination des affixes présumés les plus productifs dans une langue (voir la section 5.2).
Par ailleurs, nous proposons de prendre en considération le fait qu'un candidat terme puisse être lié à plusieurs éléments de la terminologie de référence en abandonnant l'idée de partition pour une approche, en apparence, probabiliste. Enfin, nous mettons à profit de petites listes de *stopwords* pour limiter l'influence de ces tokens dans le calcul des scores.

7.2.3.1 Construction du graphe pour l'évaluation

Pour une langue et un cadre expérimental donné, nous avons calculé nos scores en nous basant sur un graphe d'évaluation. Ce dernier est construit en 5 étapes :

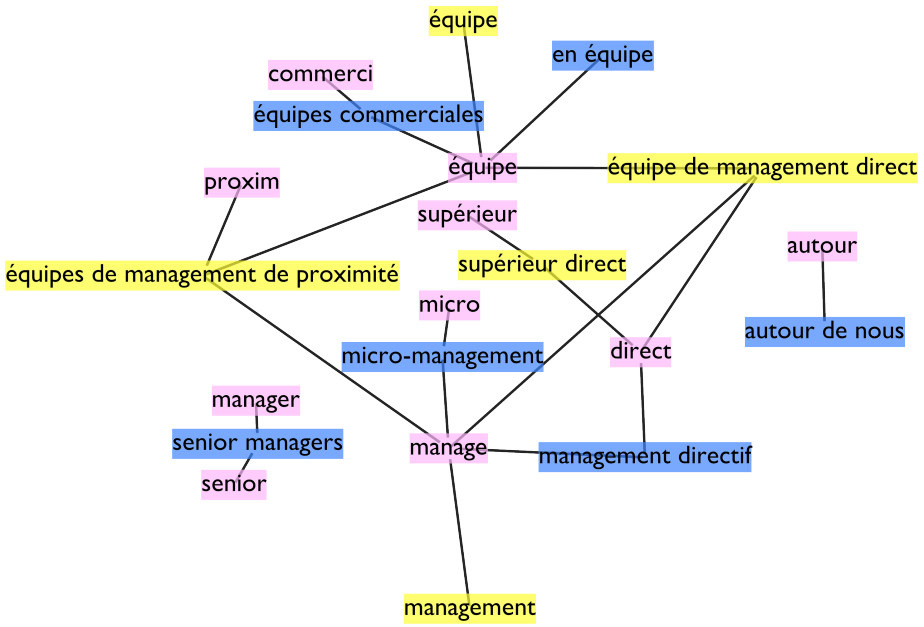
1. Dans un premier temps, les correspondances entre tokens informés et UTE modérées sont récupérés, ainsi que les listes de *stopwords*. Ces dernières ont été construites avec la méthode de référence (voir section 5.3) sur nos corpus génériques et spécifiques, et contrôlées manuellement. Leur longueur ne dépasse pas 30 *stopwords*.
2. Les listes de termes candidats et de référence issus d'une même portion de texte sont chargés dans un graphe non-orienté. En premier lieu, chaque terme, complexe ou pas, constitue un nœud du graphe. L'information concernant le statut du terme (REFERENCE, CANDIDAT) est également conservée sur chaque nœud. La figure 7.2 représente le résultat de cette étape. Les nœuds de type REFERENCE sont en jaune, les nœuds typés CANDIDAT sont en bleu.

FIGURE 7.2 – Construction du graphe d'évaluation — Deuxième étape : ajout des nœuds REFERENCE (en jaune) et CANDIDAT (en bleu).



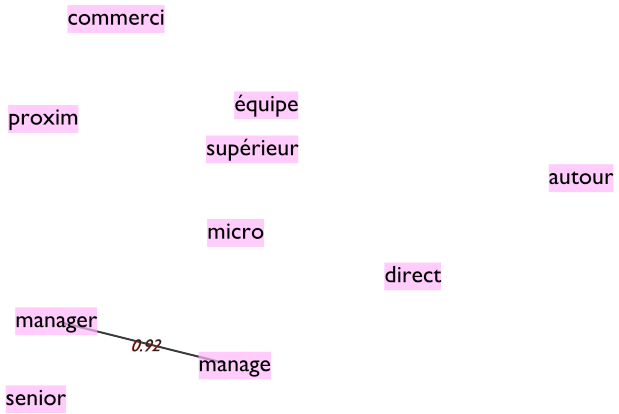
3. Pour chaque token informé qui constitue ces termes, si ce token n'apparait pas dans la liste de stopwords, l'algorithme récupère l'UTE modérée correspondante si elle existe ou conserve le terme tel quel sinon, pour créer un nœud de type UTE dans le graphe. Ce dernier est relié aux nœuds REFERENCE et/ou CANDIDAT qui le comprennent. Les arcs reliant les nœuds typés UTE aux nœuds typés REFERENCE et CANDIDAT possèdent une étiquette PARTS. La figure 7.3 représente le résultat de cette étape.

FIGURE 7.3 – Construction du graphe d'évaluation — Troisième étape : ajout des nœuds UTE (en rose pale) et de liens de type PARTS.



4. Une fois que l'ensemble des UTE représentées par les termes ont été intégrés dans le graphe sous forme de nœuds, des liens reliant ces nœuds sur la base d'une distance d'édition sont crées. Ainsi, tous les nœuds de type UTE sont comparés deux à deux avec le ratio de Levenshtein. Si ce dernier est supérieur à un seuil τ_{Lev}, un arc pondéré typé EDIT_SCORE est créé entre ces nœuds. Concernant le choix du seuil τ_{Lev}, nous avons estimé empiriquement que $\tau_{Lev} = 0,85$ était une valeur suffisamment contraignante pour prendre en compte des variations pertinentes (sur les UTE modérées, déjà amputées d'une partie de l'information morphologique) et malgré tout raisonnablement permissive pour capturer certains changement morphologiques qui resteraient à identifier. La figure 7.4 représente le résultat de cette étape, mais ne montre que les nœuds de type UTE et les arcs typés LEV_DIST (pour cet exemple, un seul arc typé EDIT_SCORE a obtenu un score suffisant pour être créé).

FIGURE 7.4 – Construction du graphe d'évaluation — Quatrième étape : ajout des arcs LEV_DIST représentant une distance d'édition acceptable entre deux nœuds typés UTE.



5. Pour chaque couple de nœuds (R,C), typés respectivement REFERENCE et CANDIDAT, on détermine leur distance terminologique $d_t(R, C)$ ainsi :
   — On établit la distance $d_s(R, C)$ en calculant simplement le ratio de Levenshtein entre les termes R et C :

$$d_s(R, C) = Levenshtein.ratio(R, C)$$

   — Si la distance $d_s(R, C)$ est suffisante [18], on calcule la distance $d_c(R, C)$. Pour cette dernière on recherche l'ensemble des chemins simples menant du nœud R au nœud C en passant par au plus deux nœuds de type UTE [19], afin de mettre à jour incrémentalement le score $d_c(R, C)$.
   Pour chacun de ces chemins $c_1, ...c_n$, on va calculer son importance I_{c_i} relativement aux deux nœuds qu'il relie en divisant le nombre d'arcs incident aux nœuds R et C concernés par ce chemin (invariablement 2) par la somme des degrés des nœuds R et C. Ce résultat sera éventuellement pondéré par le poids p_{lev} d'un éventuel arc de type LEV_DIST ($p_{lev} = 1$ si aucun arc de type LEV_DIST n'est présent dans c_i) présent dans le chemin entre deux nœuds UTE (chemins de longueur 4

18. Poser un seuil ici permet d'accélérer significativement les traitements, sans modifier outre mesure les résultats. On suppose qu'on ne souhaite pas continuer à comparer des termes si ce score n'atteint pas au minimum $\tau_{min} = 0,45$, car en deçà de ce seuil, moins de la moitié des items à comparer est comparable.

19. Il n'existe, à cette étape, aucun lien direct entre des nœuds typés REFERENCE et CANDIDAT. Les chemins concernés aurons donc une longueur l nécessairement comprise entre 3 et 4 ($3 \leq l \leq 4$), et ne serons que constitués d'arcs typés PARTS et LEV_DIST.

uniquement) :

$$I_{c_i} = \frac{2}{deg(R) + deg(C)} \times p_{lev}$$

Le score $d_c(R, C)$ est la somme de tous les scores I_{c_i} :

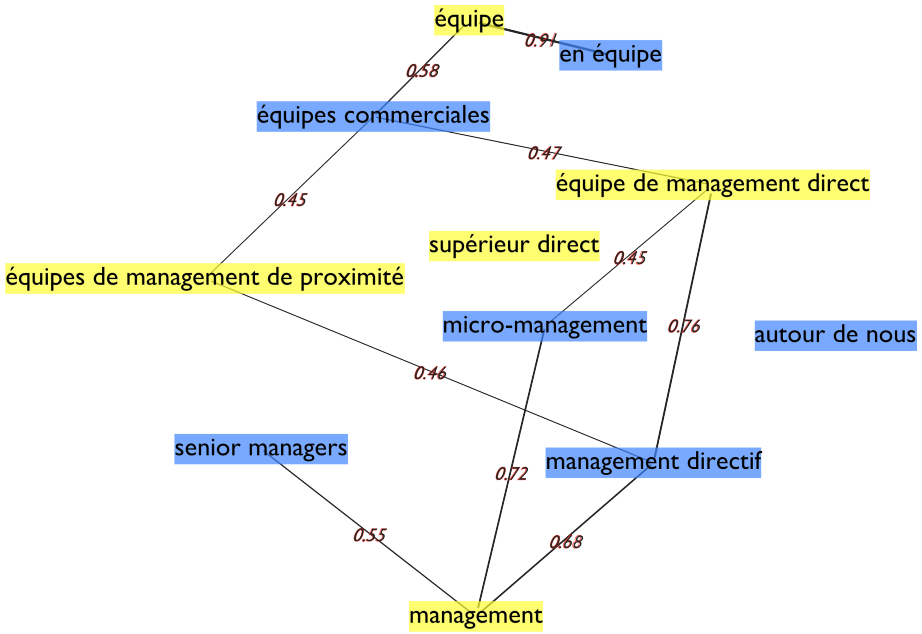
$$d_c(R, C) = \sum_{i=1}^{n} I_{c_i}$$

La distance terminologique $d_t(R, C)$ est, comme dans Nazarenko *et al.* (2009), la moyenne arithmétique de $d_s(R, C)$ et $d_c(R, C)$:

$$d_t(R, C) = \frac{d_s(R, C) + d_c(R, C)}{2}$$

Des arcs typés DISTANCE_TERMINO, pondérés avec la valeur de $d_t(R, C)$, sont alors créés entre les nœuds R et C, sans seuil minimum. Le résultat de cette étape est illustré dans la figure 7.5

FIGURE 7.5 – Construction du graphe d'évaluation — Cinquième étape : ajout des arcs DISTANCE_TERMINO représentant la distance $d_t(R, C)$ entre un nœud R typé REFERENCE et un nœud C typé CANDIDAT.



7.2.3.2 Procédure de calcul des scores

Une fois ce graphe d'évaluation obtenu, il est possible d'en déduire des scores finaux de plusieurs manières. Comme cela a été évoqué à la section 7.2.2, Nazarenko *et al.* (2009) [20] suggèrent de ne conserver pour chaque nœud typé CANDIDAT que la pertinence du sous ensemble de la partition à laquelle il appartient.

Nous souhaitons prendre en compte le fait qu'un nœud de type CANDIDAT soit très connecté, connecté avec des poids forts, ou au contraire très mal connecté au reste (peu ou avec des poids faibles). À cette fin, nous reformulons le problème en imaginant que les poids des arcs typés DISTANCE_TERMINO incidents, dont la valeur est comprise entre 0 et 1, correspondent à une probabilité d'un lien valide entre les nœuds CANDIDAT et REFERENCE qu'ils relient [21]. Soit un nœud CANDIDAT n ayant k arcs incidents $a_n^{(1)} \dots a_n^{(k)}$ de type DISTANCE_TERMINO. Si le poids $w_n^{(i)}$ de l'arc $a_n^{(i)}$ représente une probabilité $p(a_n^{(i)})$ que l'arc corresponde à un lien valide de n à un nœud de type REFERENCE, alors la probabilité que cet arc soit faux est : $p(\neg a_n^{(i)}) = 1 - w_n^{(i)}$. Suivant le même principe, la probabilité que le nœud n soit faux est égale à la probabilité que tous ses arcs incidents $a_n^{(1)} \dots a_n^{(k)}$ soient faux :

$$\begin{aligned} p(\neg n) &= \prod_{i=1}^{k} p\left(\neg a_n^{(i)}\right) \\ &= \prod_{i=1}^{k} \left(1 - w_n^{(i)}\right) \end{aligned}$$

La probabilité $p(n)$ que le nœud candidat n soit correct est donc :

$$p(n) = 1 - \prod_{i=1}^{k} \left(1 - w_n^{(i)}\right)$$

Bien entendu, il ne s'agit pas ici d'une probabilité mais d'un moyen de formaliser le calcul du score terminologique d'un nœud n (représentant un candidat terme) global, basé sur les différents scores terminologiques que ce terme entretient éventuellement avec des termes de référence. Nous continuons sur cette analogie probabiliste afin de déterminer le seuil Λ en dessous duquel nous souhaitons considérer qu'un candidat terme n'est pas correct. Pour $p(n) = \frac{1}{2}$, il y a théoriquement autant de chances que ce candidat soit bon ou mauvais. En pratique, cela signifie que seule la moitié de ce qui a été comparé était identique, ce qui en fait un seuil trop faible. Suivant le même raisonnement, nous considérons (arbitrairement) que le

20. Nous rappelons que le calcul de la distance terminologique d_t diffère entre les propositions de Nazarenko *et al.* (2009) et la notre.

21. En réalité, ce poids ne correspond pas à une probabilité, mais bel et bien à une distance terminologique.

seuil de $\Lambda = \frac{2}{3}$ est plus adapté, car il ne prend en considération que les éléments dont (significativement) plus de la moitié des choses comparées sont identiques. Nous reformulons donc la probabilité $p(n)$ que le nœud candidat n soit correct :

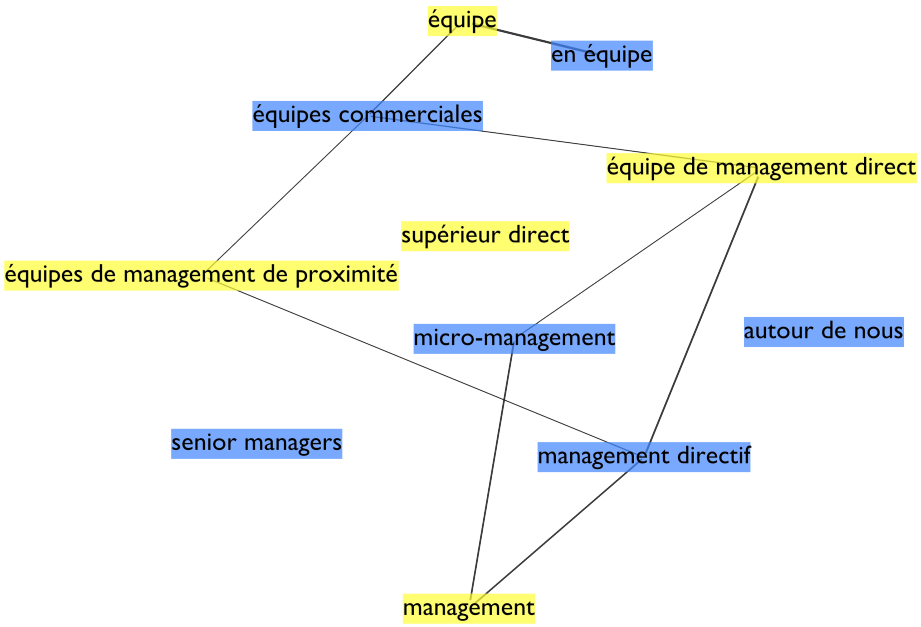
$$p(n) = \begin{cases} p(n) & \text{si} \quad \Lambda \geq \frac{2}{3} \\ 0 & \text{sinon} \end{cases}$$

Le calcul de ces pseudo-probabilités est effectué pour chaque nœud CANDIDAT n du graphe d'évaluation. Si la probabilité $p(n)$ que le nœud n soit un candidat valide est non nulle, alors l'ensemble des arcs typés DISTANCE_TERMINO incident au nœud n dont le poids est supérieur à Λ sont re-typés VALIDE. La figure 7.6 montre les liens de type VALIDE conservés dans notre exemple explicatif.

L'exemple suivant illustre l'effet du seuil Λ sur des unités similaires aux caractères chinois (ici remplacés par des lettres latines pour plus de clarté). Imaginons que le graphe d'évaluation ne contienne qu'un arc typé DISTANCE_TERMINO reliant un candidat terme « $A\ B$ » et un terme de référence « $A\ C$ ». L'arc reliant ces deux nœuds termes a un poids de $0,583$, et la probabilité [22] que le terme « $A\ B$ » soit correct est $p(A\ B) = 1 - (1 - 0,583) = 0,583$. Avec le seuil $\Lambda = \frac{2}{3}$, ce poids est jugé insuffisant. Même si ces deux termes ont un token en commun, on ne souhaite pas les comparer d'avantage, d'un point de vue lexical. Imaginons à présent que ce graphe d'évaluation compare un candidat terme « $A\ B$ » et un terme de référence « $A\ B\ C$ ». L'arc typé DISTANCE_TERMINO reliant ces deux termes a un poids de $0,775$. Étant donné que $p(A\ B) = 0,775$, le seuil Λ permet de considérer que les termes « $A\ B$ » et « $A\ B\ C$ » sont comparables. Suivant le même raisonnement pour la comparaison des termes « $A\ B$ » et « $A\ B\ C\ D$ », on obtient une distance terminologique (et une « probabilité » que le terme « $A\ B$ » soit juste) de $0,633$, à peine inférieure au seuil Λ. Le lien entre ces deux termes sera donc éliminé s'il n'existe pas d'autre preuve que ces termes peuvent être reliés. Si à ce dernier graphe d'évaluation est rajouté un nœud de référence supplémentaire « $A\ X\ Y\ B$ », qui est également à une distance terminologique de $0,633$ du terme candidat « $A\ B$ », la probabilité que ce dernier soit correct devient $p(A\ B) = 1 - \left((1 - 0.633)^2\right) = 0,865$. Par ailleurs, bien que le terme candidat « $A\ B$ » soit considéré comme correct avec une « probabilité » $p(A\ B) = 0,865$, sa « probabilité » d'être raccordé aux nœuds de référence « $A\ B\ C\ D$ » et « $A\ X\ Y\ B$ » est quant à elle toujours inférieure à Λ. Aucun de ces liens ne sera re-typé VALIDE car ils ne sont pas suffisamment similaires au nœud candidat pour que l'on souhaite les considérer avec certitude comme une variante de terme. Ce choix de ne pas valider les liens dont le score terminologique est inférieur à Λ aura son importance lors du décompte des faux négatifs.

22. Nous poursuivons ici notre évocation probabiliste tout en soulignant qu'il ne s'agit pas réellement d'une probabilité.

FIGURE 7.6 – Construction du graphe d'évaluation — Sixième étape : re-typage des arcs de DISTANCE_TERMINO en VALIDE lorsque la probabilité d'un nœud C typé CANDIDAT est suffisante.



Précision et rappel terminologiques sont ensuite calculés suivant les formules de précision et de rappel binaires données à la section 7.2.1, pour lesquelles :

- le compte de vrais positifs (vp) correspond à la somme de toutes les probabilités non nulles des candidats termes ($p(n) > 0$) ;
- le compte des faux positifs correspond à la somme de toutes les probabilités nulles des candidats termes ($p(n) = 0$) ;
- le compte de faux négatifs correspond au nombre de nœuds REFERENCE ne possédant aucun arc adjacent typé VALIDE.

Le fait de considérer pour le compte de vrais positifs les probabilités des candidats termes considérés comme corrects plutôt qu'un score de type binaire fait que les scores de précision et rappel ainsi obtenus ne sont pas binaires à proprement parler. Le score global de la terminologie à évaluer est ensuite simplement calculé avec une F-mesure, qui correspond à la moyenne harmonique de la précision et du rappel :

$$F = \begin{cases} 0 & \text{si} \quad P + R = 0 \\ 2 \times \frac{P \times R}{P + R} & \text{sinon} \end{cases}$$

Parce que certains termes extraits, même s'ils ne sont pas dans la ressource de références, peuvent être jugés pertinents par le système, notre méthode d'évaluation peut éventuellement comptabiliser plus de termes candidats corrects qu'il n'y en a dans la ressource de référence. Pour cette raison, le rappel (et par symétrie, la précision) a pour limite maximale 1.

Afin de disposer d'un élément de comparaison de notre nouvelle proposition de scores terminologiques avec la méthode proposée par Nazarenko *et al.* (2009), nous avons utilisé l'outil *Termometer*[23] pour obtenir des scores mieux comparables à ceux décrits par Nazarenko *et al.* (2009) ou Mondary *et al.* (2012) sur l'anglais, le turc et le chinois.

Le tableau 7.5 compare, pour les mêmes données, les scores terminologiques obtenus par les deux méthodes d'évaluation sur les trois meilleurs modèles anglais.
Les scores du tableau 7.5 ont été obtenus avec le seuil global 0.54061 pour l'anglais, 0.655781 pour le chinois et 0.528405 pour le turc ; ces derniers ont été déterminés par le script de *clustering* de *Termometer* sur nos données de référence. Pour l'ensemble des langues du tableau 7.5, notre système d'évaluation propose des f-score plus élevés (en moyenne $+0,68$ points de f-score). La différence est particulièrement marquée pour le chinois, pour lequel le f-score proposé par notre algorithme représente $0,182$ points en plus par rapport au score de *Termometer*, ce qui permet concomitamment au score du meilleur modèle chinois d'être plus homogène avec les scores de meilleurs modèles des autres langues, comme nous le verrons plus en détail dans le chapitre suivant. La principale différence entre ces deux algorithmes est que celui proposé par Nazarenko *et al.* (2009) favorise la précision : en moyenne $0,35$ points en plus par rapport à notre score de précision pour l'ensemble des cadres expérimentaux et des langues. Au contraire, notre algorithme avantage le rappel de $0,16$ points en moyenne.
Ces divergences sont expliquées par les différents choix d'implémentation détaillés *supra*. En premier chef, la non prise en compte d'un petit nombre de *stopwords*. Nous postulons que ce premier amendement à l'algorithme de Nazarenko *et al.* (2009) a en parti causé la baisse de précision constatée à travers toutes les langues. En second lieu, nous soupçonnons que la prise en compte de la proximité d'un terme candidat à plusieurs termes de référence a augmenté les taux de rappel, comme cela se manifeste de façon plus visible pour le chinois. Cette modification est par ailleurs exacerbée pour les langues morphologiquement riches de par l'utilisation d'UTE modérées en lieux et place des tokens originels pour la comparaison de termes complexes.

23. Téléchargé à l'adresse http://sourceforge.net/projects/termometerxd/files/latest/download.

TABLE 7.5 – Résultats obtenus par l'outil *Termometer*, avec rappel des scores obtenus par notre système sur nos meilleurs modèles (entraînés et appliqués) sur l'anglais dans les différents cadres expérimentaux.

Modèles	Scores de *Termometer*			Nos scores		
	P	R	F	P	R	F
Anglais						
Tokens (Sfreq+SIM+Dfreq+CP)	0,902	0,812	0,855	0,895	0,917	0,906
UTE M. (Sfreq+ZS+ODR)	0,849	0,772	0,809	0,742	0,938	0,828
UTE F. (Sfreq+Dfreq+MTD+ODR)	0,874	0,767	0,817	0,742	0,938	0,828
Chinois						
Tokens (ZS+Sfreq)	0.871	0.507	0.641	0.904	0.755	0.823
Turc						
Tokens (Sfreq+SIM)	0,986	0,874	0,926	0,991	0,948	0,969
UTE M. (Sfreq+DRK)	0,906	0,792	0,845	0,884	0,973	0,927
UTE F. (Sfreq+DRK)	0,908	0,798	0,849	0,887	0,974	0,928

7.3 Sélection des meilleurs modèles

Étant donné le nombre de langues à évaluer et la dimensionnalité de l'ensemble des combinaisons de traits à tester, nous avons procédé à une présélection d'un sous ensemble de traits qui maximisent les résultats pour une langue donnée. Pour estimer la validité de chaque modèle pour une langue donnée, nous nous reposons sur les scores P , R et F décrits plus haut.

Sachant qu'il y a 23 traits possibles pour l'entraînement des modèles, et que nous souhaitons utiliser inconditionnellement le trait relatif à la ponctuation [24], il nous faut déterminer, parmi les 22 traits restants quelles sont les combinaisons de traits les mieux adaptées pour chaque langue et chaque cadre expérimental (tokenisation simple, sous-spécifications modé-

24. La norme unicode permet de classer tous les caractères unicodes qui sont des marques typographiques en sept catégories universelles (voir la section 6.1.2). Cette caractéristique est donc utile quelle que soit la langue.

rée et franche). Tester l'ensemble des combinaisons (sans répétition) possibles des traits initiaux aurait amené à tester un nombre de modèles de l'ordre de $\sum_{n=2}^{22} \binom{n}{22} = 2^{22}$ (soit plus de 4 millions de modèles) par langue et par cadre expérimental, ce qui aurait été en pratique difficile à mettre en œuvre [25]. Nous avons donc opté pour une approche séquentielle ascendante partielle afin d'obtenir un protocole de sélection permettant d'identifier les meilleurs sous-ensembles de traits. Nous avons scindé l'évaluation en deux parties :

1. Phase 1 : Sélection initiale des traits. Dans un premier temps, nous avons entraîné des modèles ne reposant que sur un ou deux traits, en explorant tous les traits et tous les couples de traits possibles. Cette première phase d'entraînement a produit, pour chaque langue, 2520 modèles par cadre expérimental (22 modèles simples et 230 modèles combinés pour chacune des 10 parties de la validation croisée). Après évaluation, nous conservons, pour l'expérience concernée (langue, cadre expérimental), seulement les traits apparaissant dans les 3 meilleurs modèles.
2. Phase 2 : Recherche du meilleur modèle. La réduction drastique du nombre de traits effectuée en (1) nous permet d'envisager l'évaluation de modèles mélangeant plus de deux traits. Cette seconde phase d'entraînement va produire l'ensemble des modèles utilisant 3 des traits retenus issus de la première sélection de traits ou plus, jusqu'au modèle les utilisant tous.

À l'issue de cette seconde étape, nous considérons que parmi tous les modèles ainsi entraîné, le meilleur est celui qui obtiens le f-score le plus élevé.

25. Par « cadre expérimental », nous entendons les différentes expériences pour lesquelles les unités de traitements varient : tokens, UTE modérées et UTE franches.

Chapitre 8

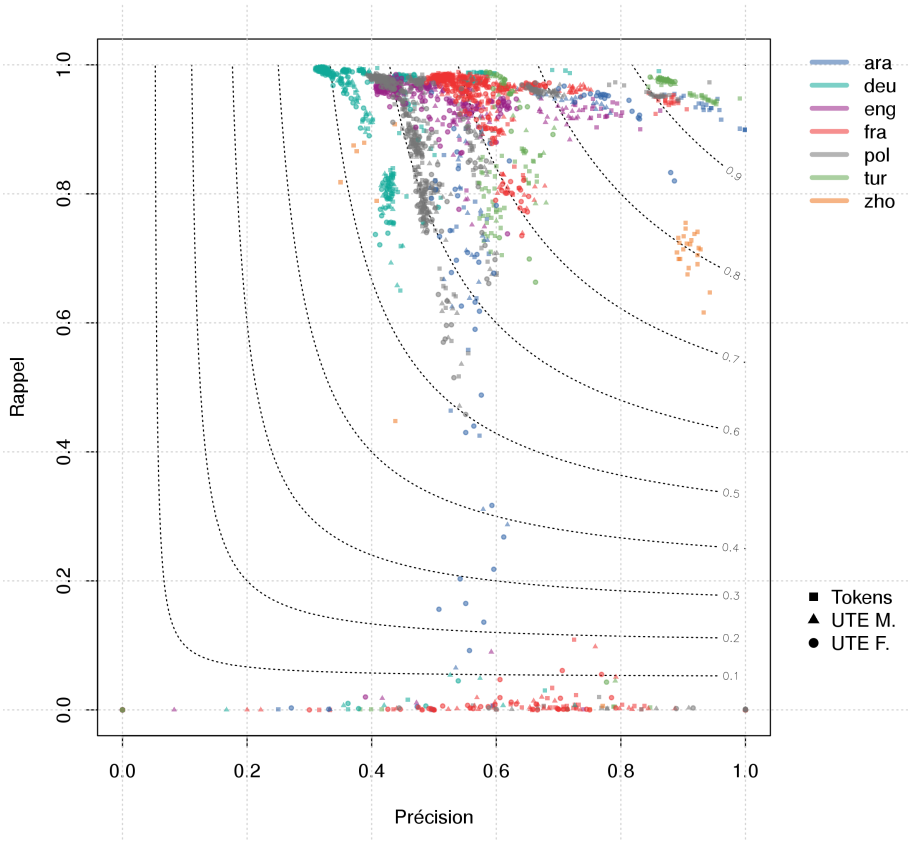
Résultats

Sommaire

Ce chapitre présente les résultats des différentes expériences qui mettent en œuvre les idées décrites aux chapitres 5 et 6, en les évaluant au moyen de la métrique proposée au chapitre précédent. Le présent chapitre est structuré en trois parties : Dans un premier temps, des résultats des expériences intra-lingues (modèles entraînés et appliqués sur une même langue) y sont présentés. Dans un second temps, ce sont les résultats des expériences inter-lingues qui sont développés. Enfin, nous discutons l'ensemble de ces résultats, y compris par mise en regard (à défaut qu'une comparaison directe soit vraiment possible) avec les résultats d'autres système d'extraction de terminologies.

8.1 Meilleurs modèles pour l'extraction de termes

Figure 8.1 – Dispersion des scores de tous les modèles. L'axe des abscisses correspond à la précision, l'axe des ordonnées au rappel, et les lignes en pointillés indiquent des iso-lignes de f-score.



L'intégralité des résultats pour l'étape de sélection initiale des traits pour nos sept langues

TABLE 8.1 – Proportions moyennes de termes complexes (comportant au minimum deux tokens ou UTE) pour l'ensemble des modèles productifs entraînés et évalués sur une même langue (cf. tableaux de l'annexe B).

Langue	*Proportions de termes complexes (%)*			Moyenne
	Tokens	UTE modérée	UTE franche	(%)
ara	7,9	8,49	5,75	7,38
deu	2,09	1,11	0,8	1,33
eng	7,44	8,47	7,04	7,65
fra	8,27	8,05	7,38	7,9
pol	2,6	2,22	1,94	2,25
tur	8,8	5,78	6,01	6,86
zho	41,97			41,97

TABLE 8.2 – Résultats des trois meilleurs modèles de la première phase d'évaluation et des modèles générés pour la deuxième phase d'évaluation pour l'arabe.

(a) Tokens

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+SIM	0.955	0.947	0.951
Sfreq+FKZY	0.999	0.9	0.947
Sfreq+BB	0.999	0.899	0.947
Phase 2 : Recherche du meilleur modèle			
Sfreq+SIM+FKZY+BB	0.964	0.952	0.958
BB+Sfreq+SIM	0.946	0.958	0.952
Sfreq+SIM+FKZY	0.909	0.968	0.937
BB+Sfreq+FKZY	0.927	0.938	0.932
BB+SIM+FKZY	∅	∅	∅

(b) UTE modérées

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+ttest	0.777	0.934	0.848
TT+Sfreq	0.783	0.925	0.848
ZS+Sfreq	0.769	0.939	0.846
Phase 2 : Recherche du meilleur modèle			
TT+Sfreq+ttest	0.855	0.904	0.879
Sfreq+ttest+ZS	0.784	0.943	0.856
Sfreq+ttest+ZS+TT	0.799	0.900	0.846
TT+Sfreq+ZS	0.744	0.933	0.828
TT+ttest+ZS	∅	∅	∅

(c) UTE franches

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
CP+Sfreq	0.830	0.922	0.874
Sfreq+ttest	0.813	0.945	0.874
PRS+Sfreq	0.828	0.918	0.871
Phase 2 : Recherche du meilleur modèle			
Sfreq+ttest+CP	0.825	0.920	0.870
PRS+Sfreq+ttest	0.859	0.820	0.839
PRS+Sfreq+CP	∅	∅	∅
PRS+ttest+CP	∅	∅	∅
Sfreq+ttest+CP+PRS	∅	∅	∅

TABLE 8.3 – Résultats des trois meilleurs modèles de la première phase d'évaluation et des modèles générés pour la deuxième phase d'évaluation pour l'allemand.

(a) Tokens

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+SIM	0.861	0.990	0.921
CP+Sfreq	0.718	0.987	0.831
Sfreq+ttest	0.711	0.991	0.828
Phase 2 : Recherche du meilleur modèle			
SIM+Sfreq+CP	0.878	0.990	0.931
ttest+Sfreq+SIM	0.863	0.989	0.922
Sfreq+SIM+CP+ttest	0.860	0.990	0.921
ttest+Sfreq+CP	0.706	0.992	0.825
ttest+SIM+CP	0.411	0.995	0.581

(b) UTE modérées

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+SIM	0.584	0.978	0.731
PMI+Sfreq	0.576	0.973	0.724
ODR+Sfreq	0.575	0.972	0.723
Phase 2 : Recherche du meilleur modèle			
Sfreq+SIM+PMI	0.581	0.981	0.730
SIM+Sfreq+ODR	0.568	0.983	0.720
Sfreq+SIM+PMI+ODR	0.559	0.980	0.712
Sfreq+ODR+PMI	0.557	0.976	0.709
ODR+SIM+PMI	∅	∅	∅

(c) UTE franches

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+ttest	0.470	0.979	0.635
Sfreq+SIM	0.468	0.985	0.635
Sfreq+Dfreq	0.468	0.984	0.634
Phase 2 : Recherche du meilleur modèle			
Sfreq+Dfreq+ttest+SIM	0.486	0.982	0.650
Sfreq+ttest+SIM	0.484	0.981	0.648
SIM+Sfreq+Dfreq	0.482	0.982	0.646
Sfreq+Dfreq+ttest	0.476	0.983	0.641
SIM+Dfreq+ttest	0.447	0.973	0.613

de tests et pour les différents degrés de sous-spécification proposés en amont (aucune, modérée, franche) est présentée dans les tableaux B.1 à B.19 de l'annexe B (p.251). La figure 8.1 donne une représentation graphique globale de la distribution des scores pour les différentes langues. La plupart de ces modèles ont un f-score supérieur à $0,5$. Le reste ont des f-scores proches de zéro. Seuls certains modèles de l'arabe dérogent à cette tendance. Cette figure permet d'estimer quelles seront les langues sur lesquelles l'application de tels modèles sera la plus prometteuse : particulièrement pour le turc, moins pour l'allemand.

Le tableau 8.1 donne les proportions moyennes du nombre de termes complexes extraits par l'ensemble des modèles produits pour les phases 1 et 2 de la sélection du meilleur modèle au sein d'une langue, pour toutes les langues de notre échantillon et dans tous les cadres expérimentaux. On observe que ces proportions moyennes vont de $1,33$ pour l'allemand à $41,97$ pour le chinois. Les langues ayant la plus faible proportion de termes complexes sont l'allemand et le polonais ; sur la figure 8.1, ce sont les langues pour lesquelles la plupart des mo-

TABLE 8.4 – Résultats des trois meilleurs modèles de la première phase d'évaluation et des modèles générés pour la deuxième phase d'évaluation pour l'anglais.

(a) Tokens

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+SIM	0.863	0.929	0.895
Sfreq+Dfreq	0.831	0.9	0.864
CP+Sfreq	0.794	0.940	0.861
Phase 2 : Recherche du meilleur modèle			
Sfreq+SIM+Dfreq+CP	0.895	0.917	0.906
Sfreq+SIM+Dfreq	0.854	0.939	0.894
CP+Sfreq+SIM	0.816	0.957	0.881
CP+Sfreq+Dfreq	0.789	0.941	0.858
Sfreq+SIM+CP	∅	∅	∅

(b) UTE modérées

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
ZS+Sfreq	0.750	0.917	0.825
ODR+Sfreq	0.733	0.929	0.819
Sfreq+ttest	0.724	0.938	0.817
Phase 2 : Recherche du meilleur modèle			
Sfreq+ZS+ODR	0.742	0.938	0.828
Sfreq+ttest+ODR	0.739	0.939	0.827
Sfreq+ttest+ODR+ZS	0.729	0.955	0.827
ttest+Sfreq+ZS	0.732	0.934	0.821
ttest+ZS+ODR	0.529	0.936	0.676

(c) UTE franches

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
MTD+Sfreq	0.632	0.947	0.758
Sfreq+Dfreq	0.633	0.932	0.754
ODR+Sfreq	0.620	0.942	0.748
Phase 2 : Recherche du meilleur modèle			
Sfreq+Dfreq+MTD+ODR	0.653	0.941	0.771
ODR+Sfreq+Dfreq	0.652	0.942	0.771
ODR+Sfreq+MTD	0.651	0.942	0.770
Sfreq+Dfreq+MTD	0.649	0.944	0.769
ODR+Dfreq+MTD	0.346	0.002	0.004

dèles se situent aux abord des iso-lignes de f-score 0, 5 et 0, 6. Parallèlement, le fait que près de la moitié des termes extraits en chinois comportent plusieurs tokens permet d'expliquer que, sur la figure 8.1, le mandarin se distingue nettement du reste des langues : alors que la plupart des langues maximisent leur f-score grâce au rappel, le chinois y parviens en optimisant sa précision.

Concernant la phase de sélection initiale des traits, les résultats issus des trois meilleurs modèles sont repris dans les tableaux 8.2 à 8.8, qui présentent également les résultats obtenus pour la sélection finale des meilleurs modèles en utilisant comme nom de traits les abréviations présentées dans le tableau 6.1 (p. 114). Sur les 22 traits présentés pour l'entraînement de modèle lors de la phase 1, seuls quatre par langues ont été retenus. Parmi ces derniers, le trait de fréquence Sfreq apparaît dans tous les modèles, quelle que soit l'expérience en jeu.

Pour l'expérience de référence sur les tokens, ce sont au total 15 traits qui ont été retenus à

TABLE 8.5 – Résultats des trois meilleurs modèles de la première phase d'évaluation et des modèles générés pour la deuxième phase d'évaluation pour le français.

(a) Tokens

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+KLO	0.906	0.931	0.918
CP+Sfreq	0.885	0.953	0.918
MD+Sfreq	0.891	0.944	0.917
Phase 2 : Recherche du meilleur modèle			
Sfreq+KLO+MD+CP	0.894	0.953	0.922
Sfreq+KLO+MD	0.886	0.947	0.916
CP+Sfreq+MD	0.874	0.960	0.915
Sfreq+CP+KLO	0.874	0.957	0.914
CP+KLO+MD	0.718	0.009	0.017

(b) UTE modérées

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
TT+Sfreq	0.752	0.961	0.844
Sfreq+ttest	0.748	0.962	0.841
PRS+Sfreq	0.740	0.969	0.839
Phase 2 : Recherche du meilleur modèle			
Sfreq+ttest+PRS+TT	0.770	0.968	0.858
TT+Sfreq+ttest	0.761	0.969	0.852
Sfreq+ttest+PRS	0.761	0.966	0.851
TT+Sfreq+PRS	0.757	0.958	0.846
TT+ttest+PRS	1.000	0.000	0.000

(c) UTE franches

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
MTD+Sfreq	0.685	0.972	0.803
Sfreq+Dfreq	0.679	0.966	0.797
PRS+Sfreq	0.678	0.967	0.797
Phase 2 : Recherche du meilleur modèle			
Sfreq+MTD+Dfreq+PRS	0.696	0.970	0.811
PRS+Sfreq+Dfreq	0.693	0.966	0.807
PRS+Sfreq+MTD	0.684	0.975	0.804
Sfreq+MTD+Dfreq	0.683	0.972	0.802
PRS+MTD+Dfreq	0.677	0.787	0.728

travers les langues à l'issue de la première phase d'évaluation (cf. 6.1 (p. 114) pour l'explication de ces traits). Les traits les plus fréquemment utilisés sont Sfreq (intégralité des langues), SIM (arabe, allemand, anglais, polonais et turc), CP (allemand, anglais, français) et SAL (polonais, turc).

On ne constate pas de similarité évidente entre les traits sélectionnés et les types de langue qui les utilisent. Il est toutefois possible de rapprocher les traits utilisés par l'allemand et l'anglais : leur seul trait discordant est le trait ttest pour l'allemand et Dfreq pour l'anglais. Ces deux traits ont toutefois la caractéristique commune de concerner des statistiques relatives aux corpus génériques. Polonais et turc partagent également les trois quart de leurs traits (Sfreq, SIM et SAL).

En ce qui concerne la deuxième phase (la recherche du meilleur modèle), cette expérience de référence obtient quasi-systématiquement les meilleurs résultats par rapport aux expériences

TABLE 8.6 – Résultats des trois meilleurs modèles de la première phase d'évaluation et des modèles générés pour la deuxième phase d'évaluation pour le polonais.

(a) Tokens

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+SIM	0.942	0.973	0.958
RCT+Sfreq	0.891	0.947	0.919
Sfreq+SAL	0.893	0.945	0.918
Phase 2 : Recherche du meilleur modèle			
Sfreq+RCT+SAL+SIM	0.951	0.975	0.963
SIM+Sfreq+RCT	0.940	0.979	0.959
SIM+Sfreq+SAL	0.938	0.976	0.956
Sfreq+RCT+SAL	0.855	0.957	0.903
SIM+RCT+SAL	0.535	0.703	0.608

(b) UTE modérées

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+SIM	0.702	0.969	0.814
ODR+Sfreq	0.709	0.953	0.813
PRS+Sfreq	0.706	0.950	0.810
Phase 2 : Recherche du meilleur modèle			
SIM+Sfreq+ODR	0.728	0.963	0.829
Sfreq+ODR+PRS+SIM	0.726	0.964	0.828
SIM+Sfreq+PRS	0.715	0.964	0.821
Sfreq+ODR+PRS	0.710	0.952	0.813
SIM+ODR+PRS	∅	∅	∅

(c) UTE franches

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
RCT+Sfreq	0.676	0.955	0.791
PRS+Sfreq	0.671	0.960	0.790
Sfreq+KLO	0.671	0.957	0.789
Phase 2 : Recherche du meilleur modèle			
Sfreq+KLO+PRS+RCT	0.678	0.958	0.794
RCT+Sfreq+KLO	0.667	0.964	0.788
RCT+Sfreq+PRS	0.668	0.959	0.787
Sfreq+KLO+PRS	0.656	0.966	0.782
RCT+KLO+PRS	0.723	0.004	0.009

menées sur des UTE sous-spécifiées. La langue obtenant le meilleur f-score est le turc $(0, 969)$ [1], suivi du polonais $(0, 963)$, de l'arabe $(0, 958)$, du français $(0, 922)$, de l'allemand $(0, 921)$, de l'anglais $(0, 906)$ et enfin, du chinois. Ce dernier affiche, contrairement aux autres langues, une précision bien supérieure au rappel, et un f-score de $0, 823$, considérablement inférieur à ceux des autres langues pour la même expérience. Comme nous le verrons, ce score sera plus comparable à ceux obtenus pour des modèles entraînés sur des UTE sous-spécifiées. Pour cette raison, bien que les modèles du chinois n'aient été entraînés que sur des tokens, nous allons comparer ses résultats aux autres langues pour les expériences menées sur les UTE modérées et franches.

L'expérience sur les UTE modérées a permis de réduire le nombre de traits utilisés à tra-

1. Deux modèles arrivent *ex æquo* pour l'experience sur les tokens : le modèle Sfreq+SIM et le modèle SIM +Sfreq+SAL. Dans ce cas là, et étant donné que notre méthode d'évaluation favorise déjà grandement le rappel, nous sélectionnons comme meilleur modèle final celui ayant la précision la plus élevée (Sfreq+SIM).

TABLE 8.7 – Résultats des trois meilleurs modèles de la première phase d'évaluation et des modèles générés pour la deuxième phase d'évaluation pour le turc.

(a) Tokens

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
Sfreq+SIM	0.991	0.948	0.969
Sfreq+SAL	0.954	0.944	0.949
FSS+Sfreq	0.951	0.943	0.947
Phase 2 : Recherche du meilleur modèle			
SIM+Sfreq+SAL	0.975	0.963	0.969
Sfreq+FSS+SAL+SIM	0.987	0.945	0.965
SIM+Sfreq+FSS	0.964	0.965	0.964
Sfreq+FSS+SAL	0.947	0.947	0.947
SIM+FSS+SAL	∅	∅	∅

(b) UTE modérées

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
DRK+Sfreq	0.884	0.973	0.927
Sfreq+KLO	0.887	0.970	0.926
JAC+Sfreq	0.883	0.972	0.926
Phase 2 : Recherche du meilleur modèle			
JAC+Sfreq+DRK	1	0.002	0.005
JAC+Sfreq+KLO	1.000	0.001	0.001
Sfreq+DRK+KLO+JAC	∅	∅	∅
Sfreq+DRK+KLO	∅	∅	∅
JAC+DRK+KLO	∅	∅	∅

(c) UTE franches

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
DRK+Sfreq	0.887	0.974	0.928
MD+Sfreq	0.882	0.974	0.926
Sfreq+FAG	0.879	0.975	0.925
Phase 2 : Recherche du meilleur modèle			
Sfreq+DRK+FAG+MD	0.888	0.959	0.922
MD+Sfreq+FAG	0.911	0.883	0.897
MD+Sfreq+DRK	∅	∅	∅
MD+DRK+FAG	∅	∅	∅
Sfreq+DRK+FAG	∅	∅	∅

TABLE 8.8 – Résultats des trois meilleurs modèles de la première phase d'évaluation et des modèles générés pour la deuxième phase d'évaluation pour le chinois sur les tokens informés.

Modèle	Précision	Rappel	F-score
Phase 1 : Sélection initiale des traits			
ZS+Sfreq	0.904	0.755	0.823
DRK+Sfreq	0.923	0.741	0.822
TT+Sfreq	0.918	0.738	0.818
Phase 2 : Recherche du meilleur modèle			
ZS+Sfreq+DRK	0.909	0.731	0.810
ZS+Sfreq+TT	0.915	0.702	0.794
Sfreq+DRK+TT	0.923	0.696	0.794
Sfreq+DRK+TT+ZS	∅	∅	∅
ZS+DRK+TT	∅	∅	∅

vers les langues de 15 à 11 à l'issue de la première phase d'évaluation, ce qui représente moitié moins de traits qu'originellement proposé. Le trait Sfreq est une fois de plus utilisé par tous les modèles dans toutes les langues, suivi des traits ODR (allemand, anglais, polonais), ttest (arabe, anglais, français), TT (arabe, français, chinois), ZS (arabe, anglais, chinois). Les autres traits ont été utilisés par deux langues ou moins. En particulier, les traits DRK (turc, chinois), PRS (français, polonais) et SIM (allemands, polonais).

On constate que le chinois, qui a l'exception du trait Sfreq, ne possédait aucun trait commun avec les autres langues pour l'expérience sur les tokens en possède à présent trois : un trait commun avec l'arabe et l'anglais, un trait commun avec l'arabe et le français, et un trait commun avec le turc. Toujours à l'exception de la fréquence en corpus de spécialité, l'allemand et l'anglais ne possède plus qu'un trait commun (ODR). Le turc et le polonais n'ont plus aucun trait commun. En revanche, le polonais et l'allemand en ont deux identiques (SIM et ODR).

Pour la deuxième phase de cette expérience (sélection des meilleurs modèles), le turc obtient une fois de plus le meilleur f-score $(0, 927)$. Ce dernier est même supérieur aux meilleurs f-scores obtenus pour le français, allemand et l'anglais sur les tokens. Puis viennent l'arabe, avec un f-score de $0, 879$, le français $(0, 858)$ le polonais $(0, 829)$, l'anglais $(0, 828)$ et l'allemand $(0, 731)$. Ici, le f-score de l'anglais est est comparable à celui obtenu pour le chinois sur les tokens ; en revanche, alors que le meilleur modèle obtenu pour l'anglais sur les UTE modérées favorise le rappel ($R = 0, 938$) sur la précision ($P = 0, 742$), c'est l'inverse en ce qui concerne le modèle chinois ($P = 0, 904$, $R = 0, 755$).

Par rapport à l'expérience sur les token, les langues dont les scores ont été les plus détériorés sont l'allemand ($-0, 2$ points) et le polonais ($-0, 134$ points). À l'inverse, les différences de scores les plus petites ont été obtenues pour le turc ($-0, 042$ points) et le français ($-0, 064$ points). Ces différences sont certainement non-significatives, même si un calcul de significativité est délicat à concevoir ici, étant donnée les caractéristiques de notre mesure d'évaluation.

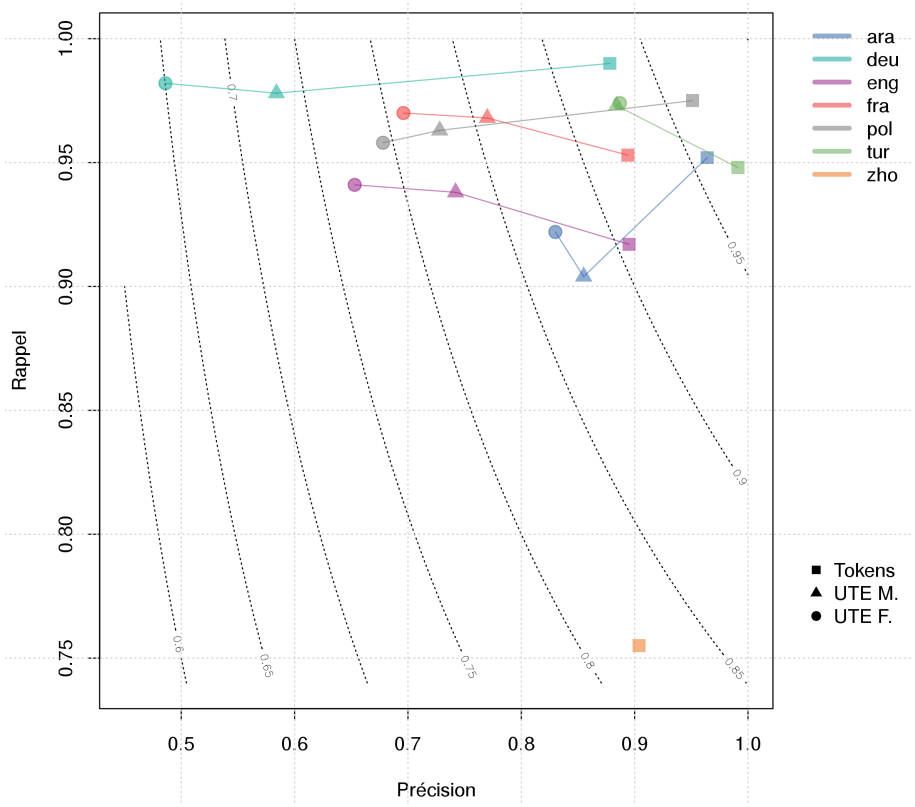
Pour la dernière expérience utilisant des unités maximalement sous-spécifiées, la première phase d'évaluation a, à l'instar de l'expérience de référence, permis de sélectionner au total 15 traits pour l'intégralité des langues concernées. Il est intéressant de noter que la réduction de 15 à 11 traits obtenus lors du passage des tokens informés à des UTE modérées n'a pas persisté, et que ce degré supplémentaire de sous-spécification a au contraire abouti à moins de « généricité ». Malgré tout, pour l'expérience de référence parmi les 15 traits retenus, le nombre de traits utilisés par une seule langue était de 11 (ttest, Dfreq, KLO, MD, RCT, TT, DRK, ZS, FSS, BB, FKZY). Pour cette expérience, ils ne sont plus que neuf (SIM, ODR, FAG, MD, ZS, TT, KLO, RCT, CP).

Comme pour les précédentes expériences, le trait utilisé universellement par tous les modèles et le trait Sfreq. Les autres traits les plus populaires à travers les langues sont les traits Dfreq

(allemand, anglais, français), PRS (arabe, français, polonais), ttest (arabe, allemand), DRK (turc, chinois) et MTD (anglais, français).
Pour l'expérience sur les tokens, turc et polonais partageaient la majorité de leurs traits. Cette tendance a totalement disparu pour les expériences sur les unités sous-spécifiées. En revanche, alors que le turc n'avait qu'un trait en commun avec le chinois sur les tokens (Sfreq), il a en plus le trait DRK en commun avec le chinois pour les expériences sur les UTE modérées et franches. C'est d'ailleurs le seul trait que le turc partage avec une autre langue pour ces deux dernières expériences. De la même manière, le français et le polonais conservent sur les deux dernières expériences le trait PRS.
Les f-scores issus de la deuxième phase pour cette expérience placent une fois de plus le turc en première position $(0, 928)$. Il est à noter que le turc est le seul exemple de langue dont le meilleur modèle obtenu pour une sous-spécification franche obtient un score supérieur au meilleur modèle obtenu pour une sous-spécification modérée ($+0, 001$ point ; cette différence n'est certainement pas suffisante pour être significative). Viennent ensuite l'arabe $(0, 874)$, le français $(0, 811)$, le polonais $(0, 794)$, l'anglais $(0, 771)$ et l'allemand et $(0, 65)$. L'allemand et le polonais sont encore les langues dont les f-scores souffrent le plus de cette sous-spécification : respectivement $-0, 281$ et $-0, 169$ points par rapport à l'expérience de référence. Les langues tolérant le mieux la sous spécification franche sont le turc et l'arabe, avec respectivement $0, 041$ et $0, 084$ points en moins par rapport à l'expérience de référence sur les tokens, mais une fois encore ces différences sont probablement non significatives.

La figure 8.2 permet de visualiser les différences de scores entre les meilleurs modèles de chaque langue et pour chaque cadre expérimental. On constate en premier lieu que le meilleur modèle chinois obtient un rappel bien inférieur à celui des autres langues, tous modèles confondus. Son f-score est en apparence comparable à ceux des modèles modérément sous-spécifiés de l'anglais et du polonais. En ce qui concerne l'évolution des scores au regard du degré de sous-spécification, toutes les langues ne sont pas également lésées. C'est pour la langue allemande que les étapes de sous-spécification provoquent les effets les plus délétères. Le polonais, l'anglais, l'arabe et le français subissent également une dégradation considérable de leur précision. Pour l'anglais et le français cependant, on observe que cette perte drastique de précision s'accompagne malgré tout d'un modeste gain de rappel. On observe le même phénomène avec le turc. Ce sont les trois langues de notre échantillon pour lesquelles on peut concevoir que la sous-spécification, au sens lexical, a effectivement fonctionné. Le turc tolère relativement bien la sous-spécification par rapport aux autres langues, mais voit tout de même son f-score et sa précision diminuer.

FIGURE 8.2 – Visualisation des meilleurs modèles entraînés (et appliqués) sur les tokens, les UTE modérées et franches pour les sept langues de notre échantillon de test.



Ces résultats, ainsi que l'influence de notre méthode d'évaluation sur les scores, seront discutés à la section 8.3.

8.2 Test sur la portabilité des modèles entre langues

La section précédente a été l'occasion de regarder dans quelle mesure des modèles entraînés sur des corpus dans une langue donnée étaient capables de retrouver des termes dans cette même langue. Or, le but de ce travail consiste également à investiguer les possibilités d'utiliser des modèles entraînés dans une langue pour proposer des termes dans une autre langue. À cette fin, nous avons mené une série d'expériences en appliquant des modèles entraînés sur une langue, dite la « langue image », sur une autre langue, dite la « langue support ». Pour poursuivre sur cette métaphore artistique, nous désignerons par « modèle calque » le meilleur modèle de la langue image (déterminé dans la section 8.1), et par « modèle contre-épreuve » le modèle entraîné sur la langue image avec une combinaison de traits identique à celle du meilleur modèle de la langue support (également determiné dans la section 8.1). Nous avons appliqué pour chaque langue de notre échantillon de test, et pour chaque alternative de sous-spécification (tokens, UTE modérées et UTE franches), les différents modèles calque et contre-épreuve entraînés sur chacune des autres langues. Un modèle entraîné sur un corpus pour un cadre expérimental donné est appliqué sur un corpus ayant un degré de sous-spécification équivalent, sauf pour le chinois.

L'intégralité des f-scores ainsi obtenus sont présentés en annexe C (p. 271). Les moyennes des meilleurs modèles calques et contre-épreuve pour chaque couple langue image–langue support sont données en marge des tableaux 8.9 (a), (b) et (c). Les résultats des meilleurs modèles sont représentés pour chaque langue support dans les figures 8.3 à 8.9. Ces figures situent les scores des différents modèles appliqués à une langue support : modèles calques et contre-épreuves pour les trois cadres de sous spécification possibles, entraînés sur chaque langue image. À titre informatif, les scores des meilleurs modèles entraînés et appliqués sur la langue support ont également été inclus dans ces graphes (symboles astérisques). Pour chaque figure, l'axe des abscisses correspond à la précision, l'axe des ordonnées correspond au rappel, et les lignes en pointillés indiquent des iso-ligne de f-score. Les lettres superposées aux symboles correspondent aux degrés de sous spécification du modèle : « F » indiquant une sous-spécification franche, « M » désignant une sous-spécification modérée et « B » (pour « *Baseline* ») faisant état d'aucune sous spécification.

Une fois de plus, nous constatons que, quelles que soient les langues supports et les langues images impliquées, les modèles obtenant les meilleures f-scores sont systématiquement des modèles entraînés et appliqués sur des corpus n'ayant pas bénéficié de sous-spécification. Pour cette raison, lorsque nous parlerons par la suite d'un modèle sans préciser si il a été entraîné sur des tokens, des UTE modérées ou des UTE franches, il faudra entendre qu'il s'agit d'un modèle non sous-spécifié.

Concernant les résultats, il est intéressant de noter qu'il est possible d'obtenir pour toutes les

TABLE 8.9 – Proportions de termes complexes obtenus par les meilleurs modèles (calques et contre-épreuve) entraînés sur différentes langues pour les trois cadres expérimentaux et appliqués sur toutes les langues support. Ces proportions (en %) sont accompagnées des f-scores moyens obtenus par l'ensemble des modèles considérés et pour l'ensemble des langues.

(a) Tokens

	Prop. de termes complexes pour les langues supports							*Prop.*	f-score	précision	rappel
Modèles	ara	deu	eng	fra	pol	tur	zho	*moy.*	moy.	moy.	moy.
ara		0.77	0.00	0.00	0.00	0.05	55.16	*9.33*	0.85	0.83	0.86
deu	21.06		37.89	29.10	10.77	14.53	66.45	*29.97*	0.85	0.83	0.89
eng	24.25	30.95		13.03	11.90	13.51	57.75	*25.23*	0.81	0.84	0.80
fra	25.46	34.25	37.29		15.51	15.57	49.94	*29.67*	0.84	0.86	0.83
pol	16.41	19.88	32.09	23.90		11.34	34.28	*22.98*	0.86	0.87	0.86
tur	6.82	7.56	12.01	10.30	4.22		22.12	*10.51*	0.94	0.94	0.95
zho	18.95	29.98	20.43	34.34	15.19	24.55		*23.91*	0.66	0.84	0.59
Prop. moy	*18.82*	*20.57*	*23.29*	*18.45*	*9.60*	*13.26*	*47.62*				
f-score moy	0.89	0.90	0.87	0.86	0.88	0.89	0.54				
précision moy.	0.98	0.88	0.89	0.84	0.91	0.96	0.54				
rappel moy.	0.83	0.92	0.87	0.90	0.87	0.84	0.55				

(b) UTE modérées

	Prop. de termes complexes pour les langues supports							*Prop.*	f-score	précision	rappel
Modèles	ara	deu	eng	fra	pol	tur	zho	*moy.*	moy.	moy.	moy.
ara		10.88	12.05	14.30	7.91	10.46	46.28	*16.98*	0.82	0.74	0.94
deu	4.33		6.78	12.90	2.33	1.21	12.85	*6.73*	0.77	0.67	0.92
eng	10.67	14.71		22.55	17.92	12.08	54.40	*22.06*	0.75	0.72	0.80
fra	8.32	5.66	9.04		7.10	3.88	44.05	*13.01*	0.78	0.75	0.83
pol	10.45	4.86	6.83	7.87		1.93	11.35	*7.21*	0.58	0.51	0.70
tur	2.87	5.12	5.59	9.00	3.64		31.79	*9.67*	0.85	0.76	0.97
zho	15.83	19.74	12.46	18.40	39.15	15.93		*20.25*	0.56	0.49	0.66
Prop. moy	*8.74*	*10.16*	*8.79*	*14.17*	*13.01*	*7.58*	*33.45*				
f-score moy	0.73	0.72	0.75	0.75	0.77	0.66	0.73				
précision moy.	0.68	0.61	0.65	0.64	0.68	0.63	0.75				
rappel moy.	0.81	0.89	0.89	0.93	0.88	0.71	0.70				

(c) UTE franches

	Prop. de termes complexes pour les langues supports							*Prop.*	f-score	précision	rappel
Modèles	ara	deu	eng	fra	pol	tur	zho	*moy.*	moy.	moy.	moy.
ara		5.09	14.74	17.30	7.14	2.36	46.11	*15.46*	0.81	0.71	0.96
deu	5.19		6.11	5.71	4.72	1.33	8.08	*5.19*	0.59	0.48	0.77
eng	5.70	4.98		11.82	8.88	3.98	37.21	*12.09*	0.58	0.52	0.66
fra	8.00	3.30	15.50		6.27	5.24	29.75	*11.34*	0.61	0.56	0.68
pol	9.00	5.12	11.21	9.63		6.41	25.53	*11.15*	0.56	0.48	0.70
tur	6.46	5.84	17.82	19.54	9.44		17.20	*12.72*	0.84	0.74	0.98
zho	9.35	12.44	18.30	18.07	21.59	20.98		*16.79*	0.54	0.46	0.68
Prop. moy	*7.28*	*6.13*	*13.95*	*13.68*	*9.67*	*6.72*	*27.31*				
f-score moy	0.51	0.69	0.73	0.71	0.75	0.57	0.56				
précision moy.	0.47	0.56	0.60	0.58	0.64	0.53	0.57				
rappel moy.	0.58	0.90	0.94	0.94	0.90	0.63	0.55				

langues de notre échantillon de tests des f-scores équivalents en utilisant des modèles entraînés sur d'autres langues. En chinois par exemple, le modèle contre-épreuve entraînée sur l'arabe obtient un score similaire au meilleur modèle entraîné sur le chinois, bien que ce modèle favorise le rappel sur la précision, à l'inverse du modèle chinois. En allemand, il existe même huit modèles, dans quatre langues différentes qui surpassent le meilleur modèle entraîné sur l'allemand.

Pour la langue support arabe (fig. 8.3), les deux seuls modèles à dépasser l'iso-ligne de f-score 0.95 ont été obtenus pour le turc (modèle calque et contre-épreuve). Le modèle contre-épreuve turc surpasse même le f-score du meilleur modèle arabe. Il est intéressant de noter que la proportion de termes complexes extraits par ces deux modèles est de 6, 82%, largement inférieure à la proportion moyenne de termes complexes extraites par l'ensemble des modèles non sous-spécifiés appliqués sur l'arabe (18, 82%).

Concernant l'allemand (fig. 8.4), les meilleurs modèles sont issus du turc, de l'arabe, du polonais et de l'anglais. Parmi ces derniers, quatre sont des modèles calques et trois sont des modèles contre-épreuves. Ils obtiennent globalement un f-score plus élevé que celui du meilleur modèle allemand. À l'instar de l'arabe, les modèles non sous-spécifiés du turc et de l'arabe extraient très peu de termes complexes au regard des autres modèles. Les modèles polonais et anglais produisent plus de termes complexes.

Les différentes expériences menées sur l'anglais (fig. 8.5) semblent indiquer qu'il est difficile de dépasser la barre de 0, 9 points de f-score dans cette langue avec notre approche : le meilleur modèle reste celui entraîné sur l'anglais. Puis viennent les modèles polonais, allemand, turc et français (au total, autant de modèles calques que de modèles contre-épreuve). Tous ces modèles extraient en moyenne 23, 29% de termes complexes, ce qui représente, en proportions, près de 3 fois le nombre de termes complexes extrait par le meilleur modèle anglais non sous-spécifié.

Pour le français (fig. 8.6), quatre modèles calques obtiennent des scores très comparables au meilleur f-score obtenu par le modèle entraîné sur le français : le polonais, le turc, l'allemand et le chinois. Ces modèles sont ceux qui proposent le plus de candidats termes complexes, exception faite du chinois.

L'ensemble des modèles des autres langues entraînés sur des tokens pour être appliqués sur le polonais produisent en moyenne moins de termes complexes que lorsque ces modèles sont appliqués sur d'autres langues (fig. 8.7). Les expériences dont la langue support est le polonais ont permis de déterminer que le meilleur modèle entraîné sur le polonais était également difficile à égaler ou surpasser en termes de f-score : seuls un modèle turc et un modèle français y parviennent.

Ce n'est pas le cas en ce qui concerne la langue support turque (fig. 8.8), pour laquelle au-

cun modèle entraîné sur une autre langue n'a permis de surpasser celui obtenu par le meilleur modèle entraîné sur le turc. Toutefois, parmi les modèles issus d'autres langues appliquées au turc, celui du chinois (contre-épreuve), du français (contre-épreuve) et les modèles calques et contre-épreuves du polonais et de l'arabe parviennent à un f-score supérieur à $0,95$. Fait remarquable, les modèles entraînés sur le turc étant issus de sous-spécification franche (f-score moyen : $0,84$) et modérée (f-score moyen : $0,85$) obtiennent des f-scores comparables à ceux des modèles non sous-spécifiés obtenus pour des langues images comme l'allemand ou l'anglais.

Enfin, pour le chinois (fig. 8.9), seul le modèle contre-épreuve non sous-spécifié de l'arabe obtient un f-score comparable à celui originairement obtenu par le meilleur modèle chinois. Contrairement aux autres langues pour lesquelles on observait un regroupement des modèles en fonction de leur degré de sous spécification et, éventuellement, de leur langue image, on constate ici que les modèles exhibent des écarts de rappel et de précision plus important. On note une tendance des modèles non sous-spécifiés à favoriser la précision plutôt que de le rappel, comme c'est globalement le cas pour le reste des modèles sous spécifié dans les autres langues. De façon intéressante, deux modèles issus de sous-spécification modérée (contre-épreuve turc et français) surpassent, en termes de f-score, des modèles non sous-spécifié. En moyenne, le chinois est la seule langue à bénéficier de modèles calques et contre épreuves modérément sous-spécifiés. Cette différence peut s'observer entre les tableaux 8.9 (a) et (b) : de $0,54$ à $0,73$ points de f-score, la sous-spécification semble faire gagner près de 20 points de f-score à la moyenne des modèles. Cette différence est causée par des modèles non sous-spécifiés ne produisant aucun candidat terme (f-score nul). En revanche, cette différence est à nuancer, car les modèles non sous-spécifiés produisant des scores non nuls sont en moyenne meilleurs que les modèles calques et contre-épreuve sous-spécifiés modérément. Le chinois est la langue qui, quels que soient les cadres expérimentaux, produit la plus grande quantité de candidats termes complexes.

En observant pour chaque langue les trois meilleurs modèles calques et contre-épreuves on constate en ce qui concerne les modèles calques, que ceux obtenant les meilleurs f-scores sont majoritairement issus du turc, et dans une moindre mesure du polonais, de l'arabe et de l'anglais. Le chinois et le français arrivent rarement à produire des modèles compétitifs. Pour les modèles contre-épreuves, il n'y a pas de langue en particulier qui soit favorisée comme le turc l'était pour les modèles calques. Néanmoins, le turc, l'arabe et le français produisent la plupart des meilleurs modèles contre-épreuve. Plus globalement, les f-scores moyens présentés dans les tableaux 8.9) indiquent que les modèles entraînés sur le turc (et dans une moindre mesure, sur l'arabe) surpassent globalement l'ensemble des autres modèles, quel que soient les cadre expérimentaux.

FIGURE 8.3 – Comparaison des meilleurs modèles *calques* et *contre-épreuve* sur la langue support arabe.

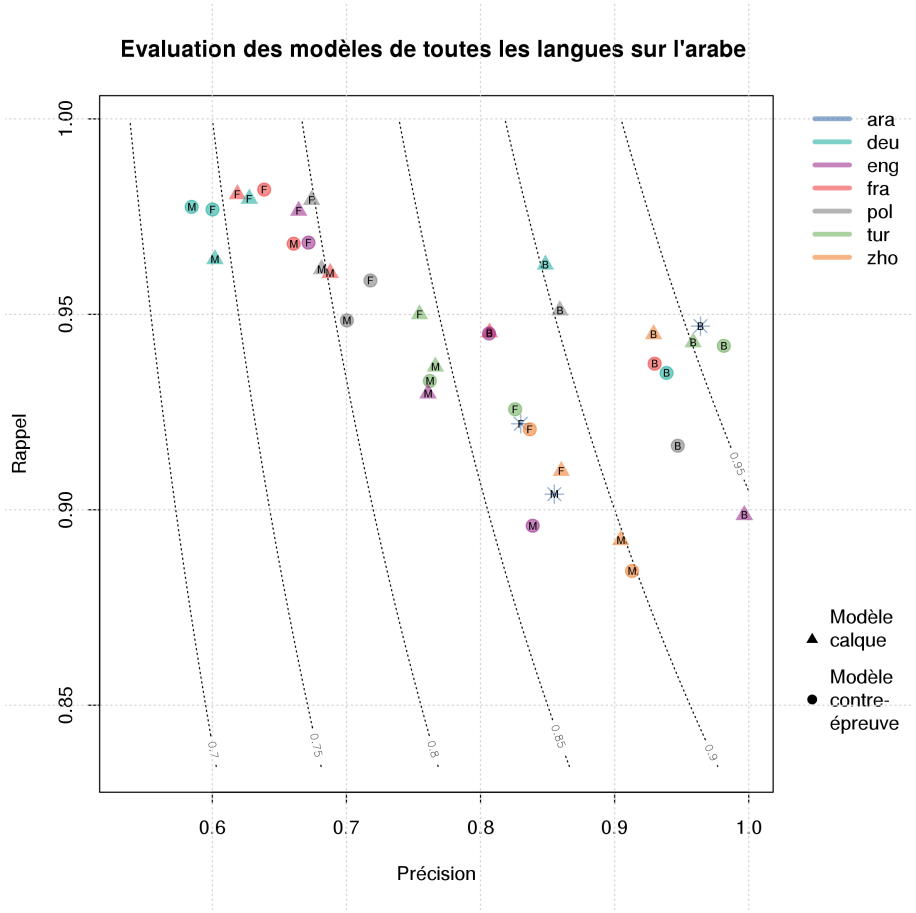


FIGURE 8.4 – Comparaison des meilleurs modèles *calques* et *contre-épreuve* sur la langue support allemande.

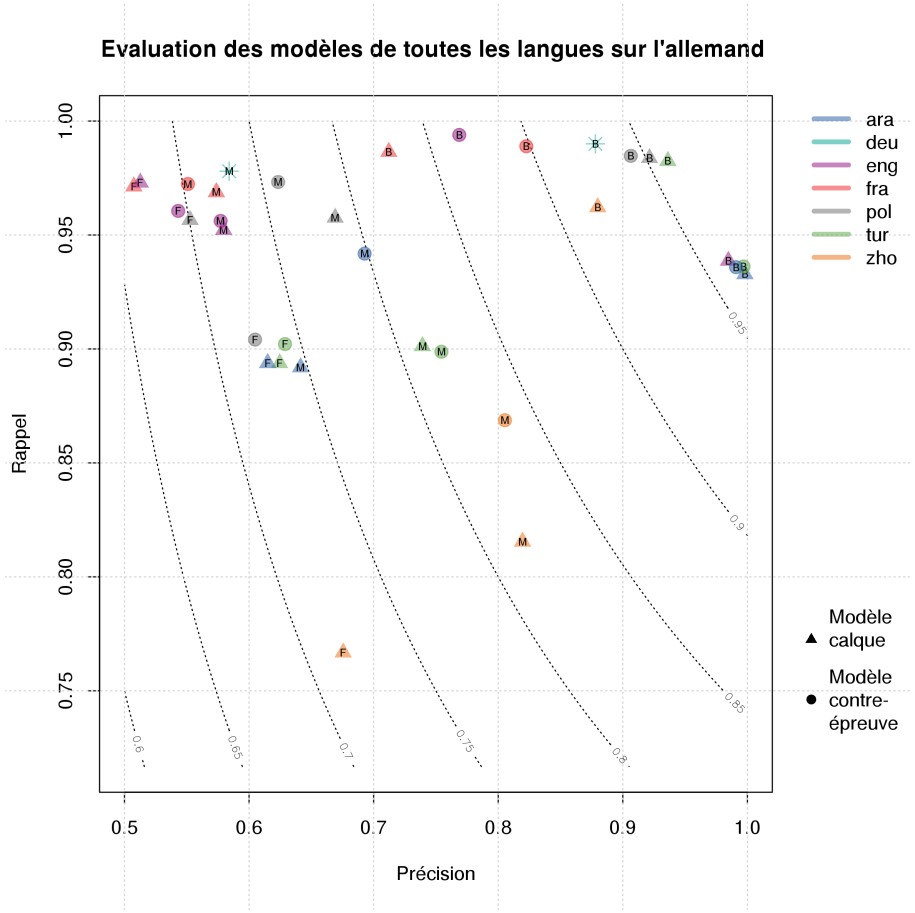


FIGURE 8.5 – Comparaison des meilleurs modèles *calques* et *contre-épreuve* sur la langue support anglaise.

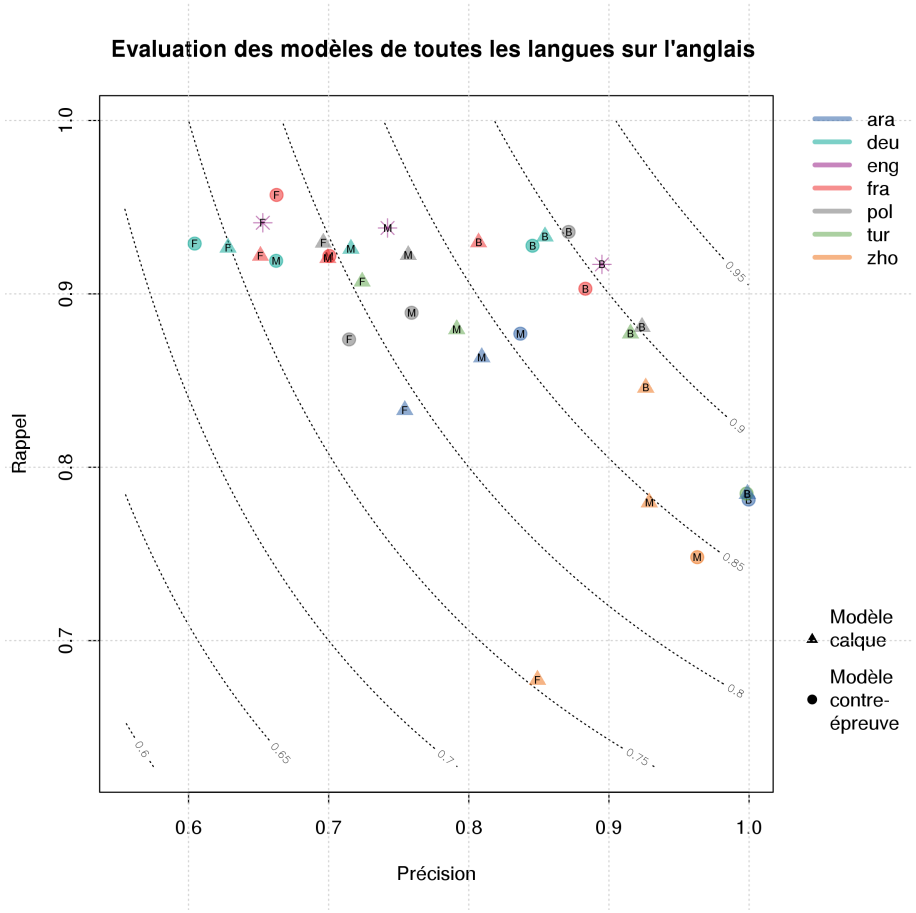


FIGURE 8.6 – Comparaison des meilleurs modèles *calques* et *contre-épreuve* sur la langue support française.

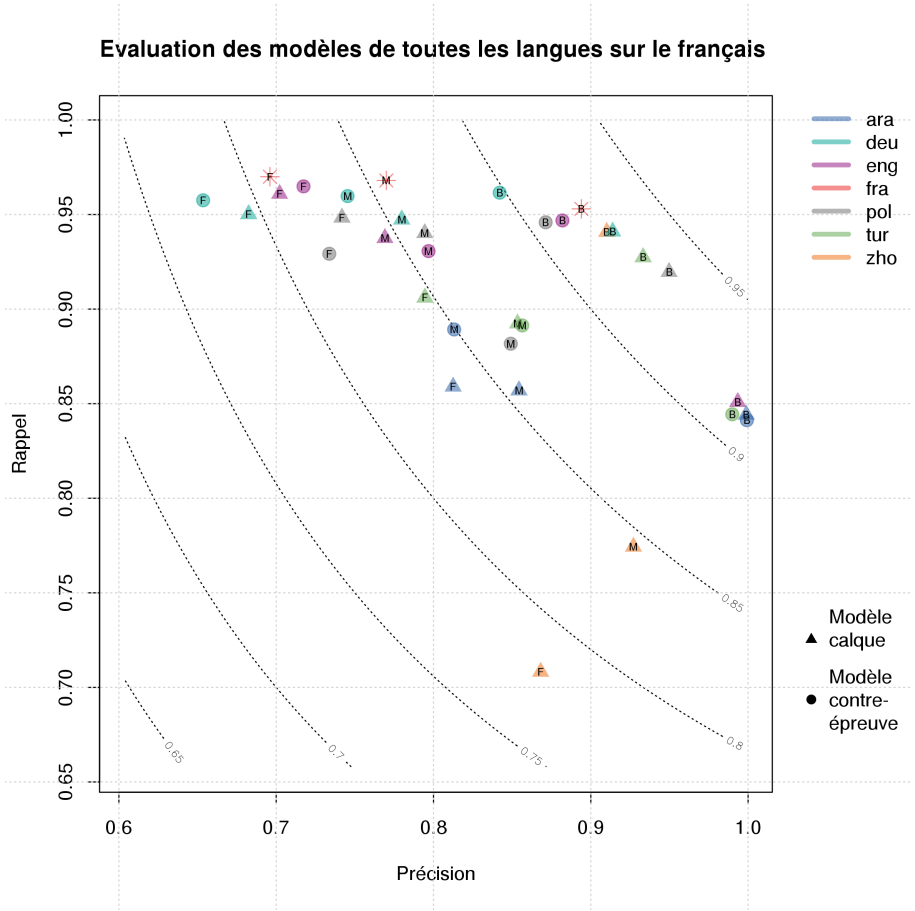


FIGURE 8.7 – Comparaison des meilleurs modèles *calques* et *contre-épreuve* sur la langue support polonaise.

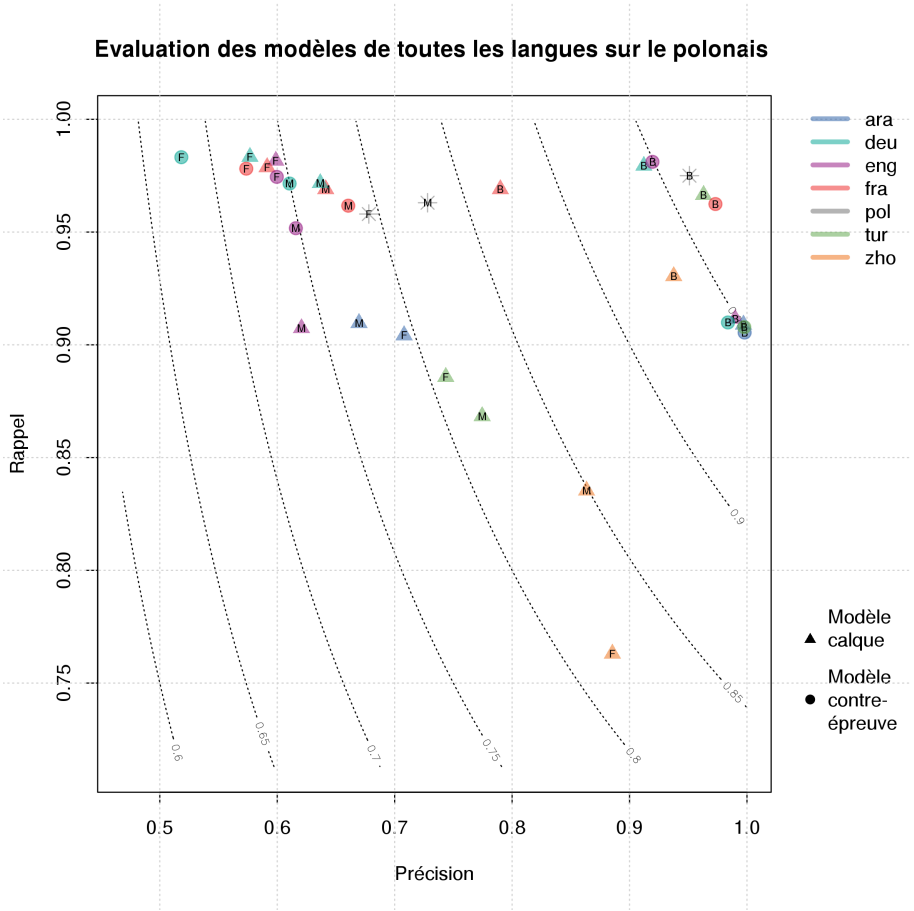


FIGURE 8.8 – Comparaison des meilleurs modèles *calques* et *contre-épreuve* sur la langue support turc.

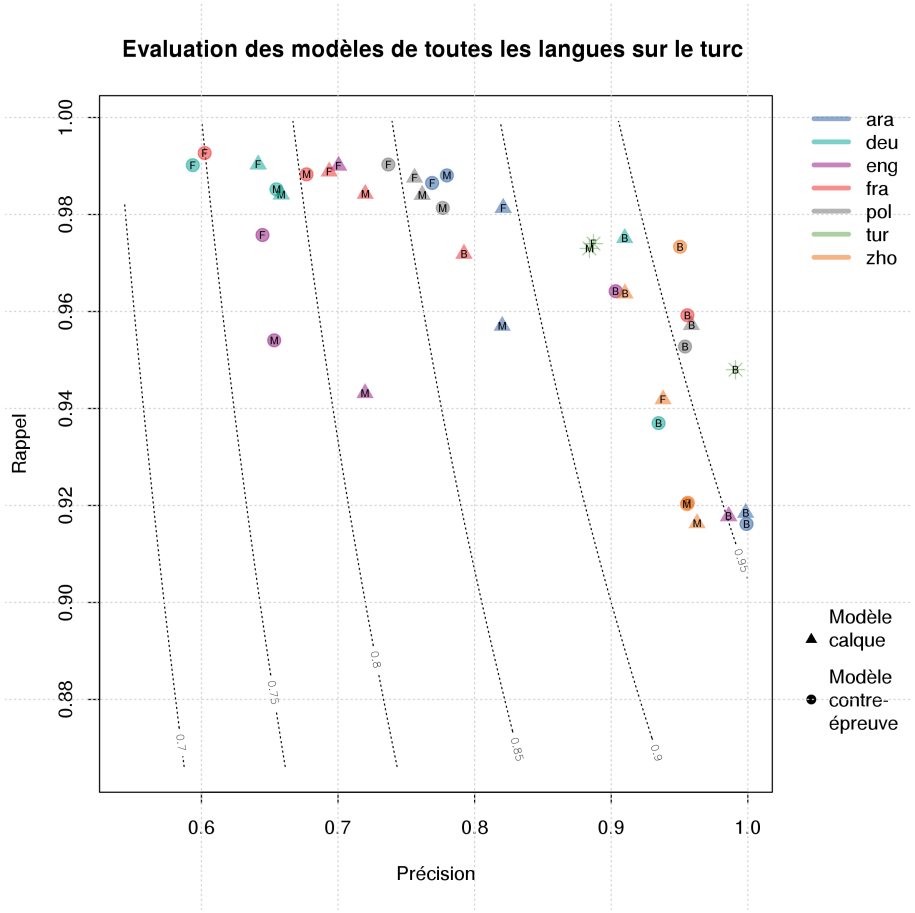
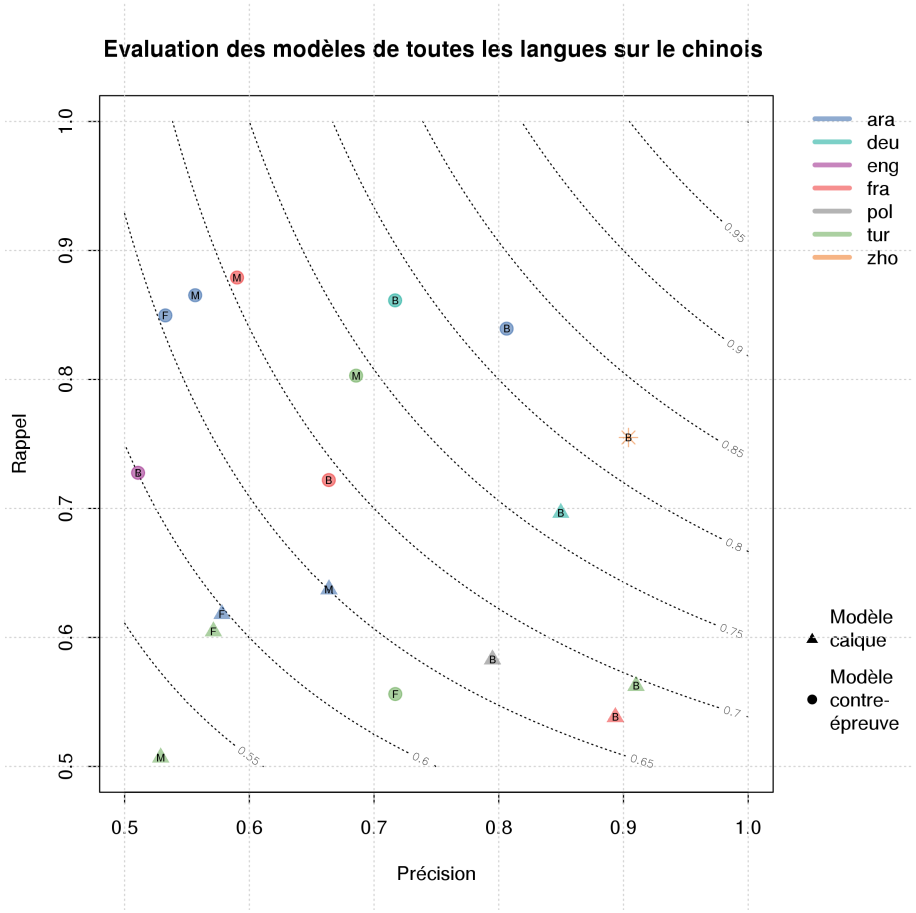


FIGURE 8.9 – Comparaison des meilleurs modèles *calques* et *contre-épreuve* sur la langue support chinoise.



8.3 Discussion

Le fil conducteur de notre recherche a consisté à étudier l'influence de différents paramètres de sous-spécification lexicale sur l'extraction de « terme de terrain » (c'est-à-dire de termes non normatifs) dans des langues typologiquement différentes, sans utiliser d'autres connaissances linguistiques que des informations typologiques.
En particulier, nous nous sommes interrogés sur la possibilité d'obtenir de meilleurs résultats ou d'améliorer la migration d'un modèle vers d'autres langues en rognant de l'information morphologique.

Pour l'ensemble des modèles entraînés, la qualité des termes extraits a pu être évaluée grâce à une version modifiée de l'algorithme de Nazarenko *et al.* (2009). Les meilleurs scores obtenus pour l'ensemble des langues sont globalement satisfaisants.
Bien entendu, les scores donnés dans le présent chapitre ne peuvent être comparés à ceux de la littérature qu'à titre indicatif, car les critères de choix de termes et les corpus peuvent faire varier les scores de façon drastique. À titre comparatif, les scores de références donnés par Nazarenko *et al.* (2009) ont été obtenus par trois systèmes sur un corpus (en anglais) recouvrant des domaines comme la biologie moléculaire et la génomique constitué d'environ 405 000 mots. Le système *Nomino* (David & Plante, 1990) obtient un f-score de 0.5, *ACABIT* (Daille, 1994) produit un f-score de $0,63$ et le f-score de *Syntex* (Bourigault & Fabre, 2000) atteint $0,8$. Pour ce qui est des scores relatés par Mondary *et al.* (2012), le meilleur f-score obtenu ($0,358$) l'est par *MIG-YaTeA* (version améliorée de YaTeA (Aubin & Hamon, 2006)) sur un corpus de résumés d'articles scientifiques en pharmacologie en anglais. Sur le même corpus et dans les mêmes conditions d'évaluation, *ACABIT* obtiens un f-score de $0,289$.
Pour toutes les langues à l'exception du chinois, nous parvenons à obtenir des modèles produisant des résultats dont le score est supérieur à $0,9$. Ces scores ont notamment bénéficié du biais positif conféré au rappel par notre algorithme, comme cela à été évoqué à la page 139. Pour le chinois, le f-score des meilleurs modèles ne franchissent pas l'iso-ligne $0,85$. Même s'il est impossible d'en avoir la certitude absolue, les comparaisons avec les scores produits par l'outil *Termometer* (c.f. tableau 7.5 p. 140) nous permettent d'estimer que nos meilleurs modèles CRF sont aptes à réaliser une extraction terminologique de niveau état de l'art pour l'anglais au moins, lorsque la nature des termes n'est pas trop contrainte.

Concernant la fiabilité des scores obtenus, nous formulons des remarques à deux égards. En premier lieu, il est difficile d'estimer à quel point les différences de scores d'une langue à l'autre sont dues à des artefacts de la méthode d'évaluation plutôt qu'à des imperfections des modèles. La figure 8.1 permet de distinguer différentes strates de scores correspondant plus ou

moins clairement à des langues différentes. Cette disposition des scores évoque justement des disparités de traitement pour le calcul des scores. On suppute que le turc, langue obtenant les meilleurs scores, est également la langue la mieux traitée par l'outil *Morfessor*. Au contraire, la gestion de la composition concaténative en allemand a sans doute occasionné une diminution plus ou moins artificielle des scores de précision des modèles. Le fait de se passer de cette marge morphologique pour le calcul des distances de Levenshtein en chinois éclaire également la distribution si caractéristique des ses scores.
En second lieu, il est intéressant de considérer les terminologies de référence utilisées pour l'entraînement des modèles sur les différentes langues. Ces dernières ne sont pas homogènes, aussi bien sur la qualité que sur la quantité. Ce paramètre a une importance non négligeable pour l'entraînement des modèles comme pour l'évaluation des termes extraits. Le tableau 6.3 (p. 115) indiquait notamment que les terminologies du chinois simplifié et de l'arabe disposaient de moins de la moitié du nombre de termes recensés pour le français. Néanmoins, reportés sur nos corpus d'évaluation avant rééquilibrage, ces termes donnent plus de 27% de classes positives (étiquettes B, I, L, U confondues) pour le chinois, alors que ce pourcentage tombe à 17,6% pour le petit corpus de l'arabe (voir le tableau 7.2 p. 120). Cela aura une incidence sur le nombre d'exemples disponibles pour l'entraînement des modèles après rééquilibrage. La langue ayant le moins d'exemples positifs avant réequilibrage est néanmoins l'allemand (85,46% d'exemples négatifs), avec en particulier une proportion dérisoire d'étiquettes balisant des termes complexes. Cela permet d'expliquer partiellement la faible proportion de termes complexes extraits pour l'allemand (tableau 8.1). Le polonais a également peu d'exemples d'étiquettes B, I et L, ce qui se répercute sur la proportion de termes complexes extraits.

Relativement aux traits privilégiés par les meilleurs modèles, le fait que le trait Sfreq apparaisse systématiquement corrobore partiellement les constatations de Wermter & Hahn (2006), dont les expériences ont suggéré que l'utilisation de mesures probabilistes, même complexes, n'apporte aucune plus-value pour l'identification de termes (ou, dans une moindre mesure, pour l'extraction de collocations) par rapport à un simple compte de fréquence, à moins d'incorporer des informations linguistiques, comme par exemple le font les mesures LSM (Wermter & Hahn, 2004) et LPM (Wermter, 2009). Seul le trait USUB n'apparaît dans aucun des meilleurs modèles quel que soit le cadre expérimental. Pour ce qui est des autres traits, nous ne nous risquerons pas à extrapoler des tendances en fonction des langues et des cadres expérimentaux. Nous nous contenterons de souligner le fait que l'expérience sur les UTE modérées aura permis, dans la section 8.1 de diminuer (de 15 à 11) l'ensemble des traits sélectionnés après la phase 1. La question de savoir si cela est le reflet d'une meilleure généricité

des modèles restera en suspend, faute de preuves supplémentaires.

Ce que l'on peut affirmer en revanche au vu de nos résultats est que la sous-spécification n'a pas produit les bénéfices escomptés : quelles que soient les expériences, les meilleurs scores ont été obtenus majoritairement lorsque toute l'information morphologique a été conservée pour l'entraînement des modèles. Conformément à nos prévisions, cette sous-spécification permet pour la plupart des langues de maximiser le rappel. Malheureusement, cela se fait au détriment de la précision. La sous-spécification n'a pas non plus été avantageuse pour appliquer des modèles entraînés sur des langues morphologiquement riches à des langues ayant un panel de phénomènes morphologiques plus réduit. Dans le tableau 8.9 (b), les modèles appliqués au chinois paraissent améliorer le f-score moyen par rapport à celui indiqué dans le tableau 8.9 (a) ; comme cela a déjà été évoqué, cette différence de score, due à plusieurs modèles non sous-spécifiés stériles (voir les lignes concernant le chinois dans les tableaux de l'annexe C), est à nuancer. Ces derniers résultats semblent confirmer le fait qu'une partie substantielle de l'information importante pour l'extraction terminologique est située dans la morphologie : plus les unités sont sous-spécifiées, moins les résultats des expériences sont satisfaisants. Cette interprétation des résultats est toutefois à nuancer pour le turc, qui semble tolérer mieux que les autres langues d'être amputé d'une partie de son information morphologique. Cette caractéristique est probablement due à la richesse morphologique du turc (dont un corollaire est une grande dispersion de la distribution lexicale), ce qui permet aux modèles sous-spécifiés de se maintenir à des niveaux de performances acceptables.

Allemand et polonais sont les langues dont les modèles supportent le moins bien cette sous-spécification. Nous apportons trois explications possibles à cela. La première est que ces deux langues ont, comparativement aux autres, un ordre des mots relativement libres, surtout en ce qui concerne le polonais. Supprimer de l'information morphologique revient également à supprimer de l'information syntaxique. La seconde explication, en ce qui concerne l'allemand, et relative au traitement de la compositionnalité concaténative. Ce procédé de composition, non binaire et très productif, est probablement mal pris en charge par notre heuristique de *pseudo-racinisation* (voir la section 5.2.2). La troisième explication concerne l'effondrement du taux de termes complexes extraits par les modèles pour ces langues, passant par exemple de près de 30% (tokens) à 5, 2% (UTE franches) pour les modèles allemands). Cette dernière explication vaut également, dans une moindre mesure, pour l'ensemble des langues sous-spécifiables : appliquée à des termes simples, la suppression de bribes morphologiques peut faire glisser le sens d'un terme plus du côté du flou que de la sous-spécification. Ce dernier voit son incertitude augmenter (section 3.4, p. 60). Pour illustrer ce problème de flou prenant le pas sur la sous-spécification modérée en français, voici deux exemples d'UTE modérées fré-

quentes en corpus :

— l'UTE ⊧vis (1009 occurrences en corpus de spécialité) regroupe les termes vis, visage(s), visaient, visât, visez, visibilité(s), visière, visiste [2], vison, visqueuses, visqueux et visserie ;

— l'UTE ⊧ban (1172 occurrences en corpus de spécialité) sous-spécifie des termes aussi hétérogènes que « bandit(s), banières [3], banque(s), banquer, banquet(s), banquette(s), ban(s) et banville ».

En ce qui concerne le traitement du chinois, nous nous serions attendus à ce que ce soit les modèles entraînés sur de l'anglais, ou toute autre langue SVO sous-spécifié modérément, qui obtiennent les meilleurs scores. Or le chinois tolère mieux des modèles non sous-spécifiés issus de l'allemand ou de l'arabe, du turc (modèle modérément sous-spécifié) et éventuellement du polonais : toutes ces langues exhibent non seulement des phénomènes morphologiques complexes et variés, et ne sont qui plus est pas SVO [4].

Qu'en est-il des modèles entraînés sur le chinois ? Nous avons déjà fait allusion au fait que ces derniers sont très peu prisés par les autres langues. Ce qui est intéressant en revanche, c'est que le seul modèle du chinois parvenant à surpasser l'ensemble des autres modèles est le modèle contre épreuve appliqué sur le corpus *non sous-spécifié*... du turc. Le même modèle appliqué aux corpus sous-spécifiés du turc obtient des scores sensiblement moins bons (en dessous de l'iso-ligne marquant $0,95$ de f-score).

Lors des expériences de portabilité des modèles d'une langue à une autre, nous avons constaté que, pour les modèles calques en particulier, la langue image la mieux tolérée était le turc (meilleur modèle pour le polonais et l'arabe, deuxième meilleur modèle pour le chinois l'anglais et le français, et troisième meilleur modèle pour l'allemand). Viennent ensuite le polonais (meilleur modèle pour l'anglais et le français, deuxième meilleur modèle pour le turc et troisième meilleur modèle pour le chinois) et l'arabe (meilleur modèle pour l'allemand et le turc, deuxième meilleur modèle pour le polonais). Cela peut signifier que la morphologie est un atout, même pour des langues qui n'en ont pas.

Ces résultats, très contre intuitifs au regard de nos attentes, nous permettent d'émettre quelques recommandations pour de futures applications de modèles d'extraction terminologique entraînés sur des langues typologiquement éloignées de celles sur lesquelles on planifie de les appliquer. En premier lieu, nous conseillons de choisir avec soin les meilleurs modèles dans des langues morphologiquement riches, comme le turc. Il n'est pas nécessaire de disposer de gros corpus ou de larges ressources terminologiques de référence, du moment que cette dernière est complète et cohérente. En second lieu, concernant les prétraitements textuels,

2. Orthographe erronée de « visite » présente dans nos corpus
3. Orthographe erronée de « bannières » présente dans nos corpus
4. Le polonais est une langue SVO, mais dont l'ordre des composants est relativement libre.

une segmentation en unités similaires au mot semble être la meilleure option, et il n'est nul besoin de procéder à une segmentation morphologique pour autre chose qu'une éventuelle évaluation des modèles. Enfin, en ce qui concerne le choix des traits pour l'entraînement des modèles, nous ne sommes pas absolument convaincus que le fait de comparer les comptes de fréquence entre un corpus générique et un corpus de spécialité apporte une réelle plus-value aux modèles. Même si une minorité des modèles les plus performants (en anglais par exemple) font usage de tels traits, ils ne sont pas suffisamment répandus dans l'ensemble des meilleurs modèles pour être jugés indispensables, au contraire du trait mesurant la fréquence simple en corpus de spécialité par exemple. Cela semble être tout l'avantage de cet algorithme reposant sur les CRF : découvrir le *termhood* d'une unité sans avoir à comparer son usage à d'autres données. Cet ensemble d'expériences a surtout permis de confirmer qu'il était possible et désirable de développer des modèles d'extraction terminologiques cross-lingue de qualité raisonnable. Par ailleurs, la facilité d'entraînement des modèles, et le fait de pouvoir disposer d'un certain nombre d'excellents modèles dans différentes langues pourrait permettre d'améliorer l'extraction plus avant en concaténant les meilleurs modèles existants par exemple. Cela pourrait faire l'objet de perspectives d'amélioration ultérieures.

Troisième partie

Complétion de terminologie multilingue structurée

Introduction de la troisième partie

La partie précédente a fait l'objet d'une réflexion et d'expériences sur une possible manière d'extraire des termes de façon indépendante de la langue. La terminologie multilingue résultante est, bien que de qualité potentiellement satisfaisante en terme d'extraction, de couverture inégalement complète selon les langues. Suivant les intuitions de Dagan *et al.* (1991), confirmées entre autres par Dyvik (1998), les différents usages d'un mot et ses sens dans différentes langues constituent une information sémantique remarquable. Pour cette raison, le fait de disposer de terminologies comparables dans plusieurs langues constitue un atout à exploiter. Il s'agira ici, de combler dans des langues l'absence de terme relevés par ailleurs pour d'autres langues, ou bien même de créer à partir de rien une terminologie comparable dans une nouvelle langue à l'aide d'un graphe de traduction multilingue (chapitre 9).

La terminologie multilingue obtenue avec l'outil d'extraction présenté dans la partie II de ce travail n'est pas structurée. Or, la structuration est un problème complexe, d'autant plus que l'utilité d'une structure de terminologie pour les ressources humaines peut plus dépendre du cadre applicatif que d'un réel lien ontologique entre des concepts. Par exemple, selon les enquêtes concernées, on peut vouloir lier les termes « poubelle » et « agent d'entretien » sous la classe « propreté des lieux » plus que les termes « poubelle » et « benne à ordures » sous une classe relevant des contenants à déchets.

Cette remarque nous permet de revenir brièvement sur la théorie terminologique. Dans la section 1.1 du premier chapitre, nous avons énuméré les différentes critiques émises par des terminologues sur le cadre théorique traditionnel, et notamment sur ce que doit représenter un concept, et comment. Nous avions été particulièrement séduits par le point de vue

de Temmerman (2000), qui prône le remplacement des concepts par des unités de compréhension (UC). Ces dernières représentent des prototypes cognitifs pouvant être structurés assez librement pour représenter les connaissances. Cette approche sémasiologique permet, et justifie même, la liberté que l'on peut prendre pour organiser les termes dans une structure conceptuelle. Ainsi il n'est pas aberrant de regrouper des termes comme « agent d'entretien » et « poubelle » sous la même unité de compréhension, comme cela peut arriver dans notre terminologie. Par la suite, nous utiliserons indifféremment les termes de *concept* et d'*unité de compréhension* pour désigner ce que Temmerman (2000) nomme *unité de compréhension*.
Pour cette partie, nous allons donc considérer que la terminologie multilingue obtenue a été structurée autour d'une *hiérarchie de concepts* cohérente d'un point de vue applicatif. Cette dernière, développée manuellement par des experts du domaine, est indispensable au travail de complétion qui va suivre. Sa structure (nombre de niveaux) restera néanmoins un paramètre mineur, du moment que des termes aux thématiques suffisamment proches dans différentes langues sont regroupés ensembles. C'est ce que nous verrons à la section 10.1.3.

L'approche générale envisagée utilisera un graphe de traduction multilingue (chapitre 9) pour traduire (dans la mesure du possible) et enrichir une terminologie multilingue structurée (chapitre 10). Cette partie détaillera les éléments du Bloc B de la figure 0.1 (p. 5). Elle n'aura pas une lecture proprement linéaire, car les travaux qu'elle présente ont vu leur développement et leur évaluation fractionnés en plusieurs étapes chronologiquement organisées :

— une phase initiale, qui a consisté à construire un premier graphe de traduction (YaMTG 0.1). Sur ce dernier, nous avons essayé un premier algorithme de complétion d'ontologie légère (présenté à la section 10.2.1) ;
— une phase de développement continu, qui a permis d'améliorer la couverture du graphe de traduction (YaMTG 1.0) en évaluant les techniques d'extraction et d'amélioration de la qualité de ce graphe. Durant celle ci, nous avons testé un second algorithme de complétion (voir la section 10.2.2).
— une phase terminale ne concernant que la couverture du graphe de traduction, pour laquelle nous avons proposé une version *a priori* définitive du graphe de traduction (YaMTG 2.2) librement utilisable par tous.

Plutôt que de rester fidèle à cet enchaînement chronologique, nous avons choisi d'organiser cette partie en deux chapitres : un premier concernant le développement des versions notables du graphe de traduction (chapitre 9), et un second relatif aux algorithmes que nous avons expérimenté pour la complétion de terminologie structurée multilingue (chapitre 10).

Chapitre 9

Construction d'un Graphe de Traduction Fortement Multilingue

Sommaire

Les bases de données de traduction multilingues généralistes à grande échelle sont utiles dans de nombreuses tâches de traitement automatique des langues. Ceci est particulièrement vrai en ce qui concerne les efforts de recherche ciblant des langues pour lesquelles il existe peu de ressources linguistiques. Les bases de données de traduction peuvent être utilisées pour adapter des ressources existantes dans d'autres langues. Cela a par exemple été le cas pour le développement de wordnets dans d'autres langues que l'anglais (voir notamment de Melo & Weikum (2009a) à ce sujet). Il existe donc un réel besoin en ce qui concerne des bases lexicales multilingues libres de droit compilant autant de traductions que possible dans n'importe quelle langue.

Ce chapitre décrit YaMTG (Yet another Multilingual Translation Graph), une base de données libre (« open source ») de traduction fortement multilingue développée entre 2012 et 2014. Ce graphe, regroupant des termes en plus de 600 langues, a été construit à l'aide de ressources couvertes par des licences permissives ou sous régime de « gauche d'auteur » (*copyleft*), comme cela sera détaillé à la section 9.3.1.

Pour le développement de YaMTG, nous avons mis l'accent sur les caractéristiques suivantes :

1. Cette ressource est destinée à un usage générique.
2. Nous avons essayé de trouver un juste équilibre entre qualité et quantité, en essayant de retirer un maximum d'erreurs présentes dans les données.
3. Nous n'avons inféré aucune traduction de façon endogène, ceci afin d'éviter la propagation d'erreurs.
4. Cette ressource doit rester facilement accessible, utilisable et extensible.

9.1 Travaux Apparentés

Les recherches axées sur la création de bases de données de traduction fortement multilingues utilisent souvent des ressources collaboratives libres issues de Wikimedia [1] (en particulier Wikipedia et Wiktionnaire). Ce faisant, la communauté tire parti de la sagesse des foules, contribuant ainsi à surmonter le manque de lexiques préexistants dans certaines langues (Meyer & Gurevych, 2012).

Etzioni *et al.* (2007) ont construit TransGraph, un dictionnaire multilingue regroupant ses entrées par sens, à partir de données issues de wiktionnaires et d'autres dictionnaires. Il couvre 100 langues, dont trois possèdent plus de 100 000 entrées lexicales. Ce graphe de traduction a été plus tard remodelé par Mausam *et al.* (2009) lors de la création du PanDictionary. Ce

1. http://www.wikimedia.org/

dernier couvre plus de 1000 langues, dont les traductions émanent toujours de wiktionnaires et d'autres dictionnaires. Cependant, cette ressource ne semble pas être librement accessible à l'heure où nous écrivons. Le projet PanLex (Kamholz *et al.*, 2014) est une ressource de traduction lemmatique qui combine un grand nombre de ressources lexicales monolingue, bilingues et multilingues. Il couvre environ 1,1 milliard de traductions reliant 20 millions de termes dans près de 9 000 variétés de langues. Il inclut également des liens de traduction inférés à posteriori de façon endogène. Cette ressource est disponible sur le site PanLex [2]. À notre connaissance, il s'agit de la plus grande base de données fortement multilingue contenant des traductions existant à ce jour. Cependant, nous n'avons pas été en mesure de trouver une évaluation de la qualité des données.

Certaines données multilingues peuvent être extraites de données initialement conçues pour autre chose que la traduction. Le succès du WordNet de Princeton (Fellbaum, 1998) en tant que ressource linguistique générique, a stimulé la création de ressources similaires dans d'autres langues. De nombreux chercheurs de pays non anglophones ont lancé différents projets, impliquant parfois plusieurs équipes, dont le but était de construire de nouveaux wordnets dans leurs langues respectives (Hamp & Feldweg, 1997 ; Farreres *et al.*, 1998 ; Pianta *et al.*, 2002 ; Diab, 2004 ; Sagot & Fišer, 2008). Une liste des ressources compatibles avec la structure des wordnets peut être trouvée sur le site Web de la *Global WordNet Association* [3]. Il existe également plusieurs initiatives s'appliquant à établir des relations entre les wordnets de différentes langues. Le projet EuroWordNet (Vossen, 1998) a été dévolu à la construction et au rattachement manuel de nouveaux wordnets dans huit langues européennes. Cette fusion d'informations lexico-sémantiques contenues dans les wordnets locaux a été proposée par Atserias *et al.* (2004). Des initiatives similaires destinées à la création de wordnets multilingues ont également fait leur apparition dans d'autres zones géographiques. C'est le cas pour BalkaNet (6 langues, Stamou *et al.* (2002)) ou IndoNet (18 langues, Bhat *et al.* (2013)), qui ont été manuellement développés par des experts, avec quelques tâches automatisées. Cependant, la qualité de l'information lexico-sémantique contenue dans ces ressources prévaut sur l'aspect multilingue.

À mi-chemin entre les bases de données lexicales multilingues et les ontologies légères apparentées à wordnet, certaines recherches ont opéré un regroupement de toutes ces données multilingues à l'intérieur d'un unique graphe de connaissances. De Melo et Weikum (2009) ont développé un WordNet Universel, couvrant des concepts indépendamment de la langue. À cette fin, ils ont utilisé des wordnets dans de nombreuses langues, des dictionnaires de traduction librement accessibles, différentes éditions du Wiktionnaire, des thesaurus et ontologies mono- et multi-lingues, et des données issues de corpus parallèles (de Melo, 2012, p.

2. http://www.panlex.org/
3. http://globalwordnet.org/?page_id=38

35-37). La version téléchargée en février 2014 (nommée « 201012 ») contient plus de 85 000 sens, instanciés par presque 1,5 millions d'items lexicaux. Au total, 419 langues sont représentées. Parmi elles, 46 possèdent plus de 10 000 instances. Ce WordNet Universel est librement téléchargeable sous licence Creative Commons.

BabelNet, développé par Navigli & Ponzetto (2010, 2012), est une base de connaissances multilingue à large couverture. Sa construction a nécessité le couplage de connaissances encyclopédiques et de données onto-lexicales (en l'occurrence wikipedia and wordnet). Les concepts ainsi obtenus ont été complétés avec des lexicalisations issues d'une chaîne de traitement de traduction automatique dans un grand nombre de langues. La version la plus récente de BabelNet (v. 2.0), contient 44 490 880 items lexicaux dans 50 langues, le tout regroupé sous un inventaire de 9 348 287 sens différents [4]. Une autre ressource de traduction exploitant les données de wikipédia pour 8 langues européennes a également été proposée par Sérasset (2012) pour des utilisations liées au web sémantique.

Parmi les ressources susmentionnées, les plus intéressantes en termes de couverture linguistique (Baldwin *et al.*, 2010 ; de Melo & Weikum, 2009a ; Navigli & Ponzetto, 2012) utilisent une architecture spécifique qui relie toutes les traductions à des nœuds dénotant du *sens* plutôt que les traductions entre elles. Cette structure de données a tendance à amplifier le bruit dans les données de traduction dans les cas où un lien erroné relierait un terme dans une langue donnée à un nœud *sens*. Comparativement à ces ressources, YaMTG (v. 2.2) contient moins d'items lexicaux (6, 5M, à comparer aux 12M de PanLex et aux 44M de BabelNet 2.0 ; voir le tableau 9.1 pour le détail des chiffres). Cela est dû au nombre réduit de ressources utilisées pour sa création, ainsi qu'au faible coût de développement mis en œuvre. Néanmoins, seul PanLex couvre plus de langues que YaMTG. La version 1.0 de YaMTG a fait l'objet d'une publication (Hanoka & Sagot, 2014).

Table 9.1 – Comparaison des chiffres de base pour les principales ressources de traduction multilingue et YaMTG.

Ressources	# Entrées Lexicales	# Langues
YaMTG 1.0	881 643	664
UWN	1 481 412	419
YaMTG 2 .2	3 466 336	675
PanLex	12 000 000	1353
BabelNet 2.0	44 490 880	50

4. Peu de temps avant la fin de la rédaction de cette thèse, nous notons la sortie d'une version 3.0 (beta) de BabelNet (disponible à l'adresse http://babelnet.org/stats), contenant 263 langues et 110 605 360 sens.

Seules les premières et la dernière versions de YaMTG sont décrites dans ce chapitre.

9.2 Versions initiales : YaMTG 0.1 et 1.0

La version initiale de YaMTG (Hanoka & Sagot, 2012) a été développée en extrayant automatiquement les traductions présentes dans des éditions de wiktionnaires en plusieurs langues, ainsi que dans la Wikipédia française, afin de déterminer dans quelle mesure un graphe de traduction multilingue peut proposer de nouveaux candidats termes dans des langues pour lesquelles il n'en existe pas ou peu. Cette première mouture du graphe nous aura permis d'élaborer un processus d'extraction de traductions candidates (section 9.3.1) et une méthodologie de filtrage à même d'améliorer la qualité des traductions (section 9.2.2, détaillée en appendice D). Ce processus aura été raffiné et évalué lors du développement de la version 1.0 de YaMTG. Cette section présente ces deux versions.

9.2.1 Sources des liens de traduction

9.2.1.1 Wiktionnaires

Wiktionnaire est un dictionnaire multilingue conçu pour être le compagnon lexical de Wikipédia. Comme décrit par Meyer & Gurevych (2012), il existe des Wiktionnaires indépendants pour chaque langue. Ces derniers sont appelés éditions. Dans chaque édition, une page wiki correspond à du matériel lexical. Ce dernier peut être soit dans la langue de l'édition donnée, soit dans d'autres langues, soit les deux. Les différentes éditions sont régulièrement mises à disposition sous format téléchargeable par la Fondation Wikimedia à l'adresse http://dumps.wikimedia.org/.

Les Wiktionnaires sont organisés comme des ensembles de pages structurées, dont le titre peut correspondre à un article décrivant une ou plusieurs entrée(s) lexicale(s). Le contenu de la page dans des éditions de langues différentes peut exhiber des structures dissemblables pour encoder l'information linguistique.

Lors du développement de la version 0.1 de YaMTG, nous avons extrait des paires de traduction et de synonymes d'un ensemble de Wiktionnaires en dix-huit langues (issus de dumps mis à disposition fin 2011) : allemand, anglais, coréen, espagnol, français, hébreu, indonésien, italien, japonais, néerlandais, polonais, portugais, roumain, russe, slovaque, suédois, tchèque et turc. Au total, 4 522 947 liens de traduction et de synonymie ont pu être extraits de ces 18 wiktionnaires pour la version 0.1 de YaMTG.

Cet échantillon de langues a évolué pour la récupération des liens de traductions/synonymie dans les versions ultérieures de YaMTG, passant de 18 à 21 langues, en abandonnant certaines au profit d'autres. Pour la version YaMTG 1.0 (décrite dans Hanoka & Sagot (2014)), les 21

éditions de langues concernées sont : allemand, anglais, bulgare, danois, espagnol, français, grec, hindi, hongrois, italien, japonais, néerlandais, polonais, portugais, roumain, russe, slovaque, suédois, tchèque, turc et vietnamien (téléchargées en septembre 2013). Cela représente au total 3 116 695 liens de traduction/synonymie.
Le processus d'extraction automatique de traductions (et chaque fois que cela était possible, de synonymes) que nous avons utilisé est décrit et évalué dans l'annexe D pour 21 éditions de langues différentes.

9.2.1.2 Traductions d'OPUS (YaMTG 1.0 et plus)

OPUS (Tiedemann, 2009) est une compilation de corpus parallèles issus de plusieurs ressources en ligne incluant des textes politiques et administratifs de l'Union Européenne, des sous-titres de film, des données biomédicales de l'Agence Européenne des Médicaments, de corpus de la Banque Centrale Européenne et d'autres. Ce projet vise à fournir à la communauté des corpus parallèles facilement exploitables dans un grand nombre de langues. Comme candidats termes à notre graphe de traduction, nous avons considéré l'ensemble les alignements de mots datant du 9 octobre 2013, ayant une fréquence supérieure ou égale à 10. Ainsi, nous avons retenu, pour les versions 1.0 et plus, 3 631 229 traductions pour 570 159 termes en 31 langues. Étant donné que les traductions issues d'OPUS sont susceptibles de contenir des erreurs d'alignement, leur qualité peut ne pas être aussi bonne que voulu. Néanmoins, ils contiennent une quantité substantielle de formes fléchies, que l'on ne retrouvera pas dans les autres ressources. Nous nous appuierons sur l'étape de filtrage (section 9.2.2) afin supprimer un maximum d'erreur.

9.2.2 Nettoyage du graphe

Suite à la récupération de ces données, nous avons agrégé l'ensemble de ces liens de traduction/synonymie au sein d'un graphe. YaMTG 0.1 comprenait 10 664 730 arcs dirigés. Chaque nœud de YaMTG représente un terme dans une langue donnée. Les arcs, reliant des termes entre eux ont conservé l'information relative à l'origine de la traduction. Le sens de la traduction *dans cette version 0.1 uniquement* a été conservé, car nous voulions initialement prévenir une propagation d'erreurs relatives à des problèmes d'extraction de traduction, relativement fréquentes. Ce choix n'a pas été maintenu dans les versions suivantes de YaMTG (1.0 et supérieures), pour mettre à profit l'ensemble des traductions, quels que soient leur sens original. Nous avons en revanche, comme nous le verrons par la suite, maintenu notre décision de ne pas inférer des liens de traductions comme ont pu le faire Mausam *et al.* (2009). La version 1.0 de YaMTG comprend quand à elle environ cinq millions de traductions.

9.2.2.1 Principes

Le graphe non filtré contient sans surprises, un certain nombre d'erreurs. Nous en avons identifié quatre classes possibles :

1. les traductions dont la langue source ou cible n'a pas été correctement identifiée ; par exemple 今日(jpn) → 今日日(ita)
2. des définitions prises à tort pour des traductions ; par exemple *pälsdjursfarm* (swe) → *farm där man föder upp djur för pälsens skull* (swe)
3. des traductions contenant du bruit (principalement des caractères indésirables) ; par exemple 木棉花(zho) → *baumwollen ??? wörtlich : Kapokblüte* (deu)
4. des erreurs variées, non répertoriées ci-dessus. par exemple *caro* (por) → *coar* (por)

La plupart des erreurs de type 1 à 3 vont pouvoir être filtrés par les procédés mis en œuvre en section 9.3.3. Les erreurs de type 4 (« erreur variées ») sont les plus pénalisantes étant donné que la plupart d'entre elles ont peu de moyens d'être éliminés automatiquement. Les erreurs de type 4 sont principalement le fait d'un processus d'extraction parfois trop permissif des traductions issues des wiktionnaires.

L'ensemble des traductions candidates récupérées précédemment doivent donc être filtrées afin d'améliorer la qualité globale de la ressource. Ce processus repose sur la suppression de traductions candidates répondant à au moins un des critères suivants :

— Un des termes de la traduction est soit *trop long* (en général une définition), soit *trop court* [5] (en général des *stopwords*) ;
— Ils contiennent de la ponctuation et/ou des caractères numériques.
— Leur jeu de caractères est incohérent avec la langue déclarée.
— Un des termes de la traduction apparaît dans moins de X autres traductions.
— Un des termes de la traduction apparaît dans plus de Y autres traductions. Lorsque trop de traductions sont proposées pour un même terme, il apparaît souvent que ce dernier est une erreur.

Les trois premiers critères sont conçus pour éviter de garder les erreurs relevant des classes 1 à 3. Les deux autres critères ont quant à eux vocation à limiter également le nombre d'erreurs de type 4.

Concernant les choix du paramère X, il s'est avéré, à travers toutes les expériences que nous avons pu faire sur l'ensemble des versions de YaMTG, que supprimer les *hapax* et les *dis lego-*

5. Ce filtre ne concerne que les mots faisant moins de trois lettres écrits dans les alphabets cyrilliques, grecs et latins.

menon (noeuds apparaissant strictement moins de trois fois dans l'ensemble des traductions) permettait un bon équilibre entre la qualité des traductions et la couverture du graphe. Ce seuil minimum $X = 3$ permet de ne pas trop pénaliser les langues rares tout en gardant un seuil peu compatible avec le hasard (une erreur peut difficilement apparaître 3 fois dans les données par pur hasard)

Pour ce qui est du choix du paramètre Y, nous avons supposé qu'il pouvait être lié au nombre de langues présentes dans le graphe une fois les quatre premiers filtres appliqués. Imaginons qu'un graphe partiellement filtré comporte n langues pour lesquelles il existe 100 termes ou moins. Nous supposons arbitrairement que parmi ces termes, il ne peut raisonnablement pas y avoir en moyenne plus de deux traductions pour chacune de ces n langues. Dans ce cas, un terme ne peut pas être impliqué dans plus de $Y = 2n$ traductions. En pratique, ce paramètre de filtre peut supprimer une infime portion de termes corrects mais qui sont suffisamment polysémiques et bien traduits dans l'ensemble des langues pour déroger à cette règle. C'est par exemple le cas du terme anglais *dog*. Nous conservons toutefois ce critère de filtrage car il permet malgré tout de supprimer de nombreuses traductions erronées ou trop polysémiques.

9.2.2.2 Évaluation sur YaMTG 1.0

Les performances de cette méthode de filtrage ont été évaluées sur une version postérieure du graphe (version 1.0), incluant en plus des sources déjà présentes dans la version 0.1 de YaMTG des traductions issues d'alignement de corpus parallèles OPUS (voir section 9.2.1.2). Une différence majeure entre les versions 0.1 et 1.0 de YaMTG est que cette dernière est un graphe non orienté. Cette version du graphe [6] est décrite dans Hanoka & Sagot (2014). A cette occasion, une évaluation du filtre pour la version en fonction des types d'erreurs a été accomplie. Nous redonnons dans cette section les principaux résultats obtenus pour l'évaluation du filtrage.

À l'échelle du graphe de traduction initial, environ trois quart des termes et des langues présentes dans le graphe non filtré ont été exclus du graphe final par cette heuristique de filtrage. Malgré cela, 34% des liens de traduction/synonymie ont été conservés (cf. Table 9.2). Parmi ces derniers 51,2% sont extraits uniquement d'OPUS et 42,1% de l'une des 21 éditions de wiktionnaires utilisées pour l'occasion. La proportion de traductions issues à la fois d'OPUS et des wiktionnaires atteint 3,2%.

Nous avons évalué la qualité du graphe filtré en évaluant 1 639 liens de traduction/synonymie sélectionnés comme suit :

6. Il s'agit de la version 1.0, disponible à l'adresse http://alpage.inria.fr/~hanoka/yamtg.html.

Table 9.2 – Nombre de termes, traductions/synonymes et langues avant et après filtrage lors de la construction de la version 1.0 de YaMTG.

	Avant filtrage	Après filtrage (YaMTG 1.0)
Termes	6 121 187	881 643
Trad./Syn.	17 176 698	5 842 279
Langues	4 324	664

— 1 439 liens issus d'une précédente évaluation concernant la qualité de l'extraction de traduction des différents Wiktionnaires (cf. Annexe D, section D.0.0.1) ;
— 200 liens supplémentaires choisis aléatoirement dans le graphe de traduction filtré pour ne pas biaiser la précision globale.

Ces liens de traduction/synonymie couvrent au total 3 332 termes dans 72 langues.

◇ Précision globale :

Nous avons estimé la précision globale des liens de traductions/synonymie dans le graphe filtré en procédant à l'évaluation manuelle de 200 traductions aléatoirement sélectionnées. Les chiffres ainsi obtenus sont présentés dans le tableau 9.3 : la précision globale y est estimée à 91%.

Table 9.3 – Estimation de la précision globale de la qualité du graphe de traduction filtré (YaMTG v.1.0) reposant sur une évaluation manuelle d'un échantillon aléatoire de 200 liens de traduction/synonymie.

Source	total	Erronés				Corrects	
		Misc.	lang.	def.	bruit		
OPUS	99	10	0	0	0	89	90%
Wiktionaires	85	2	4	0	2	77	91% [7]
OPUS+Wikt.	16	0	0	0	0	16	100%
Total	200	12	4	0	2	182	91%

Parmi les 18 traductions incorrectes (9%) présentes dans cet échantillon aléatoire, quatre ont vu un de leurs termes se faire assigner une mauvaise langue, deux contiennent toujours du

7. Ce chiffre est plus petit que ce que l'on peut attendre étant donné les résultat obtenus dans le tableau D.3 de la page 280. Il s'agit là d'un artefact relatif au petit nombre de liens sélectionnés aléatoirement pour l'évaluation, qui a pour conséquence un nombre encore plus faible d'erreurs parmi les traductions extraites de wiktionnaires uniquement (8 erreurs). Comme nous le verrons plus loin, la précision des traductions (filtrées) extraites des Wiktionnaires est plus proche de 97 %.

bruit et 13 sont des erreurs variées (principalement des erreurs d'alignements issues d'OPUS). Il est à noter que toutes les définitions prises à tort pour des traductions dans le processus d'extraction du wiktionnaire ont été supprimées avec succès par l'heuristique de filtrage.

◇ Évaluation de l'heuristique de filtrage :

Sur les 3 150 liens de traduction évalués à l'occasion de l'ensemble des évaluations (filtre et qualité globale de la ressource), 1 439 (45,3%) ont pu être conservés dans la version filtrée du même graphe. Le tableau 9.4 (reproduisant partiellement le tableau D.3 (p. 280)), compare les chiffres avant et après l'application du filtre pour une évaluation des liens de traduction/synonymie extraits de plusieurs éditions de wiktionnaires.

Table 9.4 – Résultats globaux de l'évaluation de traductions/synonymes non-filtrés et filtrés issues de 21 éditions de wiktionnaires. Pour les chiffres détaillés, voir tab. D.3 (p. 280).

Avant filrage						
	Erronnés			Corrects		
Misc.	lang.	déf.	bruit	total	%	
84	62	175	69	2760	—	
2.7%	1.9%	5.6%	2.2%	87.6%	—	
Après filtrage						
	Erronnés			Corrects		Conservés
Misc.	lang.	déf.	bruit	total	%	%
19	11	17	3	1389	—	—
1.3%	0.8%	1.2%	0.2%	96.5%	—	—

Pour les éditions de langues concernées, ces chiffres indiquent que les liens de traduction/synonymie après application du filtre ont une précision moyenne de 95,8%. Dès lors que l'on pondère les scores par les proportions du nombre de traductions disponibles pour chaque édition de langues différentes présentes dans le graphe filtré (Table 9.5), on obtient une estimation de la précision finale de 97,2% pour les traductions extraites des Wiktionnaires. En moyenne, ce filtre est parvenu à améliorer les scores de précision (non pondérés et pondérés) de près de 8 points.

9.2.3 Remarques

Ces évaluations quantitatives menées sur la version 1.0 de YaMTG ont permis de constater que les procédés utilisés pour son développement permettent d'obtenir une ressource de

TABLE 9.5 – Nombre de liens de traductions/synonymes pour l'ensemble des sources de la version de YAMTG (v. 1.0) sur laquelle a été menée l'évaluation de la méthode de filtre (sous-tableau a). Il est à noter qu'un même lien de traduction peut être extrait de plus d'une source à la fois, et peut donc être comptabilisé plus d'une fois (voir sous-tableau b).

Source	édition de langue	#arcs de trad./syn.
Wiktionnairesy	Slovaque	826
	Hindi	4,012
	Danois	18,281
	Italien	38,031
	Japonais	38,332
	Suédois	42,842
	Tchèque	61,798
	Bulgare	68,809
	Turc	72,037
	Grec	80,328
	Vietnamien	97,622
	Polonais	137,058
	Allemand	139,136
	Roumain	147,717
	Espagnol	151,981
	Hongrois	152,564
	Néerlandais	160,144
	Portuguais	166,669
	Russe	309,209
	Français	362,404
	Anglais	866,895
	Total	3,116,695
OPUS		3,178,247

(a)

#Sources	#arcs
1	5,449,753
2	333,179
3	58,192
4	1,142
5	13

(b)

qualité : pour cette version, les liens de traduction/synonymie ont été estimés corrects dans approximativement 91 % des cas. Ce score est toutefois susceptible d'être amélioré relativement simplement, en récupérant plus de traductions candidates : une particularité relative à notre heuristique de filtrage concerne l'écrémage d'items lexicaux (et des liens de traduction/

synonymie dans lesquels ils apparaissent) sur la simple base qu'ils constituent des *hapax* ou *dis legomena* dans l'ensemble des traductions recueillies. Pour cette raison, il est à prévoir que plus le nombre de traductions candidates est élevé et leurs sources variées, plus il sera possible de conserver des traductions/synonymes corrects en augmentant les chances qu'un terme dans une langue apparaisse dans plus de deux traductions. C'est ce que nous avons essayé de mettre en œuvre en continuant développement de YaMTG. La section suivante présente la ressource qui émane de cet effort d'amélioration. Il est à noter que nous n'avons pas procédé au même ensemble d'évaluations sur la ressource finale, comme nous allons le voir.

9.3 Version finale : YaMTG 2.2

La version 2.2 propose une augmentation du nombre de traductions de 780% et du nombre de termes de 293% par rapport à la version 1.0. Le nombre de langues représentées n'augmente en revanche que de 1,6%. Cette section détaille les moyens mis en œuvre pour parvenir à une telle inflation : la récupération de nouvelles traductions libres (section 9.3.1), les étapes de création (section 9.3.2) ainsi que certaines propriétés intéressantes du graphe final (section 9.3.4).

9.3.1 Sources

Nous avons voulu améliorer la couverture du graphe en utilisant d'autres ressources, mais en continuant à nous limiter à des ressources libres. Nous allons donc décrire les ressources nouvellement utilisées pour la construction de YaMTG 2.

9.3.1.1 Dictionnaires Bilingues

Il existe un nombre considérable de dictionnaires de traduction open-source disponibles sur Internet. Une liste de dictionnaires bilingue de cet acabit est proposée par de Melo (2012, p. 36). Parmi ces ressources, nous avons utilisé Apertium, CEDICT, HanDeDict et Free Dict (références et détails ci-après). Nous avons également extrait des traductions d'OmegaWiki, DictionaryForMIDs , des dictionnaires extraits par V. Solomko [8] et d'un dictionnaire bilingue indépendant anglais-bengali.

◇ Lexiques Apertium :

Apertium est une platforme de traduction automatique open-source (licence GNU GPL) développée par le groupe de recherche *Transducens* de l'université d'Alicante. Ce projet est

8. Téléchargeables à l'adresse http://www.slovnyk.org/.

disponible en ligne à l'adresse http://www.apertium.org/. Les données linguistiques développées et mises à disposition incluent un grand nombre de paires de langues, souvent apparentées (par exemple espagnol-catalan, danois-norvégien, etc). Le nombre de paires de langues proposées augmente régulièrement. Nous avons utilisé les paires de langues disponibles en février 2014. Parmi elles, nous avons retenu 843 616 termes pour 860 175 traductions dans 29 langues.

◇ CEDICT, HanDeDict, CFDICT, ChEDICC :

CEDICT (disponible à l'adresse http://www.mdbg.net/chindict/chindict.php?page=cedict) est un dictionnaire chinois-anglais mis à disposition sous licence CC BY-SA 3.0. Ce projet a inspiré quelques dictionnaires similaires dans d'autres langues : Pour l'espagnol, le ChEDICC peut être téléchargé à l'adresse http://cc-chedicc.wikispaces.com/. HanDeDict, la version allemand-chinois, est quant à elle disponible sur le site http://www.handedict.de/chinesisch_deutsch.php?mode=dl. La version français-chinois est téléchargeable sur http://www.chine-informations.com/chinois/open/CFDICT/. Tous ces dictionnaires de traduction sont sous licence CC BY-SA 3.0. Pour ces cinq langues, nous avons gardé 558 388 traductions impliquant 579 096 termes.

◇ Freedict :

Le projet Freedict (http://www.freedict.org/) recense un ensemble varié de ressources bilingues disponibles sous licence GNU GPL ou sous des licences moins restrictives. Des 81 dictionnaires proposés, nous avons conservé un ensemble de 981 150 traductions reliant 1 049 504 termes en 18 langues.

◇ DictionaryForMIDs :

DictionaryForMIDs est un projet open source (licence GPL) conçu pour donner librement accès à des dictionnaires polyvalents pour toutes sortes d'appareils électroniques (ordinateurs, téléphones portables, PDA etc.). Les dictionnaires proposés au téléchargement sont disponibles à l'adresse http://dictionarymid.sourceforge.net/. Nous avons extrait les traductions à partir des dictionnaires suivants : le dictionnaire anglais-tchèque (créé par Peter Kmet), le dictionnaire allemand-tagalog (de Piero Peruzzi), le dictionnaire anglais-hindi *Shabdanjali* et le dictionnaire anglais-khmer *SSBIC*. Ces derniers ont été distribués soit sous licence

GPL, soit sans mention particulière concernant la licence. De ces quatre dictionnaires, nous avons extrait et gardé 193 942 traductions reliant 180 046 termes dans six langues.

◇ OmegaWiki :

Omegawiki (http://www.omegawiki.org) est un projet collaboratif visant à produire un dictionnaire multilingue libre. Le dictionnaire Omegawiki est conjointement sous licence GFDL et CC-BY. Il est accessible depuis le site de DictionaryForMIDs. Nous en avons retenu 386 984 termes pour 410 872 traductions en 34 langues.

◇ Dictionnaires de Valentyn Solomko :

Un nombre considérable de dictionnaires bilingues sont mis en ligne par V. Solomko à l'adresse http://www.slovnyk.org/. Tous sont sous licence GNU GPL. De ces derniers, nous avons conservé 56 404 927 traductions connectant 6 476 909 termes en 31 langues.

◇ Dictionnaire Bilingue anglais-bengali :

Le dictionnaire anglais-bengali Ankur est offert au téléchargement sous licence GNU GPL à l'adresse http://www.bengalinux.org/english-to-bengali-dictionary/. Il recense 36 541 traductions reliant 22 869 termes.

9.3.1.2 Wordnets Libres

Le Princeton WordNet (PWN, Fellbaum (1998)) est une base de données lexicale structurée, pour la langue anglaise, développée manuellement depuis plus d'une décennie. Il consiste en un ensemble de termes, rassemblés en groupe de sens appelés *synsets*. Le PWN relie sémantiquement les *synsets* et/ou les mots entre eux. Pour ce travail, nous avons considéré uniquement les termes et les *synsets* auxquels ils appartiennent. Peu après sa publication, plusieurs initiatives pour créer des ressources similaires dans d'autres langues, en utilisant le PWN comme référence, ont été entreprises (Vossen, 1998 ; Pianta *et al.*, 2002 ; Stamou *et al.*, 2002 ; Diab, 2004 ; Sagot & Fišer, 2008). Plusieurs travaux se sont même employés à l'unification des données ainsi produites dans des structures de données multilingues de Melo & Weikum (2009a) ; Navigli & Ponzetto (2010).
Nous avons mis à profit les wordnets librement disponibles dans différentes langues, et uti-

lisant des identifiants de synsets cohérent (*synset IDs* PWN 3.0) [9], en considérant ces synsets IDs comme étant des termes d'une langue artificielle. Intégrer des traductions reliant des synsets IDs à des termes du graphe de traduction améliore la fiabilité en cas d'ambiguïté.

Nous avons utilisé six wordnets :

— Princeton WordNet (Fellbaum, 1998) (anglais) —license WordNet 3.0.
— Wordnet Bahasa (Noor *et al.*, 2011) (indonésien) —License MIT.
— OpenWordnet-Pt (de Paiva *et al.*, 2012) (portugais brésilien) —CC-BY-SA 3.0.
— Hebrew WordNet (Ordan *et al.*, 2007) (hébreu) —© Hebrew WordNet 2007.
— Japanese WordNet (Isahara *et al.*, 2008) (japonais) —© NICT, 2009-2010 .
— Wordnet Libre du Français (WOLF) (Sagot & Fišer, 2008) —License CeCILL-C.

Le rassemblement de ces wordnets a permis de regrouper 223 475 termes dans 6 langues naturelles reliés aux 122 838 synset IDs formant une « interlingua ». Au total, 385 029 liens de « traductions » entre 346 313 termes et éléments de l'interlingua sont retenus comme candidats pour intégrer le graphe de traduction final.

9.3.1.3 Wikipedia Française

Wikipédia est une encyclopédie multilingue au contenu libre rédigée collaborativement par des contributeurs bénévoles. Les articles (pages) Wikipédia contiennent des informations encyclopédiques sous forme de paragraphes de texte, d'images, de listes ou de tableaux. Les articles sont reliés entre eux, et certains articles peuvent également disposer de liens vers le même article dans une langue différente (*liens interlangues*). Ces liens interlangues fournissent des bases de traductions faciles à extraire pour les titres d'articles. Nous avons utilisé la version XML extraite en 2011 par Sagot & Stern (2012) dans le cadre de la construction de la base d'entités nommées Aleda. Les textes, images, contenus multimédias intégrés, bien que libres, peuvent être licenciés différemment à la demande des éditeurs de la page en question. En revanche, il n'est pas clair de savoir à quelle licence sont soumis les liens interlingues. Nous considérerons donc que ces derniers sont soumis à une licence conjointe CC BY-SA et GNU Free Documentation License (GFDL) [10]. Cette extraction a permis de récupérer 6 141 783 candidats termes (parmi lesquels presque 11% sont des termes français) impliquées dans 7 496 959 traductions, allant d'un terme français à un autre terme dans une des 267 autres langues représentée.

9. Dans l'éventualité où la version native des synset IDs d'un wordnet n'était pas en PWN 3.0, leurs identifiants ont été automatiquement reliés aux synsets IDs 3.0 correspondants au moyen d'un script d'alignement développé par Tomaž Erjavec.
10. Plus de détails à l'adresse http://en.wikipedia.org/wiki/Wikipedia:Copyrights

9.3.1.4 Wiktionnaires

Pour la description des wiktionnaires et du processus d'extraction de traductions, nous renvoyons à la section 9.2.1.1. Il a été effectué sur les éditions de wikitionnaires (téléchargées en février 2014) dans les même 21 langues que pour la version 1.0 de YAMTG (allemand, anglais, bulgare, danois, espagnol, français, grec, hindi, hongrois, italien, japonais, néerlandais, polonais, portugais, roumain, russe, slovaque, suédois, tchèque, turc et vietnamien). Certains bogues d'extraction de traduction ont été progressivement réglés, ce qui a permis d'extraire 8 151 765 traductions candidates reliant 6 074 411 termes dans 3 756 langues.

9.3.2 Création du Graphe de Traduction

Les étapes précédentes ont permis de regrouper aux alentours de 78.10^6 traductions candidates dans près de 3 760 langues [11].

Toutes ces traductions candidates sont modélisées par un graphe de traduction $G_{cand} = (S_{cand}, A_{cand})$ où :

- S_{cand} est un ensemble fini de sommets candidats. Un sommet représente un terme dans une langue donnée ;
- A_{cand} est un sous-ensemble de $S_{cand} \times S_{cand}$, représentant l'ensemble des liens (candidats) de traduction ou de synonymie [12] entre les sommets du graphe. Chaque arc de traduction candidat est étiqueté par les informations sur son ou ses origine(s).

Il s'agit d'un *graphe simple*, c'est-à-dire non-dirigé, sans boucles et au sein duquel il existe au maximum une arête entre deux sommets. Au total, il contient près de 18.10^6 sommets et 78.10^6 arêtes. Au total, presque un million des traductions identiques issues de sources différentes qui ont été récupérées en première intention ont pu être fusionnées. Le graphe multilingue regroupant ces traductions est encodé en UTF-8. Le tableau 9.6 récapitule le nombre de termes, traductions et langues par origine, et indique la taille du graphe unissant toutes ces ressources.

11. À titre de comparaison, l'association *Ethnologue* (http://www.ethnologue.com/) recense plus de 7 105 langues encore utilisées aujourd'hui.

12. Nous modéliserons donc ici la synonymie comme un cas particulier de traduction reliant des termes de même langue.

TABLE 9.6 – Nombre de traductions candidates (avant filtrage) par sources réunies dans YAMTG 2.2.

Source	# Termes	# Traductions	# Langues
Apertium	843 616	860 175	29
*DICT(zho)	579 096	558 388	5
Freedict	1 049 504	981 150	18
MIDs	180 046	193 942	6
Omega	386 984	410 872	34
Solomko	6 476 909	56 404 927	31
Ankur	22 869	36 541	2
OPUS	874 468	3 631 229	31
Wordnets	346 313	385 029	7
Wikipedia fr	6 141 783	7 496 959	262
Wiktionnaires	6 074 411	8 151 765	3 756
TOTAL UNION	18 144 861	77 957 379	3 756

La plupart des traductions proposées sont issues de ressources lemmatiques. Certaines sources, comme OPUS ou les wiktionnaires, favorisent l'introduction de formes fléchies ou d'expressions idiomatiques dans le graphe.

9.3.3 Filtrage

Concernant les paramètres de filtres (section 9.2.2), nous avons conservé le seuil minimal $X = 3$. Le seuil maximal Y a quant à lui été fixé à 1000. Après avoir appliqué les premiers critères de filtrage (c'est-à-dire ceux non liés à la fréquence d'un terme au sein des traductions candidates), le graphe contient des items lexicaux dans plus de 3 700 langues. Pour ces langues, seulement 500 apparaissent dans au moins 100 traductions. Pour rappel, ce chiffre a été déterminé parce que nous avons estimé qu'un terme ne devrait raisonnablement pas avoir en moyenne plus de deux traductions pour ces 500 langues. Il est à noter que ce filtre peut éliminer des termes pertinents. Le nœud du graphe représentant l'entrée lexicale « dog (eng) », ayant initialement un degré de 1 241, est par exemple évincé du graphe final. Cela est dû au fait que ce terme possède un certain nombre de synonymes, et est traduit dans de nombreuses langues, y compris de façon erronée.

Le tableau 9.7 présente l'effet du filtre sur le graphe de traduction en fonction des différentes sources.

TABLE 9.7 – Nombre de traductions (après filtrage) par sources réunies dans YAMTG 2.2.

Source	# Termes	# Traductions	# Langues
Apertium	224 347	396 854	29
*DICT(zho)	113 286	169 904	5
Freedict	350 285	487 533	18
MID	115 491	118 096	6
Omega	142 955	171 153	34
Solomko	2 700 731	33 545 790	31
Ankur	1 171	757	2
OPUS	371 258	3 158 613	31
Wordnets	84 085	123 246	7
Wikipedia fr	1 453 276	1 255 834	247
Wiktionnaires	1 048 830	3 568 002	675
TOTAL UNION	3 466 336	39 743 127	675

Les traductions les plus filtrées proviennent du dictionnaire bilingue anglais-bengali Ankur : seulement 5% des termes et 2% des traductions originales issues de cette ressource sont conservées. Les sources dont la majeur partie des traductions sont conservées sont OPUS (87%), MID (61%) et Solomko (59%). Alors que la plupart des ressources conservent un nombre de langues fixe à toutes les étapes, les traductions de la Wikipédia française et celles issues des wiktionnaires voient leur nombre diminuer respectivement de 6% et 82%.

Une fois les traductions candidates triés, il s'avère que le graphe ainsi obtenu n'a globalement conservé que 19% des termes et 18% des langues initialement présentes dans le graphe non filtré. Cependant, 51% des liens de traduction initialement présents ont été maintenus (cf. Table 9.8).

TABLE 9.8 – Nombre de termes et traductions avant et après filtrage durant la construction du graphe de traduction.

	Avant Filtrage	Après Filtrage (YaMTG 2.2)	% Conservés
Termes	18 144 861	3 466 336	19
Trad./Syn.	77 957 379	39 743 127	51
Langues	3 756	675	18

La plupart des traductions représentées dans le graphe ne possèdent qu'une source (93,5%). 3,8% ont deux sources, 0,6% ont trois sources, près de 0,2% des traductions ont quatre sources et les traductions ayant cinq sources ou plus représentent 0,07% de l'ensemble des traductions.

9.3.4 Propriétés du Graphe

Les « graphes de terrain » (issus de données empiriques) sont souvent des graphes dits « *petit monde* » (Gaume, 2004). Il est raisonnable de penser que notre graphe de traduction n'échappe pas à la règle. Le phénomène du petit monde fait référence aux travaux menés par Stanley Milgram (Milgram, 1967), qui suggèrent que deux personnes prises au hasard n'importe où dans le monde ne sont en moyenne qu'à six degrés de séparation l'une de l'autre. Autrement dit, il suffit au premier individu de passer par l'intermédiaire d'un petit nombre de personnes se connaissant pour atteindre le second individu.

Il n'existe pas de critères absolus permettant de classifier de façon certaine un graphe dans la catégorie des petits mondes. Néanmoins, deux propriétés indiquent qu'un graphe est petit monde (Watts & Strogatz, 1998) : (1) il contient un grand nombre de clusters (c'est-à-dire ensemble de nœuds très connectés entre eux, et relativement isolés du reste du graphe) et (2) la distance moyenne entre deux nœuds est proportionnelle au logarithme du nombre de nœuds (cela est du à l'existence d'un petit nombres d'arcs « à longue portée » reliant des nœuds ne faisant pas partie d'un même cluster). On peut également rajouter les critères suivants, découlant peu ou prou des précédents (Gaume, 2004) : (3) il est peu dense : le nombre d'arcs est de l'ordre de $|S| \log |S|$ et (4) sa distribution des degrés d'incidence suit une loi de puissance.

Humphries & Gurney (2008) ont proposé une mesure à même de juger du degré de « petitesse du monde » d'un graphe. Cette dernière consiste à comparer les coefficients de clustering et la longueur de chemins du graphe petit monde G à ceux d'un graphe d'Erdös-Rényi G_{rand} aléatoire. Soit L_G (resp. $L_{G_{rand}}$) la moyenne des plus courts chemins dans G (resp. G_{rand}) ; soit n_G^{Δ} le nombre de triangles dans G and $n_G^{(2)}$ le nombre de chemins de longueur 2 dans G (la même définition valant pour G_{rand}). On définit le coefficient de clustering de G comme étant $C_G^{\Delta} = 3\frac{n_G^{\Delta}}{n_G^{(2)}}$ (la même définition valant pour G_{rand}). G a les caractéristiques d'un graphe petit monde si $C_G^{\Delta} \gg C_{G_{rand}}^{\Delta}$ (propriété (1)) et $L_G \geq L_{G_{rand}}$ (propriété (2)). Calculer la moyenne des plus courts chemins pour un graphe contenant $|S|$ sommets et $|A|$ arêtes est, avec les algorithmes habituels, de l'ordre de $\mathcal{O}(|S| \log |S| + |A|)$ (Cao *et al.*, 2011). Cette opération n'est donc pas appropriée à la taille de notre graphe, qui contient plusieurs millions de nœuds.

Pour déterminer si ce graphe de traductions est effectivement de type petit monde, nous avons calculé les différentes métriques concernées sur la plus grande composante connexe du graphe (PGCC par la suite). Cette dernière comporte 2 152 347 nœuds et 34 395 481 arcs, soit respectivement 62% des nœuds et 86,5% des traductions contenus dans le graphe global. Ces termes proviennent de 669 langues différentes.

Le coefficient de clustering de la PGCC, vérifiant la propriété (1), est bien conforme à ce que l'on pourrait attendre d'un graphe petit monde : $C_G^{\Delta} = 0.284 \gg C_{G_{rand}}^{\Delta} = 1,4.10^{-5}$.

La propriété (2), relative à une distance moyenne entre deux nœuds, n'a pas pu être vérifiée en calculant la valeur moyenne du chemin le plus court comme nous l'avions annoncé plus haut. En revanche, il nous a été possible de confirmer que la PGCC du graphe est très peu dense (propriété (3)) : $D_{PGCC} = 1,48.10^{-5}$. Le nombre d'arcs de la PGCC est conforme à ce que l'on peut généralement observer pour les graphes de terrain selon le critère (3).
La figure 9.1 présente la répartition des degrés dans le PGCC. Le degré moyen du graphe est de $31,9$ (avec un écart type de ± 49). On constate figure 9.1 que la distribution des degrés d'incidence du PGCC correspond bien à une loi de puissance (propriété (4)).

FIGURE 9.1 – Répartition des degrés dans la plus grande composante connexe de YAMTG 2.2. L'axe des ordonnées représente le log-nombre de nœuds.

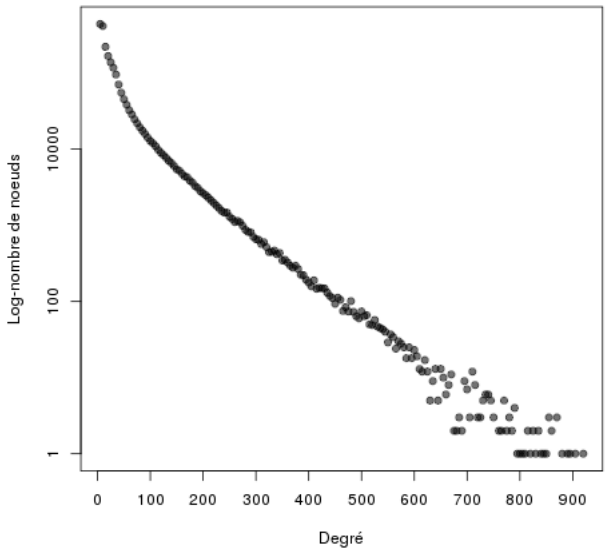
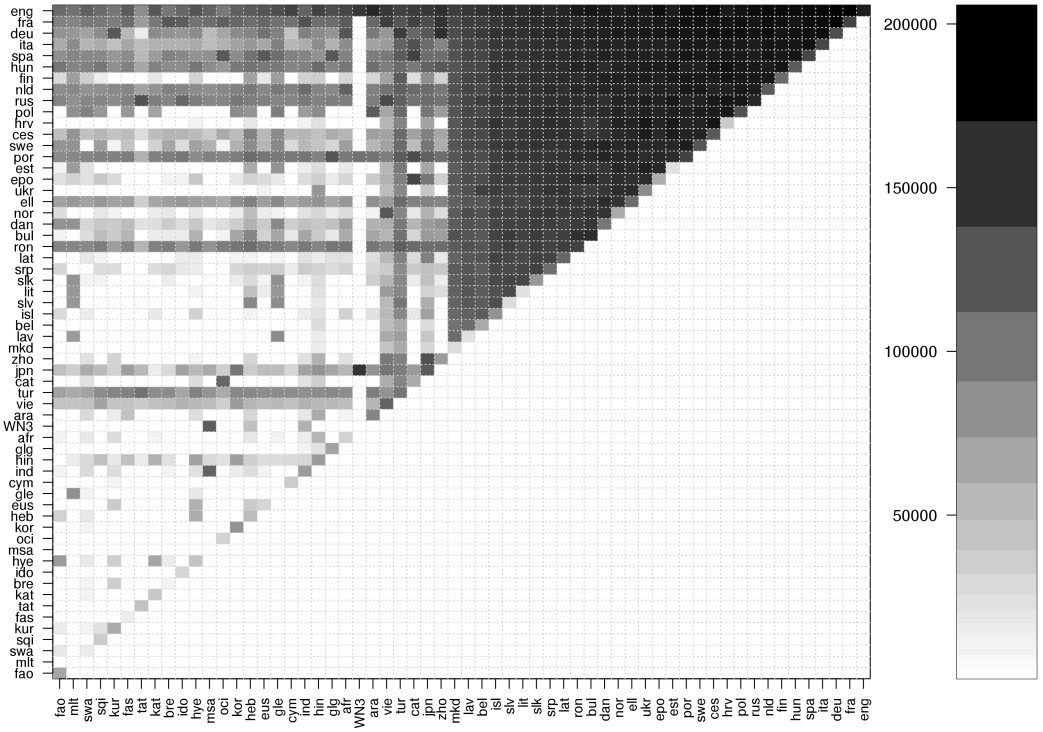


FIGURE 9.2 – Représentation (avec palette de niveaux de gris logarithmique) de la proportion de traductions reliant les termes dans les 60 langues les plus fréquentes de YAMTG 2.2. Plus une cellule est sombre, plus il y a de traductions/synonymes dans cette langue. Les langues sont indexées avec leur code terminologique ISO 639-2.

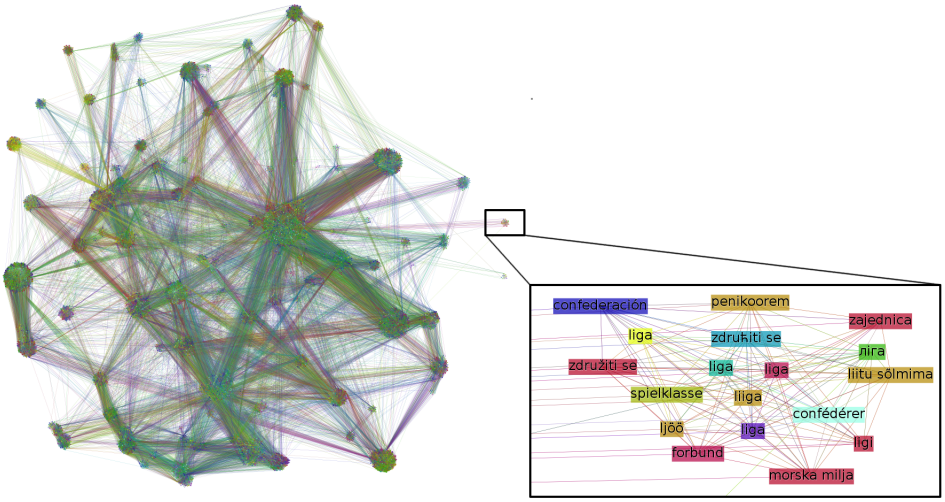


La figure 9.2 illustre la distribution des langues impliquées dans les traductions. Comme le graphe n'est pas dirigé (c'est-à-dire les traductions n'ont pas de sens), la demie matrice suffit à représenter les relations entre les langues dans YAMTG 2.2. Les liens de synonymie se trouvent dans la diagonale de cette matrice, si l'on considère qu'il s'agit d'un cas particulier de traduction d'un terme dans une même langue. On peut y voir que la majorité des traductions impliquent des termes dans des langues européennes. Ces dernières, ainsi que certaines autres langues pour lesquelles un effort d'extraction supplémentaire a été consenti (japonais, turc, vietnamien), constituent également un bon panel de « langues pivot » pour des langues plus rares (lignes plus sombres de la matrice).

On peut par ailleurs anticiper le fait que les traductions sont socioculturellement polarisées dans les ressources dont elles sont issues. Par exemple, une ressource de traduction multilingue issue de la sphère vietnamienne offrira sans doute plus de traductions (et de meilleure qualité)

dans des langues qui sont géographiquement proches (khmer, thaï, malaisien, laotien, chinois, indonésien,...) ou historiquement liées (français, anglais). Ainsi, l'ajout de langues issues de pôles éloignés permettra en théorie de combler les écarts entre des langues considérées comme peu documentées sans certaines sphères linguistiques, et moins dans d'autres.

FIGURE 9.3 – Portion du graphe YAMTG 2.2, spatialisé avec l'algorithme OpenOrd (Martin *et al.*, 2011) disponible dans Gephi v.0.8.2 (Bastian *et al.*, 2009), avec agrandissement illustratif d'un cluster de termes. Chaque couleur représente une langue.



La figure 9.3 donne une idée de l'allure générale du graphe. Il est possible d'y voir les différents clusters, plus où moins reliés entre eux, indiquant l'existence de liens polysémiques.

9.4 Synthèse

YAMTG a fait l'objet de plusieurs phases d'améliorations qualitatives et quantitatives. Ces dernières ont abouti à la mise à disposition libre [13] d'un graphe de traduction non orienté comprenant 39 743 127 traductions impliquant 3 466 336 termes en 675 langues. Au terme des différentes étapes visant à améliorer la qualité du graphe par le truchement de filtres heuristiques, la qualité de ce dernier a été estimée à plus de 95% sur la version 1.0.

En règle générale, l'utilisation d'une donnée soumise à une certaine licence oblige sa redistribution sous la même licence. Légalement, YAMTG 2.2 se qualifie d'« œuvre composite » : il s'agit d'une œuvre nouvelle à laquelle est incorporée une œuvre préexistante sans la collabo-

13. Les versions 1.0 et 2.2 de YAMTG sont disponibles librement à l'adresse http://alpage.inria.fr/~hanoka/yamtg.html.

ration de l'auteur de cette dernière [14]. Les traductions issues de sources hétérogènes qui sont agglomérées dans YaMTG 2.2 sont régies par différentes licences en fonction de leur(s) source(s) originales. Si l'on souhaitait donner un encadrement légal convenable à la ressource globale, il faudrait produire une nouvelle forme de licence qui permette de superposer les droits et obligations de chacune des licences impliquées. Il s'agirait d'un travail juridique complexe, qui dépasserait le cadre de cette thèse. Pour cette raison, nous avons fait le choix de distribuer le graphe YaMTG sous un principe de « gauche d'auteur » (*copyleft*) sans licence globale. En revanche, étant donné que chaque traduction garde trace de son origine dans le graphe, elle sera de façon individuelle selon sa source, soumise aux obligations de sa licence initiale (ou la plus permissive des licence dans le cas où ladite traduction aurait plusieurs origines).

14. Article L113-4 de la loi n° 92-597 du 1 juillet 1992

Chapitre 10

Complétion de terminologie multilingue

Sommaire

Les termes extraits dans la partie II ne l'ont été que dans un nombre restreint de langues et, selon les corpus utilisés pour leur extraction, les terminologies monolingues qu'ils constituent ont une couverture inégale de l'ensemble des thèmes pouvant être abordés dans un domaine donné. Pour cette raison, nous avons examiné dans quelle mesure il est possible de combler des lacunes terminologiques pour une langue et à travers les langues.

Comme cela a déjà été évoqué dans l'introduction de cette partie, l'atout majeur dont nous disposons ici relève de la composante fortement multilingue de notre terminologie comparable. Cet atout, pour être exploitable, doit satisfaire à une condition : que les termes appartenant à la même unité de compréhension soient mis en rapport entre eux, quelles que soient les langues considérées. La section 10.1 abordera cette question aussi bien du point de vue des techniques couvertes par la littérature pour traiter ces problématiques (section 10.1.1) que de la forme à donner à la structure de données représentant une terminologie multilingue dans la perspective des traitements qui suivront (section 10.1.3).

La section 10.2 présentera deux essais de complétion de terminologie multilingue structurée qui ont été expérimentées l'une après l'autre sur différentes versions de YaMTG.

10.1 Contexte applicatif et motivations

Selon les termes de Gruber (1993), « une ontologie est la spécification explicite d'une conceptualisation ». Uschold & Gruninger (2004) complètent cette définition en y ajoutant deux critères supplémentaires : elle est constituée (1) d'un vocabulaire de termes qui font référence aux sujets d'intérêt d'un domaine particulier et (2) de spécifications sémantiques sur les termes, reposant idéalement sur une forme de logique.

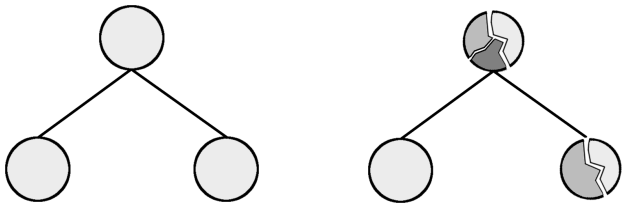
Il est communément admis qu'il existe un continuum au niveau de la quantité d'information donnée pour spécifier les termes et les relations qu'ils entretiennent (Uschold & Gruninger, 2004) : moins les spécifications sont détaillées, plus l'ontologie est dite légère. Elle ne consiste alors qu'en des termes, reliés entre eux par des liens mono-typés (traditionnellement des liens hyperonymiques de type « A est-un B »). Comme le rappellent Guarino *et al.* (2009), ce qui caractérise *a minima* une ontologie relève d'un lien hiérarchique (entre ce qui est plus générique et ce qui est plus spécifique) suffisamment général pour être compris par tous. À l'autre extrémité du spectre, les ontologies formelles [1] cristallisent plus d'informations sémantiques et logiques associées à des contraintes rigoureuses.

Dans notre cas, le type de connaissances que nous traitons a un caractère taxonomique, mais les thèmes abordés peuvent relever d'une grande subjectivité : les liens hiérarchiques ne sont

1. Au sens strict, seules ces dernières peuvent se qualifier d'ontologies.

pas, comme traditionnellement, de type « A est-un B » mais plutôt « A est-relatif-à B ». Les termes regroupés dans un même ensemble (nœud ou synset) ne sont pas nécessairement (quasi-)synonymes. Ces nuances ont une forte répercussion sur la nature de la structure de donnée résultante. C'est ce que schématise la figure 10.1.

FIGURE 10.1 – Représentation schématique de structures d'unités de compréhension monolingues dans le cas canonique (a) et pour notre terminologie structurée (b). Dans le cas (a), nœuds et unités de compréhension coïncident. Dans le cas (b) , un nœud peut regrouper, conformément à l'intuition de l'expert du domaine, plusieurs concepts distincts au sein d'une thématique large mais cohérente.



(a) Cas canonique (ontologie légère de type wordnet)

(b) Cas de notre terminologie structurée

Dans quelle mesure la complétion de terminologie structurée semblable à celles schématisée dans la figure 10.1 (b) peut-elle s'envisager comme un cas particulier d'une problématique plus globale relative à la construction et à l'enrichissement de structures ontologiques légères (comme celle schématisée dans la figure 10.1 (a)) ? Répondre à cette question justifierait que nous puissions tester nos propositions sur des wordnets plutôt que sur notre ressource originale. C'est cette problématique que nous allons aborder dans la section 10.1.1.

10.1.1 Revue des approches pour la complétion de taxonomies

Pour initier cette réflexion, nous allons aborder brièvement la question des stratégies possibles pour la création et l'extension d'ontologies légères (de type canonique comme dans la figure 10.1 (a)).

Le champ de recherche en détection automatique de liens taxonomiques est vaste, et s'adresse à l'extraction de relations sémantiques ou lexicales variées : « A est-un B » , « A est-une-partie-de B », « A est-similaire-à B », « A est-synonyme-de B », « A cause B », etc. (Yang & Callan, 2009). L'approche la plus ancienne (monolingue) utilisée pour identifier ces différentes relations emploie des règles (pré-définies ou heuristiques) reposant sur des patrons lexico-syntaxiques supposés identifier certains types de relations à partir d'un corpus (cf. entre autres Navigli *et al.* (2011)). Par exemple, le patron « NP_1 est un NP_2 » peut identifier

un lien d'hyponymie entre les syntagmes nominaux NP_1 et NP_2 comme dans la phrase « le [rouge-gorge] est un [oiseau] ». Bien que ce genre d'approche parvienne en règle générale à extraire des types de relations précis et à rajouter des termes à la bonne place dans la hiérarchie, elle est inefficace dès lors que des relations de ce type n'apparaissent pas explicitement dans le corpus. En ce qui nous concerne, elle serait inapplicable sur nos données, qui ont un registre particulier : par exemple, si plusieurs employés courroucés ont déclaré « le directeur est un abruti », cela pourrait créer ou rajouter des liens ontologiquement problématiques. Une seconde approche pour la découverte de relations sémantiques utilise une approche distributionnelle qui envisage cette tâche comme un problème de classification ou de *clustering* et permet ainsi de découvrir des relations non-explicites dans les corpus (Navigli *et al.*, 2011). Il est à craindre que, à l'instar des approches par patrons, cette dernière associe des termes qui co-occurent beaucoup dans les corpus sans indiquer de liens ontologiques pertinents. Qui plus est, il est fréquent que les verbatim soient trop courts pour pouvoir tirer parti du contexte. Certaines publications proposent d'étendre des structures ontologiques en utilisant le web comme un corpus (Liu *et al.*, 2005 ; Kozareva & Hovy, 2010). Cette approche paraît plus adaptée au cas de figure qui nous intéresse ici. Toutefois, le coût de développement pour chaque langue de cette méthode envisagée dans un cadre fortement multilingue rend cette option malcommode à appliquer pour cette thèse.

Or justement, la question du multilinguisme pour créer ou étendre de wordnets a éclos peu de temps après la publication du WordNet de Princeton (WNP par la suite), comme en témoigne l'initiative EuroWordNet (Vossen, 1998). Ce domaine de recherche s'est considérablement développé durant les années suivantes (Pianta *et al.*, 2002 ; Stamou *et al.*, 2002 ; Diab, 2004 ; Sagot & Fišer, 2008 ; de Melo & Weikum, 2009a ; Navigli & Ponzetto, 2010).
Les approches multilingues utilisant le WNP comme référence pour la création de nouveaux wordnets dans d'autres langues se fondent sur l'utilisation d'un inventaire de synsets similaires. Ce procédé n'est pas dénué de difficultés car il existe des divergences conceptuelles entre les langues et il n'est pas toujours aisé de trouver une traduction correcte pour un terme donné en fonction de sa place dans la taxonomie. Néanmoins, le fait que chaque synset répertorie un sens de façon non-ambiguë, comme schématisé dans la figure 10.1 (a), constitue un avantage : la caractérisation sémantique est nette et bien définie.
À l'instar des raisonnement présentés dans les travaux de Resnik & Yarowsky (1997) et Ide *et al.* (2002) sur la désambiguïsation sémantique multilingue, Dyvik (1998) a proposé de trouver des traductions pertinentes pour des termes issus de synsets en exploitant des traductions issues de corpus parallèles bilingues. Pianta *et al.* (2002) ont quant à eux suggéré d'utiliser des (contre-)traductions [2] bilingues associées à une information contextuelle sous forme de

2. « Contre-traduction » traduit l'anglais *back-translation*.

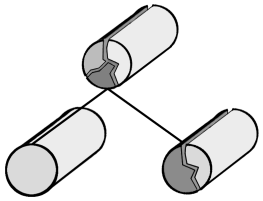
gloses afin de faciliter le travail lexicographique mis en œuvre pour la construction du pan italien de MultiWordNet. Une autre approche proposée par Sagot & Fišer (2008) consiste à fusionner les résultats de deux techniques différentes pour le remplissage des synsets : d'une part, la récupération de traductions extraites de corpus parallèles en cinq langues, et d'autre part des traductions extraites de lexiques bilingues issus de ressources libres. Davidov & Rappoport (2009) ont quant à eux mis au point une approche fondée sur la (contre-)traduction de termes spécifiant un même concept (en l'occurrence, un synset) dans un grand nombre de langues intermédiaires (45 au total) pour compléter ledit concept.

Au delà de ces techniques, certains auteurs ont pris le parti d'utiliser des ressources de type wiki pour la création et l'extension de wordnets dans plus d'une nouvelle langue à la fois : Navigli & Ponzetto (2010) ont établi un lien entre les informations multilingues de type encyclopédique contenus dans Wikipedia et le WNP, qui est utilisé comme squelette pour organiser le reste des connaissances ainsi regroupées. Dans un état d'esprit approchant, de Melo & Weikum (2009b) ont entraîné une machine à vecteur support (Vapnik, 1995) sur un graphe de connaissances multilingue (dérivé de resources mono- et bi-lingues) et un ensemble de wordnets pré-existants pour construire un wordnet universel (*universal WordNet*, UWN).

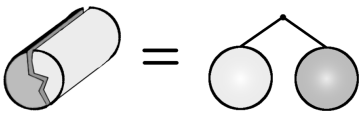
10.1.2 Compatibilité avec nos données

La section précédente a présenté des solutions de construction et d'extension pour les bases de données terminologiques, organisées au sein d'une partition structurée de concepts bien délimités (figure 10.1 (a)) dans les cadres mono- et multi-lingues. Or, ces dernières sont moins bien applicables pour les terminologies structurées semblables à celles schématisée dans la figure 10.1 (b), pour lesquelles des termes afférents à des concepts distincts coexistent et forment des unités de compréhension hétérogènes quoique cohérentes. Fort heureusement, notre terminologie structurée a une dimension sémantique supplémentaire : la composante multilingue. Cet aspect a été représenté dans la figure 10.2 (a), dans laquelle chaque fragment de nœud, dénotant une unité de compréhension prend corps grâce à la traduction de ses termes en plusieurs langues.

FIGURE 10.2 – Représentation schématique de notre terminologie structurée, avec l'ajout de la dimension multilingue (figure (a)). La figure (b) réduit la dimension multilingue en réajustant l'organisation sémantique implicite d'un nœud grâce à l'apport de sens offert par la composante multilingue.



(a) Cas de notre terminologie structurée multilingue



(b) Aplanissement de la dimension multilingue

Ainsi, un fragment d'unité de compréhension voit sa description sémantique aussi bien définie par sa position dans la hiérarchie que par le nombre et la pertinence des traductions des termes qui l'instancient. Ce glissement est schématisé par la figure 10.2 (b). Nous avons évoqué le fait que, dans les graphes de traduction YAMTG, les liens de synonymie étaient modélisés comme des liens de traduction ; ici, le corollaire prévaut : les traductions sont traités comme des (quasi-)synonymes, créant ainsi des « synsets implicites ». D'un point de vue théorique, cette opération justifie que l'on puisse envisager d'appliquer les heuristiques de complétion présentées à la section 10.2 aussi bien à des ontologies légères comme les wordnets qu'à des terminologies structurées multilingues ayant une structure conceptuelle non canonique. La section suivante évoquera des solutions pratiques en terme de structures de données pour unifier la représentation des taxonomies à compléter.

10.1.3 Structure hiérarchique multilingue

Pour tous les types d'ontologies, il existe des formalismes normalisés comme comme SKOS (*Simple Knowledge Organisation System*), RDF (*Resource Description Framework*) ou OWL (*Web Ontology Language*), pour ne citer que les plus connus [3], conçus pour faciliter la portabilité des ontologies existantes pour des applications apparentées. Ces formats sont notamment utilisés dans le cadre du web sémantique. Or, le problème de la représentation idéale d'une terminologie multilingue dépend, dans le cas présent, de la structure de données permettant de simplifier au maximum les traitements ultérieurs, à savoir la comparaison de ces terminologies multilingues avec un large graphe de traduction. À ce titre, nous n'avons pas jugé indispensable d'utiliser les formalismes du W3C. Afin de garder les choses simples, il nous a

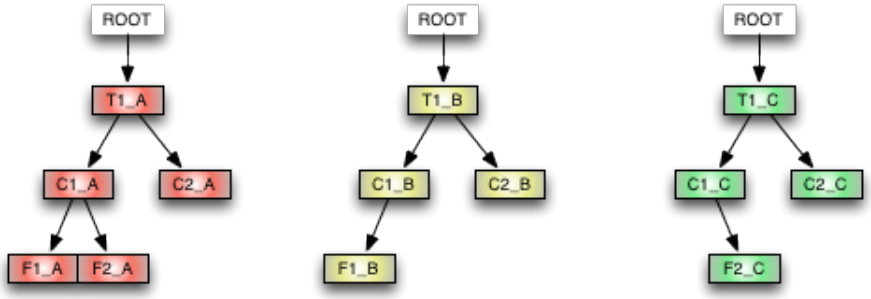
3. Pour plus de détails, consulter les rubriques correspondantes sur le site du W3C (*World Wide Web Consortium*) : http://www.w3.org/.

semblé approprié d'utiliser une structure de donnée directement comparable à notre graphe de traduction.

10.1.3.1 Cas général : choix de la structure de données

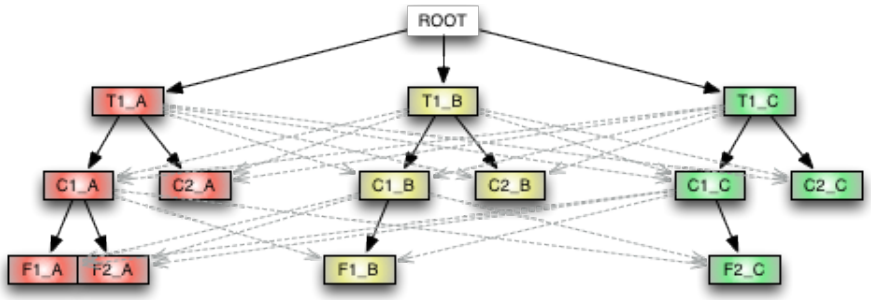
Le choix de la structure de données dépend de la spécificité que l'on désire conférer au traitement : dans l'idéal, on ne souhaite pas donner d'influence au nombre de niveaux dans la hiérarchie et ne traiter que les termes pour pouvoir conserver un aspect générique. Mais la question la plus impérative concerne la représentation du multilinguisme. La figure 10.3 schématise des terminologies structurées comparables en trois langues (A, B et C).

FIGURE 10.3 – Schéma de trois terminologies structurées comparables dans les langues fictives A, B et C. Les nœuds ROOT représentent les racines des différentes arborescences, les différents codes préfixant chaque nœud correspondent à des identifiants désignant différents niveaux d'une hiérarchie fictive de profondeur 3 : T pour Thème, C pour Classe, F pour Feuille (les feuilles sont, en l'occurrence, des termes).



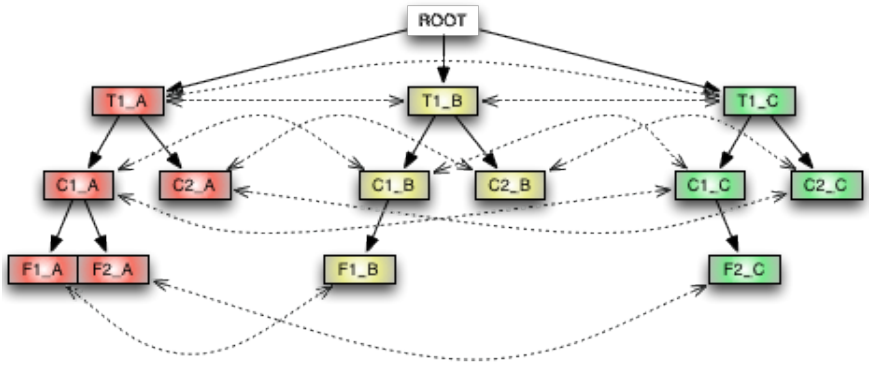
Parmi les structures de données candidates pour représenter le regroupement de ces terminologies multilingues comparables, il y a les cas particuliers de graphes à même de permettre l'identification de niveaux de spécialisation tels que les arbres, les graphes dirigés acycliques (aussi appelés DAG – pour *Directed Acyclic Graphs*), les treillis, les hypergraphes, etc.
Les représentations qui nous ont semblé les plus pertinentes sont les graphes dirigés acycliques (DAG), les graphes hiérarchiques et les arbres. La représentation à l'aide de DAG permet de connecter directement dans la hiérarchie chaque nœud à ses enfants dans la même langue ainsi qu'aux traductions de ses enfants. Un nœud dans une langue donnée est connecté, au niveau supérieur de la hiérarchie, au nœud parent ayant la même langue que lui ainsi qu'aux traductions de son parent direct. Néanmoins, les traductions à un même niveau ne sont pas indiquées. Cette représentation est schématisée dans la figure 10.4.

FIGURE 10.4 – Représentation schématique d'une terminologie comparable multilingue en trois langues (A, B et C) avec un graphe dirigé acyclique (DAG).



Une alternative possible à cette représentation consiste à relier entre elles (à l'aide de sous-graphes) les traductions de même niveau dans la hiérarchie. C'est ce que représente la figure 10.5, qui peut être vue comme un graphe hiérarchique. Cette structure de donnée peut être transformée en hypergraphe si les nœuds de même niveau dans la hiérarchie qui sont reliés entre eux sont regroupés en un hyper-nœud.

FIGURE 10.5 – Représentation schématique d'une terminologie comparable multilingue en trois langues (A, B et C) avec un graphe hiérarchique.

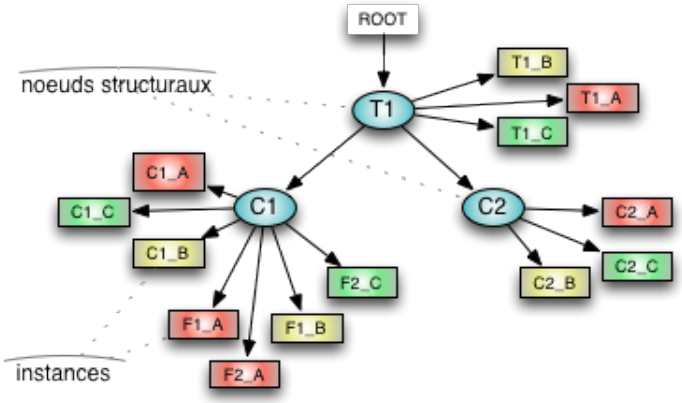


Bien que mieux adaptée à la tache de complétion que l'agencement en DAG, la représentation utilisant un graphe hiérarchique manque de compacité. Une solution hybride entre cette dernière et l'hypergraphe correspondant consiste à utiliser un arbre [4] typé distinguant les nœuds structuraux (correspondant presque à des hyper-nœuds) des nœuds instances (titres des classes ou termes). Cette solution, que nous avons préférée à toutes les autres, est schéma-

4. En réalité, cette structure est un arbre dans le cas canonique, pour lequel un terme n'est pas rattaché à plusieurs classes. En pratique, un terme peut être polysémique et se rattacher à plusieurs classes : la structure n'est donc plus arborescente mais se transforme en DAG.

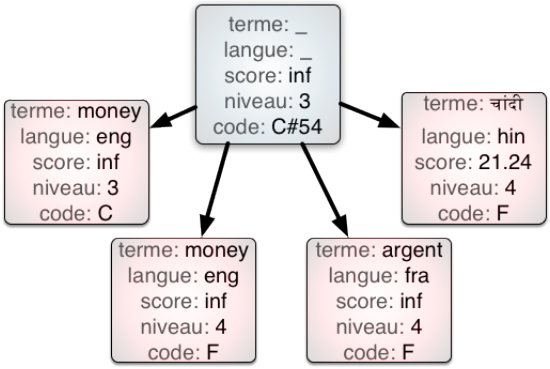
tisée dans la figure 10.6.

FIGURE 10.6 – Représentation schématique d'une terminologie comparable multilingue en trois langues (A, B et C) avec un arbre typé.



La figure 10.7 propose un aperçu du contenu possible de différents nœuds dans une portion de hiérarchie.

FIGURE 10.7 – Aperçu du contenu possible des nœuds de la terminologie multilingue : un nœud structural (nœud de niveau 3 dont les champs langue et terme sont non instanciés), une instance de nœud structural en anglais (niveau 3, instancié) et trois nœuds représentant des termes (nœuds de niveau 4).



Par la suite, on ne considérera comme indispensable que les champs terme, langue et score. Le score peut prendre *à priori* n'importe quelle valeur réelle ainsi que la valeur inf dénotant que le score du terme n'est pas sujet à questionnement car il s'agit d'un terme de référence. La langue est systématiquement désignée par son code terminologique ISO 639-2 [5].

5. La liste correspondante est disponible à l'adresse http://www.loc.gov/standards/iso639-2/php/

10.1.3.2 Cas particulier : wordnets alignés

Afin de contrôler l'efficacité des algorithmes de complétion qui vont suivre, nous avons pris le parti d'évaluer leurs propositions pour la construction d'un wordnet jouet dans une nouvelle langue, à partir de wordnets alignés en plusieurs autres langues. Comme nous l'avons vu à la section 9.3.1.2, il existe différentes éditions de wordnets reliant de façon unifiée des termes en une langue à un même inventaire de synsets. Ces derniers sont discernés par des identifiants communs à certaines éditions (celles énumérées dans la section 9.3.1.2 ou encore celles issues d'autres projets comme BalkaNet (Tufiş *et al.*, 2004)). Dans la chronologie des travaux que nous présentons ici, nous avons utilisé deux groupes de wordnet alignés différents lors des tests de nos algorithmes :

— Pour l'algorithme de contre-traduction (section 10.2.1), nous avons utilisé un ensemble de wordnets (version 2.0) synset-alignés en 4 langues (bulgare, tchèque, anglais, roumain). La version anglaise est le WordNet 2.0 de Princeton (Fellbaum, 1998), et les wordnets des autres langues sont issus du projet BalkaNet (Tufiş *et al.*, 2004). Au total, 115 424 synsets permettent de regrouper 145 626 termes en anglais, 32 510 termes en bulgare, 18 848 termes en roumain, 30 174 termes en tchèque.

— Pour l'algorithme de clustering local via recuit simulé (section 10.2.2), nous avons utilisé la même base de wordnets que pour la précédente expérience (dont les identifiants de synsets ont été actualisés conformément à la version 3.0 du WNP), ainsi que des wordnets (synsets du WNP 3.0) en cinq autres langues : hébreu, indonésien, japonais, malais, portugais brésilien. Ces derniers sont présentés à la section 9.3.1.2. Au total, nous avons ainsi utilisé pour cette expérience une base de wordnets synset-alignés en 9 langues, dont 4 langues non européennes. Ce regroupement comprend 122 838 synsets réunissant 148 729 termes en anglais, 32 372 termes en bulgare, 18 960 termes en roumain, 31 774 termes en tchèque, 4 843 termes en hébreu, 211 termes en indonésien, 58 693 termes en japonais, 11 471 termes en malais, 1 745 termes en portugais brésilien.

Dans les deux cas de figure, la hiérarchie représentant les wordnets alignés est simple : chaque nœud-terme, quelle que soit sa langue, est relié à un (ou plusieurs) nœud(s) structural(aux) ayant pour identifiant le(s) numéro(s) de synset(s) original(aux) au(x)quel il est initialement rattaché. Chaque nœud structural est quant à lui rattaché à la racine de la hiérarchie. Cette dernière ne contient que deux niveaux : les nœuds structuraux dénotant les synsets et les nœuds feuille représentant les termes.

---

code_list.php.

10.2 Algorithmes de complétion

Nous avons considéré deux méthodes s'appuyant sur le graphe de traduction pour proposer de nouveaux candidats termes dans une langue. Plutôt que de tester ces dernières sur la terminologie structurée relative au domaine des ressources humaines dont nous disposons dans le cadre industriel de cette thèse, nous avons préféré relater ici nos expérimentations sur la complétion de wordnets. Ce choix a été fait principalement pour deux raisons : (i) les wordnets sont une ressource standard dans le domaine, et (ii) il nous est possible de comparer l'efficacité de nos algorithmes à d'autres méthodes de complétion. L'évaluation théorique de ces méthodes a été menée sur une extension en conditions réelles du WOLF (wordnet libre du français), passant initialement par la création *ab initio* d'un wordnet « jouet » du français.

La première méthodologie mise en œuvre, décrite à la section 10.2.1, repose sur une approche très simple de contre-traduction pondérée. Elle a fait l'objet d'une publication (Hanoka & Sagot, 2012). La seconde approche expérimentée repose sur un algorithme de clustering local via recuit simulé. Cette dernière est détaillée à la section 10.2.2.
Les algorithmes présentés ci-après opèrent sur un synset à la fois : ils prennent *en entrée* les termes dans l'ensemble des langues disponibles pour le dit synset, et offrent *en sortie* un ensemble de termes pondérés, pouvant contenir les termes initiaux (ces derniers prennent alors un poids maximal). Plus il y a de termes dans des langues différentes, meilleures seront les propositions faites par les algorithmes. Les candidats termes ainsi proposés sont issus du graphe de traduction YaMTG et peuvent provenir de n'importe quelle langue.

10.2.1 Contre-traduction pondérée

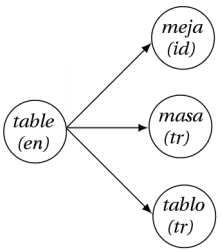
Cet algorithme présenté par Hanoka & Sagot (2012) est inspiré de l'approche miroir de Dyvik (1998), validée par Muller & Langlais (2011) pour l'appariement de synonymes et, plus largement, pour la découverte de similarité sémantique entre des termes.
En ce qui nous concerne, nous avons tiré parti du fait que nous disposions initialement d'un graphe de traductions dirigé fortement multilingue. Cet algorithme a été développé et testé que sur une version préliminaire de YaMTG (v. 0.1) ayant conservé le sens initial des traductions (voir section 9.2). Par la suite, nous désignerons ce graphe par la lettre G. L'algoritme sera détaillé à la section 10.2.1.1 et exemplifié à la section 10.2.1.2.

10.2.1.1 Principes

Nous allons préalablement à toute explication définir le *degré par langue* dans le graphe multilingue. Un noeud n de G a, au regard d'une langue l, un *degré par langue* $deg_l(n)$ égal au nombre d'arcs sortant de n incidents à d'autres noeuds représentant des termes dans la langue

l. Par exemple dans la figure 10.8, $deg_{tr}(table_{en}) = 2$ et $deg_{id}(table_{en}) = 1$.

FIGURE 10.8 – Exemple simplifié de nœuds voisins dans le graphe de traduction orienté G.



L'algorithme considère les termes initialement présents dans un synset, sans distinction de langue. Ces termes, qui constituent l'ensemble Γ_0 des nœuds feuille ayant un nœud structurel commun, sont des termes de référence (nœuds *gold*). Le score ω_{gold} qui leur est associé vaut initialement 100 (choix arbitraire). L'ensemble Γ_0 est donné en entrée à l'algorithme de complétion.
À partir de cet ensemble Γ_0, on va utiliser le graphe de traduction G pour construire un ensemble Θ_1 de termes candidats. On regroupe ensuite les ensembles Θ_1 et Γ_0 dans un nouvel ensemble Γ_1 en réajustant les scores de façon heuristique. Puis Γ_1 est à son tour donné en entrée à l'algorithme de complétion pour produire un ensemble Γ_2, de la même façon que Γ_0 a produit Γ_1. Et ainsi de suite, l'ensemble de termes candidats (et de référence) pouvant représenter la sémantique du synset considéré grossit par inflations successives.

Plus formellement, à chaque étape k de l'inflation et pour un synset donné, l'algorithme prend en variable d'entrée un ensemble multilingue Γ_k de termes représentés par des nœuds. Ces derniers contiennent toute l'information indispensable sur le terme concerné pris parmi l'ensemble de tous les termes possibles, lesquels sont indexés de 1 à N. Cette information concerne notamment la forme graphique γ_i, la langue l_i et un score $\omega_i^{(k)}$, où i est compris entre 1 et N (l'ensemble des N termes n'étant pas nécessairement tous représentés dans Γ_k).

L'ensemble Γ_k est utilisé pour rechercher des traductions dans G ; toutes les traductions ainsi trouvées sont regroupées dans un nouvel ensemble Θ_{k+1} contenant toute l'information sur ces nœuds candidats : leur forme graphique γ_j, leur langue l_j et leur score $\omega_j^{(k+1)}$. Une fois l'ensemble Θ_{k+1} complet, les termes qu'il contient sont intégrés à Γ_{k+1} sous certaines conditions.

Pour chaque γ_i initialement présent dans l'ensemble de termes Γ_k, on ajoute à Θ_{k+1} toutes les traductions de γ_i (peu importe la langue) présente dans G. Chaque terme candidat $\gamma_j \in \Theta_{k+1}$ se voit assigner un score $\omega_j^{(k+1)}$, valant initialement $\frac{1}{deg_{l_j}(\gamma_i)}$: ce score correspond à la probabilité d'aboutir sur le nœud γ_j dans la langue l_j depuis le nœud γ_i. Le poids

Algorithme 1 Contre-traduction pondérée

Pré-conditions : Γ_k; $a, b \in \mathbb{N}$; $a < b$
  pour tout $(\gamma_i, l_i, \omega_i^{(k)}) \in \Gamma_k$ faire
    $\Theta_{k+1} \leftarrow$ getTranslationsFor(γ_i, l_i)
    pour tout $(\gamma_j, l_j, \omega_j^{(k+1)}) \in \Theta_{k+1}$ faire
      $\omega_{back} \leftarrow 0$
      $B \leftarrow$ getTranslationsFor(γ_j, l_j)
      pour tout $(\beta_x, l_x) \in B$ faire
        si $\beta_x \in \Gamma_k$ alors
          $\omega_{\beta_x} \leftarrow$ getScoreIn$\Gamma_k\big((\beta_x, l_x)\big)$
          si $\omega_{\beta_x} \geq \omega_{gold}$ alors
            $\omega_{back} = \omega_{back} + a$
          sinon si $\omega_{\beta_x} \geq \omega_{gold}/5$ alors
            $\omega_{back} = \omega_{back} + b$
          sinon si $\omega_{\beta_x} < \omega_{gold}/50$ alors
            $\omega_{back} = \omega_{back} - b$
          fin si
        fin si
      fin pour
      si $(\gamma_j, l_j) \in \Gamma_k$ alors
        $\omega_j^{(k+1)} \leftarrow \omega_j^{(k+1)} +$ getScoreIn$\Gamma_k\big((\gamma_j, l_j)\big)$
      fin si
      $\omega_j^{(k+1)} \leftarrow \omega_j^{(k+1)} + \omega_{back}$
    fin pour
  fin pour
  $\Gamma_{k+1} \leftarrow \Gamma_k \cup \Theta_{k+1}$
  retourner Γ_{k+1}

$\omega_j^{(k+1)}$ est par la suite modifié en fonction de l'adéquation de ses contre-traductions. Soit B l'ensemble de contre-traductions obtenues à partir des membres de Θ_{k+1} qui apparaissent aussi dans l'ensemble Γ_k (termes de référence et termes ajoutés lors de l'inflation $k-1$ de Γ). Pour chaque $\beta_x \in B$ ayant un score ω_x :

— si $\omega_x \geq \omega_{gold}$, on ajoute un bonus de contre-traduction a à $\omega_j^{(k+1)}$;
— si $\omega_x \geq \frac{\omega_{gold}}{5}$, on ajoute un bonus (réduit) de contre traduction b à $\omega_j^{(k+1)}$;
— si $\omega_x < \frac{\omega_{gold}}{10}$, applique à $\omega_j^{(k+1)}$ un malus de contre traduction b ;

Les seuils et les valeurs des bonus/malus peuvent varier en fonction de la taille du graphe G. Pour notre expérience, nous avons empiriquement choisi $a = \frac{\omega_{gold}}{10}$ et $b = \frac{\omega_{gold}}{50}$. Une fois que l'ensemble de candidats termes Θ_{k+1} est constitué, on crée un nouvel ensemble mul-

tilingue Γ_{k+1} en réalisant l'union de Γ_k et Θ_{k+1}.
Comme indiqué plus haut, cet algorithme peut être appliqué plusieurs fois en donnant en entrée à la passe courante l'ensemble résultant de la passe précédente. Dans notre expérience, nous nous en sommes tenus à deux passes. Appliqué à l'ensemble des synsets des wordnets en 4 langues alignés pour l'occasion (c.f. section 10.1.3.2), cette routine a permi de créer un wordnet « jouet » du français pour lequel chaque candidat terme possède en outre de ses informations essentielles (terme, langue) un score supérieur à un seuil minimum.

10.2.1.2 Exemple

Cet algorithme a été appliqué et évalué pour la proposition de termes en français complétant un wordnet multilingue synset-aligné en quatre langues (bulgare, tchèque, anglais, roumain), mais ne comptant aucun terme français. Le graphe de traduction utilisé était la version 0.1 de YAMTG, comprenant alors des traductions et de synonymes extraits d'un ensemble de Wiktionnaires en 18 langues (tchèque, néerlandais, anglais, français, allemand, hébreu, indonésien, italien, japonais, coréen, polonais, portugais, roumain, russe, slovaque, espagnol, suédois et turc), ainsi que des liens interlangues de la Wikipédia française (cf. section 9.2).

TABLE 10.1 – Synset (v. 2.0) ENG20-04543367-n aligné en 5 langues après complétion. Les termes candidats sont suivis de leur score. La colonne du tchèque étant vide, elle a été omise.

Anglais	Roumain	Bulgare	Français
inequality triangle inequality(2.0) dissimilarity(1.5)	imparitate inegalitate	неравенство	inégalité (12.0) dissemblance (1.0) inéquation (1.0)

La figure 10.1 offre un exemple de sortie alignée pour le synset ENG20-04543367-n {inequality} . Pour ce cas de figure, l'ensemble des termes de références données en entrée algorithme lors de la première passe était $\Gamma_0 = \{$(inequality, eng, 100), (imparitate, ron, 100), (inegalitate, ron, 100), (неравенство, bul, 100)$\}$.

La requête de traduction du terme *inequality (eng)* dans le graphe a abouti à la proposition de sept candidats termes dans six langues différentes (allemand, italien, français, suédois, anglais et espagnol). Parmi ces derniers, trois candidats ont gagné le bonus de contre traduction a : *Ungleichung (deu), inégalité (fra), olikhet (swe).*

Aucune traduction a été trouvée pour les termes de référence *imparitate (ron), inegalitate (ron) et неравенство (bul).*

L'étape précédente a permis d'obtenir un nouvel ensemble Γ_1, plus gros que Γ_0 , étant donnés que sept candidats termes ont pu y être ajoutés. Pour la seconde passe de l'algorithme :

La recherche de traductions et contre-traductions pour le terme *inequality (eng)* a permis de mettre à jour les scores des sept candidats termes ajoutés plus tôt, en octroyant notamment un bonus au même trio de terme que durant la passe précédente.

La requête de traduction pour le nouveau candidat terme *Ungleichheit (deu)* (dont le score est de $0,8$) a permis d'améliorer le score d'un terme. Quatre nouveaux candidats termes (en turc, français, anglais et russe) sont ajoutés. Parmi ces derniers, le score du terme turc *eşitsizlik (tur)* gagne le bonus a.

La requête de traduction pour le nouveau candidat terme *Ungleichung (deu)* a permis de mettre à jour le score d'un terme. Deux nouveaux candidats termes (un russe et un français) ont ainsi été ajoutés à l'ensemble des candidats. Parmi ces derniers, le terme français *inéquation (fra)* remplit les conditions nécessaires pour faire gagner le bonus a à son score.

La requête de traduction pour le nouveau candidat terme *inégalité (fra)* , dont le score était de $12,0$ a permis d'appliquer des bonus aux scores de deux termes anglais.

La requête de traduction pour le nouveau candidat terme *olikhet (swe)* (score de $12,0$) a permis d'augmenter le score d'un terme, et parallèlement d'intégrer sans bonus un nouveau candidat terme en suédois.

Aucune traduction n'a été trouvée pour le terme *неравенство (bul)* et les cinq autres candidats termes ajoutés durant la première passe.

Au final, comme cela est visible dans le tableau 10.1 pour les langues qui nous intéressent ici, il a été possible de découvrir des candidats termes pour l'anglais et le français. Le candidat terme le plus approprié (*inégalité*) a reçu le meilleur score parmi les candidats termes en français. En ce qui concerne les candidats termes proposés en anglais, leur score faible montre qu'ils n'ont pu prétendre à aucun bonus durant les deux passes de l'algorithme. Il ne devrait donc pas, en toute logique, être accepté dans le synset.

10.2.1.3 Résultats et comparaison au WOLF

En fixant le seuil minimum sur le score pour la conservation d'un candidat terme à 30, cette méthode nous a permis de récupérer 10 568 candidats termes en français. Parmi ces derniers, 6 119 (58%) n'étaient pas présents dans le WOLF (inventaire de synsets PWN 2.0).

Le tableau 10.2 montre un échantillon aléatoire issu de ces candidats termes, alignés avec les informations contenues dans le PWN pour le synset correspondant, le score du candidat terme, son statut lors de notre évaluation manuelle (OUI si correct, NON sinon), ainsi que l'information indiquant si le candidat terme se trouvait déjà ou non dans le WOLF (OUI si le candidat terme est déjà dans le WOLF, NON sinon).

TABLE 10.2 – Exemples de candidats termes proposés par l'algorithme de contre-traduction pour différents synsets.

Candidat Français	Synset	Score	termes PWN dans le synset	définition PWN	Correct	Dans WOLF
harmonie	06738523-n	54.7	harmony, concord, concordance	agreement of opinions	OUI	NON
ivre	00879266-a	100.0	intoxicated, drunk	as if under the influence of alcohol	OUI	NON
alphabet	06096415-n	39.6	alphabet	a character set that includes letters and is used to write a language	OUI	NON
ensemble	00515754-b	66.3	together	in each other's company	OUI	NON
jeunesse	10099908-n	47.0	young person, youth, younker, spring chicken	a young person (especially a young man or boy)	NON	OUI
tête	08134688-n	100.0	head	the top of something	OUI	OUI
salamandre	03825556-n	35.6	poker, stove poker, fire hook, salamander	fire iron consisting of a metal rod with a handle ; used to stir a fire	NON	NON
périlleux	01991204-a	40.6	hazardous, risky, venturesome venturous	involving risk or danger	OUI	OUI
accord	06733497-n	34.7	agreement	the verbal act of agreeing	OUI	OUI
électricité	07054143-n	71.8	electricity	keen and shared excitement	NON	OUI

Afin de valider notre approche, nous avons procédé à l'évaluation manuelle de 400 candidats termes associés à leur synset. Cette évaluation s'est faite en fonction de deux paramètres : le premier, t, est un seuil appliqué au score de chaque candidat terme : fixer ce seuil à 30 permet de retenir tous les candidats, alors que lui donner une valeur supérieure permet d'écrémer les candidats ayant un score inférieur à t. Le second paramètre, n_{max}, et le nombre maximum de candidats termes pouvant être retenu pour chaque synset : si ce paramètre est fixé à 3, les candidats termes retenus pour un synset sont les trois candidats possédant les meilleurs scores (ou moins si ces scores sont inférieurs à t).

Pour chaque valeurs de t et de n_{max}, nous avons envisagé :

— la *précision* des candidats termes, qui correspond au rapport de candidats termes corrects sur le nombre total de candidats termes ; si l'on retient l'intégralité des 10 568 candidats ($t = 30$ and $n_{\text{max}} = \infty$), la précision et de 74, 1%.

— l'*estimation du nombre de candidats corrects*, que nous avons calculé en effectuant le produit du nombre total de candidats termes et de la précision, nous fait présumer que, toujours dans le cas où l'ensemble des candidats seraient retenus, on pourrait obtenir aux alentours de $10568 \times 0,741 = 6465$ candidat corrects. Parmi ces derniers, 6 119 ne sont pas dans le WOLF, et leur précision est de 65%. Ce résultat montre que cette

approche a permis de récupérer des candidats termes qui n'ont pas pu être générés par d'autres approches utilisées pour créer et étendre le WOLF précédemment, et ce avec une bonne précision malgré tout. Cette précision peut justement être améliorée en utilisant des paramètres plus stricts, comme le montrent les chiffres du tableau 10.3.

— Un « *rappel estimé* », mesuré par rapport aux 6 465 candidats corrects conservés dans le cas où les paramètres choisis ne filtreraient aucun candidat terme (c'est-à-dire avec $t = 30$ et $n_{\text{max}} = \infty$).

— Un « *f-score estimé* », se fondant sur les chiffres de précision et de rappel estimé.

Le tableau 10.3 donne les couples précision / nombre de candidats termes obtenus en faisant varier les valeurs des paramètres t et n_{max}. Comme attendu, augmenter la valeur des seuils n_{max} ou t provoque une diminution du nombre de termes candidats tout en améliorant la précision.

Pour $n_{\text{max}} = 1$ et $t = 50$, la précision atteint 86% pour 3 353 candidats termes. Parmi ces derniers, 1 601 candidats termes ne sont pas dans le WOLF ; leur précision est de 82%. Pour ce qui est de ces candidats n'apparaissant pas dans le WOLF, leur précision dépasse même les 89%, à comparer aux 87% de précision globale obtenus pour l'ensemble des candidats termes qui sont également dans le WOLF.

Toutefois, de tels paramètres sont très restrictifs. En se fondant sur les « f-scores estimés », les valeurs optimales des paramètres sont $(t = 30, n_{\text{max}} = 3)$. Avec ces valeurs, il est possible de conserver 10 403 candidats termes (soit la quasi totalité) avec une précision de $74,8$%, légèrement mais significativement plus que pour $n_{\text{max}} = \infty$. Une validation manuelle de ces candidats termes ajouterai au WOLF près de 7 780 termes corrects.

Le tableau 10.3 montre que les scores de précision varient avec t et n_{max}, ce qui confirme la pertinence des scores calculés par l'algorithme 1. En conséquence, un seuil t très élevé associé à un seuil $n_{\text{max}} = 1$ permet d'obtenir les meilleurs scores de précision. Par exemple, pour les paramètres $(t = 60, n_{\text{max}} = 1)$, la précision culmine à $90,5$%, en proposant 2 245 candidats termes parmi lesquels environ 2 031 sont corrects.

Nous avons examiné manuellement les candidats termes corrects n'étant pas dans le WOLF

TABLE 10.3 – Nombres de termes candidats retenus pour différentes valeurs des seuils t et n_{max}, associés à la mesure de *précision* correspondante.

	$t = 30$	$t = 40$	$t = 50$	$t = 60$
$n_{\text{max}} = 1$	$8362/77,3$	$5340/81,5$	$3353/85,6$	$2245/90,5$
$n_{\text{max}} = 3$	$10403/74,8$	$6298/80,6$	$3890/85,1$	$2582/89,6$
$n_{\text{max}} = \infty$	$10568/74,1$	$6357/80,3$	$3917/85,2$	$2594/89,6$

obtenus avec les paramètres ($t = 30, n_{\max} = 3$). Le plus souvent, ces candidats correspondent à des termes très ou moyennement fréquents, mais polysémiques, parfois très polysémiques. En d'autres termes, ces candidats représentent l'information la plus importante pour la plupart des usage que l'on peut vouloir faire d'un wordnet. Ces candidats termes représentent également les cas de figure pour lesquels les approches précédentes utilisées pour développer le WOLF ont le moins bien fonctionné.

L'approche fondée sur les alignements utilisée pour désambiguïser les termes polysémiques (Sagot & Fišer, 2008) est très fortement dépendante des corpus utilisés et échoue pour les termes moyennement fréquents. L'approche utilisant des lexiques était initialement restreinte aux termes monosémiques (Sagot & Fišer, 2008). Des efforts ultérieurs pour utiliser cette méthode sur des termes plus polysémiques (Sagot & Fišer, 2012) n'ont toutefois pas obtenus de résultats significatifs sur des termes très polysémiques. Des exemples de candidats termes proposés par l'algorithme 1 mais absents du WOLF incluent des termes à la fois basiques et fréquents comme *manger*, *taper*, *lent*, *faim* ou *dehors*.

Le wordnet jouet du français obtenu suite à l'application de cet algorithme aura permis d'intégrer de nombreux candidats termes à la version officielle du WOLF. Dans le WOLF version 1.0b4 (la dernière à ce jour), parmi les 117 658 synsets (homologues à ceux du Princeton WordNet 3.0), 56 479sont non-vides dans le WOLF. Parmi ces synsets contenant au moins un terme (littéral) français, 33 347 (59% des synsets non-vides) contiennent au moins un littéral qui a été proposé par l'approche à base de graphe de traduction (exemple de ce genre de synset à la figure 10.9) . Parmi eux, 8 999 (15,9% des synsets non-vides) ne contiennent que des littéraux qui n'ont été proposés que par cette approche, sans laquelle ils seraient restés vides. 1 536 sur ces 8 999 synsets (17%) ont 2 littéraux ou plus (exemple de ce genre de synset à la figure 10.10). Pour plus de détails à ce sujet, voir Sagot (2013).

FIGURE 10.9 – Exemple d'un synset du WOLF (v. 1.0b4) rempli à l'aide de différentes approches. Les techniques ayant proposé un littéral sont listées dans l'attribut « lnote » ; celle utilisant la contre-traduction pondérée est « lrec12mllexwn » le score final obtenu par notre algorithme étant indiqué entre parenthèses).

```
<SYNSET>
 <ID>eng-30-15271008-n</ID>
 <ILR type="hypernym">eng-30-15269513-n</ILR>
 <ILR type="eng_derivative">eng-30-00363493-v</ILR>
 <ILR type="eng_derivative">eng-30-00779061-v</ILR>
 <ILR type="eng_derivative">eng-30-02535716-v</ILR>
 <ILR type="eng_derivative">eng-30-02641035-v</ILR>
 <POS>n</POS>
 <BCS>2</BCS>
 <SYNONYM>
   <LITERAL lnote="7/1:fr.roen;gwa2012(0.309);lrec12mllexwn(3.589)">pause</LITERAL>
   <LITERAL lnote="gwa2012(0.266);lrec12mllexwn(1.368)">interruption</LITERAL>
   <LITERAL lnote="gwa2012(0.146)">repos</LITERAL>
   <LITERAL lnote="gwa2012(0.104)">intermission</LITERAL> <!-- erroné -->
   <LITERAL lnote="lrec12mllexwn(3.393)">trêve</LITERAL>
 </SYNONYM>
 <DEF>a time interval during which there is a temporary cessation of something</DEF>
</SYNSET>
```

FIGURE 10.10 – Exemple d'un synset du WOLF (v. 1.0b4) rempli à l'aide de l'algorithme de contre-traduction pondérée (attribut « lrec12mllexwn »).

```
<SYNSET>
 <ID>eng-30-14936010-n</ID>
 <ILR type="hypernym">eng-30-14935555-n</ILR>
 <ILR type="eng_derivative">eng-30-01603732-v</ILR>
 <SYNONYM>
   <LITERAL lnote="lrec12mllexwn(1.399)">chaux</LITERAL>
   <LITERAL lnote="lrec12mllexwn(1.354)">hydroxyde de calcium</LITERAL>
 </SYNONYM>
 <DEF>a caustic substance produced by heating limestone</DEF>
 <POS>n</POS>
</SYNSET>
```

:

10.2.2 Clustering via recuit simulé

L'heuristique décrite à la section 10.2.1 consistait à récupérer des traductions candidates dont les traductions se montraient elles-même pertinentes au regard de l'ensemble de termes de référence données pour initier la complétion. Dans un graphe orienté, cela permettait de favoriser les candidats termes faisant partie du voisinage immédiat des termes de référence. Afin de mieux formaliser ce mécanisme, il nous a semblé que la suite logique à donner à l'heuristique de contre-traduction résidait dans la détection de clusters d'intérêt dans le graphe de traduction, en fonction d'un ensemble de termes de référence. La section 10.2.2.1 explique le choix effectué, en s'appuyant notamment sur le tour d'horizon minutieux proposé par Schaeffer (2007) des méthodologies couramment appliquées pour les tâches de clustering dans les

graphes.

10.2.2.1 Justification du choix

Avec la perspective de développer la couverture de YaMTG 0.1 (puis YaMTG 1.0), plusieurs choix techniques ont dû être faits :

— en premier lieu, afin de permettre de garder une représentation de graphe la plus compacte possible et de simplifier les traitements ultérieurs, nous avons pris le parti de ne plus conserver l'information relative au sens des traductions, mais seulement leurs origines ;

— en second lieu, il a fallu s'assurer que d'éventuels algorithmes applicables à YaMTG offraient une garantie de scalabilité, c'est-à-dire qu'il seraient toujours applicables si la taille du graphe de traduction augmentait considérablement.

Les algorithmes de clustering appliqués à des graphes permettent de regrouper certains nœuds sur la base de leurs liens de proximité. De façon générale, on peut définir un cluster comme un ensemble de nœuds ayant des connexions internes (au sein du cluster) plus fournies que des connections externes : s'il n'existe aucun chemin interne à un cluster reliant deux nœuds, alors ces deux nœuds ne doivent préférablement pas appartenir à un même cluster (Schaeffer, 2007).

Un nombre important d'algorithmes de clustering (ainsi que des variantes) sont proposés dans la littérature, et chacun explore et découvre des structures topologiques variées. Il existe beaucoup de graphes aux propriétés topologiques hétérogènes, pour lesquels des « clusters naturels » ne sont pas facilement identifiables. C'est la cas par exemple pour les graphes dont la topologie tend à être uniforme. Même dans des cas où il est possible de distinguer intuitivement des clusters, il n'a jamais pu être établi de définition consensuelle des propriétés qu'un cluster doit satisfaire (Edachery *et al.*, 1999 ; Estivill-Castro, 2002). Reste, en ce qui nous concerne, que la nature de notre graphe de traduction nous permet de choisir, parmi tous les algorithmes existant, les plus élémentaires. En effet, notre graphe est simple, non dirigé, non pondéré et, comme nous l'avons vu précédemment, de type petit monde. Toutes ces caractéristiques sont de nature à faciliter la recherche de clusters, car la plupart des questions théoriques sur la nature des clusters à déterminer sont, dans notre cas, déjà en partie résolues.

Notre graphe de traduction contient près de 8.10^5 nœuds et 5.10^6 arcs dans sa version 1.0 (ces chiffres montant à 3.10^6 nœuds et 4.10^7 arcs pour YaMTG 2.2). Or, pour la complétion de terminologie multilingue structurée, nous ne sommes intéressés que par une portion infime des nœuds du graphe. Il n'est donc pas rentable d'y appliquer des procédés globaux. Pour cette raison, nous avons restreint notre choix à des méthodes de clustering locales, car ces dernières ne nécessitent qu'une vue partielle de la topologie du graphe. Ainsi les problèmes de

scalabilité à des graphes plus gros sont évités.

Les versions récentes de YAMTG n'étant pas dirigées, un algorithme de clustering symétrique [6] suffit : si un nœud n_1 appartient au cluster $\mathcal{C}(n_2)$, alors le nœud n_2 appartient au cluster $\mathcal{C}(n_1)$.

Il existe de nombreux algorithmes de clustering locaux utilisant des « fonctions de pertinence » (*fitness functions* en anglais) qui mesurent la qualité d'un cluster dans un graphe pour rechercher les clusters optimaux. Parmi les fonctions les plus utilisées (Šíma & Schaeffer (2006), Schaeffer (2007)), il y a :

— la *conductance* (aussi appelée *ratio de Cheeger*), qui calcule pour un cluster le ratio du nombre de ses connexions externes sur le total de ses connexions. Cette fonction a été utilisée pour calculer des clusters locaux notamment par Chung (2007), et pour développer l'algorithme Nibble (Spielman & Teng, 2013).

— les *densités locale et relative*, et les combinaisons possibles de ces deux mesures. La *densité locale* d'un cluster $\mathcal{C}$ calcule le ratio du nombre d'arcs effectivement présents dans $\mathcal{C}$ par le nombre maximum d'arcs possibles dans $\mathcal{C}$. La *densité relative* d'un cluster C dénote le ratio du degré interne de $\mathcal{C}$ (c'est-à-dire la somme des degrés de tous les nœuds $n \in \mathcal{C}$) par le nombre d'arcs incidents à tous les nœuds du cluster $\mathcal{C}$ (aussi bien internes qu'externes). Un cluster idéal maximise ses densités locale et relative. Schaeffer (2005) propose d'utiliser comme fonction de pertinence à maximiser le produit des densités locale et relative.

Une fois la fonction de pertinence choisie, la tâche de clustering s'apparente à un problème d'optimisation résolu par des algorithmes d'approximation (Estivill-Castro, 2002 ; Schaeffer, 2007) : on suppose qu'une valeur élevée pour une fonction de pertinence indique un cluster de qualité. Les algorithmes locaux qui nous intéressent ici cherchent le meilleur cluster qui contienne un nœud de départ (*seed*) s, qui constitue l'état initial du cluster $\mathcal{C}$. Reste à sélectionner la technique qui choisit quels nœuds ajouter à $\mathcal{C}$ en priorité et quand arrêter l'ajout de nouveaux nœuds pour calculer l'état optimal de $\mathcal{C}$ (celui qui maximise la fonction de pertinence). Là encore, les options possibles sont nombreuses, pouvant aller d'une sélection aléatoire à la sélection du nœud voisin ayant le plus fort degré ou l'ensemble des nœuds situés à des distance géodésiques successivement éloignées de la source (Fortunato, 2010).

Parce que l'algorithme de clustering local via recuit simulé (*Simulated Annealing* en anglais (Kirkpatrick *et al.*, 1983)) décrit dans Schaeffer (2005) repose sur un ensemble d'idées simples à implémenter et ne nécessitant aucun paramètre autre que ceux donnés à l'heuristique de recuit simulé, c'est celui que nous avons choisi d'utiliser pour procéder à la complétion de

6. Par opposition au clustering asymétrique, qui est par exemple essentiel pour traiter des graphes modélisant des réseaux sociaux ; dans ces derniers, les relations (comme l'« amitié ») peuvent être à sens unique : le graphe sera alors dirigé.

sysnsets. Ce choix a notamment été rendu possible par la structure petit monde de YAMTG, dont les clusters sont plus aisés à détecter.

10.2.2.2 Principes

Clustering via recuit simulé de Schaeffer (2005)

Le recuit simulé est une méthode probabiliste capable de trouver un optimum global d'une fonction pouvant avoir plusieurs optimums locaux. Cette heuristique a été inspirée par le processus éponyme utilisé en métallurgie pour permettre à des solides de refroidir suffisamment lentement pour permettre à leur structure de se figer en minimisant l'énergie du matériau.
Sachant une fonction de pertinence Φ, des paramètres t_0, α et I ainsi qu'une solution candidate initiale S, cet algorithme retourne une solution S' qui maximise la fonction Φ. Pour ce faire, chaque solution candidate évaluée par l'algorithme de recuit est acceptée si elle améliore le score de Φ, ou rejetée avec une probabilité de $\exp \frac{-|\Phi(\mathcal{C}')-\Phi(\mathcal{C})|}{t}$ (avec t la température courante) sinon. Cette possibilité d'accepter des résultats dégradant occasionnellement (et de moins en moins souvent) la pertinence permet d'éviter le confinement dans la recherche d'un optimum trop local.

Recuit simulé

$T \leftarrow$agenda_de_refroidissement(t_0, α)
pour $t \in T$ faire
  pour $i = 1$ à I faire
    $\mathcal{C}' \leftarrow$ mise-à-jour de $\mathcal{C}$
    si $\Phi(\mathcal{C}') > \Phi(\mathcal{C})$ OU rand$([0, 1]) < \exp \frac{-|\Phi(\mathcal{C}')-\Phi(\mathcal{C})|}{t}$ alors
      $\mathcal{C}' \leftarrow \mathcal{C}$
    fin si
  fin pour
fin pour

Dans notre cas, nous avons choisi l'ensemble des paramètres pour l'expérience relatée dans la section 10.2.2.3 en partant des choix standards, puis de façon heuristique. Concernant l'agenda de refroidissement, qui est une fonction décroissante représentant la baisse de la température, nous avons utilisé la version de Kirkpatrick *et al.* (1983), qui refroidit exponentiellement la température suivant un ratio α (ayant pour valeur $0,9$ ici) : $t_k = t_0 \cdot \alpha^k$. Ces auteurs avaient également fixé la température initiale t_0 à 10. Le paramètre I indiquant le nombre maximal d'itérations possibles a été fixé à 1000.

Schaeffer (2005) propose une mise à jour du cluster candidat $\mathcal{C}$ fondée sur l'information localement disponible dans le graphe $G = (S, A)$ comprenant un ensemble S de sommets (ou nœuds) éventuellement reliés par un ensemble A d'arcs . Pour l'obtenir, il s'agit de maintenir à jour en permanence les trois informations suivantes :

— la liste des sommets inclus dans $\mathcal{C}$;
— le degré interne de $\mathcal{C}$: $deg_{int}(\mathcal{C}) = |\langle u, v\rangle \in A|\ u, v \in \mathcal{C}$;
— le degré externe de $\mathcal{C}$: $deg_{ext}(\mathcal{C}) = |\langle u, v\rangle \in A|\ u \in \mathcal{C}, v \in S \setminus \mathcal{C}$.

Le cluster initial est composé que du nœud *seed* et de tous ses voisins directs dans le graphe. Par la suite, nous considérons [7] en priorité pour l'ajout au cluster $\mathcal{C}$ le sommet $v \notin \mathcal{C}$ apparaissant le plus fréquemment dans le voisinage dans les nœuds qui constituent $\mathcal{C}$. Puis pour confirmer ou infirmer son ajout dans $\mathcal{C}$, il faut :

— récupérer l'ensemble $\Xi(v)$ des voisins de v ;
— calculer les degrés internes et externes du nouveau cluster candidat $\mathcal{C}' = \mathcal{C} \cup \{v\}$.
  Soit $k = |\Xi(v) \cup \mathcal{C}|$ le nombre de voisins de v déjà présents dans $\mathcal{C}$. On a :
  • $deg_{int}(\mathcal{C}') = deg_{int}(\mathcal{C}) + k$;
  • $deg_{ext}(\mathcal{C}') = (deg_{ext}(\mathcal{C}) - k) + (deg(v) - k)$;

Une fois ces informations obtenues, la fonction de pertinence $\Phi(\mathcal{C}')$ (représentant le produit des densités locale et relatives du cluster) est calculable selon la formule :

$$\Phi(\mathcal{C}) = \frac{2 \cdot deg_{int}(\mathcal{C}')^2}{|C'| \cdot (|C'| - 1)(deg_{int}(\mathcal{C}') + deg_{ext}(\mathcal{C}'))}$$

qui est la fonction sur laquelle l'algorithme de recuit simulé appuie son optimisation afin de retourner l'estimation du meilleur cluster qui contient le nœud *seed*. Dans notre implémentation, cette fonction renvoie en résultat le cluster obtenu en même temps que la valeur de la fonction de pertinence Φ obtenue par ce dernier. Cette dernière prend une valeur comprise entre 0 (pire des cas) et 1 (clique isolée).

Inflation de synsets

À l'instar de l'algorithme de contre-traduction pondérée, celui-ci (algorithme 2) prend en entrée les termes initialement présents dans un synset, sans distinction de langue. Ces termes, qui constituent l'ensemble Γ_0 des nœuds feuille ayant un nœud structurel commun, sont des termes de référence (nœuds *gold*). Chacun de ces terme de référence va être envoyé à l'algorithme de clustering via recuit simulé en qualité de nœud *seed*. À ce titre, chaque synset est envisagé comme un ensemble de nœuds *seed*. Ainsi, si aucun des nœud de Γ_0 n'existe aussi

7. Schaeffer (*ibid.*) sélectionne aléatoirement n'importe quel sommet $v \notin \mathcal{C}$ relié à au moins un nœud de $\mathcal{C}$.

dans le graphe de traduction, il est impossible de proposer des candidats termes pour le synset courant et l'algorithme de complétion passe au synset suivant. En revanche, si pour un ou plusieurs nœuds de Γ_0 il existe un nœud représentant le même terme dans la même langue dans le graphe de traduction, alors il est possible d'appliquer l'algorithme de clustering à chaque nœud *seed* pour la complétion.

Pour chaque nœud *seed* qui représente un terme γ dans une langue l, on récupère le cluster $\mathcal{C}$ afférant grâce à notre implémentation de l'algorithme de clustering via recuit simulé de Schaeffer (2005). Ces nœuds prennent symboliquement un poids infini afin de pouvoir être facilement distinguables des candidats termes par la suite. Tous les nœuds n de $\mathcal{C}$ sont mémorisés, et le score qui est associé à n est mis à jour en additionnant au score précédent de n le score du cluster dont il est issu. De la même façon, des compteurs spécifiques à chaque langue sont incrémentés dès lors que le nœud n représente un terme dans cette langue. Une fois recensés tous les nœuds des clusters obtenus pour l'ensemble des termes *seed* pour le synset courant, une passe de normalisation leur est appliquée avant leur ajout au synset. Cette dernière dépend :

— du nombre de nœuds *seed* ayant été considérés pour la recherche des clusters ;
— du nombre de candidats termes dans une langue donnée ; Dans le cas où il existe de nombreuses traductions dans une langue, il est légitime de soupçonner qu'une ambiguïté sémantique est en cause.

Algorithme 2 Complétion avec l'algorithme de clustering via recuit simulé

Pré-conditions : Γ_0 : l'ensemble des nœuds *seed* d'un synset
  pour $(\gamma_i, l_i, \infty) \in \Gamma_0$ faire
    $(\mathcal{C}_{\gamma_i}$,score_$\mathcal{C}_{\gamma_i})$ $\leftarrow$ clusteringViaRecuitSimulé (γ_i, l_i)
    pour $(\theta_j, l_j) \in \mathcal{C}_{\gamma_i}$ faire
      Ajouter (θ_j, l_j) à la liste des candidats termes
      score$[(\theta_j, l_j)]+=$score_$\mathcal{C}_{\gamma_i}$
      compte$[l_j] + +$
    fin pour
    $k = |\Gamma_0|$
    $\Gamma_1 \leftarrow \emptyset$
    pour $(\theta, l) \in$ liste des candidats termes faire
      score_final $=$ score$[(\theta, l)]/(k \times$ compte$[l])$
      $\Gamma_1 \leftarrow \Gamma_1 \cup \{(\theta, l,$ score_final$)\}$
    fin pour
  fin pour
  $\Gamma_z \leftarrow \Gamma_0 \cup \Gamma_1$
  retourner Γ_z

10.2.2.3 Exemple

Cette algorithme a été appliqué et évalué pour la proposition de termes en français complétant un wordnet multilingue synset-aligné en 9 langues (bulgare, tchèque, anglais, hébreu, indonésien, japonais, malais, portugais brésilien et roumain [8]). Le graphe de traduction utilisé était la version 1.0 de YaMTG (décrite dans le chapitre 9).

Le tableau 10.4 reprend le même exemple que celui utilisé dans la section 10.2.1.2, mais complété avec l'algorithme de recuit simulé. Pour cet exemple, étant donné que le seul nœud *seed* trouvé dans le graphe correspondait au terme anglais *inequality*, la recherche de cluster n'a eu lieu qu'une fois. Pour cette raison, tous les candidats termes dans une langue obtiennent le même poids. Il s'agit d'ailleurs d'une limitation de cet algorithme.

TABLE 10.4 – Synset (v. 3.0) ENG30-04752221-n aligné en 9 langues après complétion. Les termes candidats sont suivis de leur score. Les colonnes de l'hébreu, de l'indonésien et du malais étant vides, elles ont été omises.

Anglais	Roumain	Bulgare	Tchèque	Japonais
inequality inequalities (7534,8) inequation (7534,8)	imparitate inegalitate	неравенство	nerovnice (22604,4)	不均等 不均衡 不平等 不等 凸凹 不同(22604,4)

Portugais	Français
as desigualdades (7534,8) desigualdade (7534,8) desigualdades (7534,8)	inégalité (5651.1) inégalités de revenu (5651.1) inégalités (5651.1) inéquation (5651.1)

Le tableau 10.5 présente une seconde illustration de la complétion de synset pour laquelle plusieurs nœuds *seed* ont pu être utilisés. Pour cet exemple : 4 nœuds *seed* (les termes anglais *avarice*, *cupidity*, *covetousness* et *avariciousness*) ont servi à déterminer des frontières de cluster.

On constate en premier lieu que les scores ne sont pas homogènes d'un synset à l'autre, et que la fonction qui se charge de leur calcul demanderait à être améliorée. Néanmoins, étant donné la disparité des ressources, il n'y a aucune raison que les scores soient comparables d'un sysnset à l'autre et d'une langue à une autre ; ainsi, il n'est pas envisageable de fixer un seuil global : l'approche de sélection des candidats sera locale à un synset et paramétrable pour chaque langue.

8. Les wordnets issus de BalkaNet (bulgare, tchèque, roumain) ont été convertis de 2.0 à 3.0 par un script de Tomaž Erjavec.

TABLE 10.5 – Synset (v. 3.0) ENG30-04945758-n aligné en 9 langues après complétion. Les termes candidats sont suivis de leur score. Les colonnes de l'indonésien et du malais étant vides, elles ont été omises. Les candidats termes qui sont des erreurs dans la graphe de traduction sont préfixés d'une astérisque.

Anglais	Roumain	Bulgare	Tchèque	Japonais	Portugais
avarice cupidity covetousness avariciousness greediness (206,5) avidity (651,2) *havesyke (444,8) *sannt (444,8) greed (851,4)	cupiditate (482,3) lăcomie (1780,1) avariție (431,7) zgârcenie (431,7) aviditate (1780,1)	алчност (5115,2)	chamtivost (3410,1) lakomství (1534,1)	強慾 強欲 欲ばり 欲張り 欲心 欲深 貪慾 貪欲	ganância (3916,3) avareza (920,4)

Hébreu	Français
קנאה(4924,5)	cupidité (2175,7) avidité (2175,7) avarice (511,3)

En second lieu, il est intéressant d'observer que plus de candidats termes pour des langues à moindre couverture que l'anglais ou le français dans le graphe ont pu émerger, avec une qualité variable. Sans avoir procédé à une évaluation concrète de ce phénomène, nous avons pu observer que cette approche permettait d'émuler grossièrement une inférence de liens de traduction.

10.2.2.4 Résultats

Au total, 30599 candidats termes ont été proposés pour le français durant la complétion, parmi lesquels 63% ne sont pas dans le WOLF. Le tableau 10.6 montre un échantillon aléatoire des candidats termes non filtrés issus de la complétion fondée sur le clustering. À l'instar du tableau 10.2 les candidats termes sont mis en correspondance avec les informations contenues dans le PWN pour le synset correspondant, le score du candidat terme, son statut lors de notre évaluation manuelle, ainsi que l'information indiquant si le candidat terme se trouvait déjà ou non dans le WOLF v. 3.0 (OUI si le candidat terme est déjà dans le WOLF, NON sinon).
Cette méthode a été évaluée sur un échantillon de 715 candidats au total validés manuellement. Deux paramètres vont varier pour l'évaluation : le seuil t_1 sur le score du meilleur candidat terme, et le seuil t_{2+} pour le reste des candidats.

Les résultats ci-dessous sont pour $n_{max} = 2$, c'est-à-dire qu'on ne s'autorise qu'à rajouter au plus deux candidats pour un même synset, afin de filtrer un maximum de bruit dans les

TABLE 10.6 – Exemples de candidats termes proposés par l'algorithme de contre-traduction pour différents synsets.

Candidat Français	Synset	Score	termes PWN dans le synset	définition PWN	Correct	Dans WOLF
chapeau de cow-boy	03124170-n	64180	ten-gallon hat, cowboy hat	a hat with a wide brim and a soft crown ; worn by American ranch hands	OUI	NON
las	02432428-a	9164,6	aweary, weary	physically and mentally fatigued	OUI	NON
adresser la parole à	00990655-v	807,5	address, accost, come up to	speak to someone	OUI	NON
comme prévu	00150243-b	3481,7	sure enough	as supposed or expected	OUI	NON
recuit	00402951-n	34176	annealing, tempering	hardening something by heat treatment	OUI	OUI
nitrate de sodium	14699441-n	110697	caliche	nitrate-bearing rock or gravel of the sodium nitrate deposits of Chile and Peru	NON	NON
emboutir	01531265-v	1222,8	emboss, boss, stamp	raise in a relief	NON	OUI
rocheux	00747910-a	21574	rocky, rough	full of hardship or trials	NON	NON
plongée sous-marine	00326677-n	19959	dive, nose dive, nosedive	a steep nose-down descent by an aircraft	NON	NON
émailler	01374587-v	753	slosh, slush, slosh around, slush around	spill or splash copiously or clumsily	NON	NON

propositions de candidats termes. Comme nous l'avons vu, l'approche de sélection de candidats est locale, et effectue en l'occurrence un choix spécifique au français.

L'expérience montre que les candidats ayant un score très élevé sont souvent faux, probablement en raison de mots très polysémiques et donc à fort degré de connectivité dans le graphe de traduction. On se donne donc un t_{max}, score maximum autorisé pour un candidat (ceux de score plus élevé sont éliminés). Nous avons fixé empiriquement t_{max} à 100 000.

Nous avons par ailleurs deux valeurs de score minimum t_1 et t_2, une pour le meilleur candidat (t_1) et une pour le second (t_2). Par exemple, si on a pour un synset donné un candidat de score 12 000 suivi d'un candidat de score 5 000, et que $t_1 = 5000$ et $t_2 = 10000$, on ne gardera que le meilleur candidat et on éliminera le second. Si on veut obliger à n'avoir qu'un seul candidat, il suffit de fixer le score minimum t_2 du deuxième candidat à une valeur supérieure à la valeur de score maximum t_{max} (noté $t_2 = \infty$). En revanche, $t_1 > t_{max}$ ne conserverait aucun candidat.

Enfin, il serait peu pertinent d'avoir $t_2 < t_1$ (c'est-à-dire d'être plus permissif pour le second candidat que pour le premier), car cela pourrait conduire à rejeter le meilleur en conservant le second. Ces configurations ne sont donc pas étudiées.

Plusieurs stratégies de réglage des paramètres sont possibles : deux stratégies proposant moins de candidats mais de meilleure qualité (stratégies rapides), et une stratégie plus large offrant plus de candidats mais un peu moins bons. Les stratégies rapides consistent à ne prendre qu'au plus un candidat par synset ($t_2 = \infty$) et un seuil t_1 à respectivement 2000 (un peu plus de candidats un peu moins bons) ou 15000 (un peu moins de candidats, mais ils sont meilleurs). On obtient alors respectivement une estimation de 10785 candidats corrects parmi 15132 candidats (précision de 71, 27%) ou 8356 candidats parmi 11027 (précision de 75, 78%). La stratégie large consiste à poser $t_1 = 1000$ et $t_2 = 2000$. On obtient alors 14318 candidats corrects parmi 20988 (précision de 68, 22%).
Bien que les scores ne soient pas comparables en raison de différences entre les ressources utilisées entre cette expérience et celle décrite dans la section 10.2.1 , on peut constater qu'on obtient environ le même nombre de candidats que dans les résultats du tableau 10.3 (p. 215) en mode $t_{max} = \infty$ et $t = 30$, soit environ 10500, avec une précision de l'ordre de 76% (au lieu de 74%). À l'inverse, cette précision de 74% est obtenue avec environ 12000 candidats au lieu de 10500. Dans tous les cas, ces scores sont pour $t_2 = \infty$ ou proches de l'infini (presque toujours un seul candidat par synset).

10.3 Conclusion

Nous avons considéré, lors de la réflexion qui a précédé les travaux exposés ici, que des solutions réellement multilingues pour la tâche qui nous intéressait ici ne pouvaient être atteignables qu'avec des données de traduction récoltées sur Internet. Dans notre cas, nous nous sommes limités à récupérer ces dernières depuis des ressources libres, et à les compiler dans un large graphe de traduction fortement multilingue, dont la couverture (en particulier pour des langues ne faisant pas partie du « pôle » indo-européen) a encore une marge d'amélioration

TABLE 10.7 – Nombres de termes candidats retenus pour différentes valeurs des seuils t_1 et t_2, associés à la mesure de *précision* correspondante.

	$t_1 = 1000$	$t_1 = 2000$	$t_1 = 15000$	$t_1 = 50000$
$t_2 = 2000$	20988/68,22	20357/68,41	—	—
$t_2 = 20000$	18276/68,27	17729/68,86	13156/72,36	—
$t_2 = \infty$	15637/70,72	15132/71,27	11027/75,78	4859/76,66

considérable.

L'algorithme de contre-traduction pondérée (décrit p. 211) a été utilisé avec succès pour étendre le WOLF. Les deux algorithmes de complétion ont également été utilisés à divers degré pour compléter la terminologie multilingue relative au traitement d'enquête dans le domaine de Ressources Humaines développée par l'entreprise *Verbatim Analysis – VERA* (présentée dans la section 6.2.1, p. 109). À l'issue de ces expériences, nous considérons que la question posée à la section 10.1 a effectivement trouvé une réponse possible ici : la complétion de terminologie structurée semblables à celles schématisée dans la figure 10.1 (b) (p. 201) peut effectivement s'envisager comme un cas particulier d'une problématique plus globale relative à la construction et à l'enrichissement de structures ontologiques légères (comme celle schématisée dans la figure 10.1 (a)), et cela grâce au fort aspect multilingue des données.

D'autre part, notre ambition première pour la complétion de terminologies multilingues structurées était de garder des traitements simples. C'est notamment pour cette raison que la première heuristique de complétion par contre-traduction pondérée (algorithme 1) a été développée. La seconde heuristique (algorithme 2) visait quand à elle à explorer dans quelle mesure une généralisation du principe de contre-traduction (en l'occurrence, étendu à un cluster) était un atout pour la proposition de candidats termes. Les évaluations produites sur ces deux expériences, même si elles ne sont pas parfaitement comparables, semblent indiquer que dans l'espace de recherche de traductions pour la proposition de nouveaux candidats termes, il est préférable de rester local. En effet, l'algorithme 2 fondé sur un clustering local utilise plus de données pour la complétion que l'algorithme 1, (wordnets alignés en 9 langues et YaMTG v. 1.0 *vs.* wordnets alignés en 4 langues et YaMTG v. 0.1). Néanmoins, le nombre de candidats termes ainsi que les scores de précision estimés (présentés dans les tableaux 10.3 (p. 215) et 10.7 (p. 226) et récapitulés sous forme graphique dans la figure 10.11) indiquent que l'algorithme 2 affiche globalement de moindres performances. La méthode de calcul des scores que nous avons essayée ici est inadaptée : la récupération du contenu d'un seul cluster pour un petit nombre de nœuds *seed* donnera le même poids au « bruit » qu'aux traductions pertinentes. Par ailleurs, plus le cluster sera gros, plus il y aura des chances qu'il contienne du bruit. Dans notre proposition de calcul de score, nous n'avons pris en compte, pour le calcul du score d'un candidat terme, que les scores (cumulés) des clusters dont ils est issu, avant d'y appliquer un facteur normalisant relatif au nombre de nœuds *seed* et au nombre total de candidats proposés dans cette langue. Il serait également possible, afin d'améliorer la pertinence de ces scores, de garder une indication sur la distance géodésique des candidats termes par rapport aux nœuds *seed*. Nous en concluons que, même sur un graphe de traduction fortement multilingue de type petit-monde, le clustering simple n'est pas une méthode adapté, et qu'il faut se restreindre à des approches plus locales encore, à l'instar de l'algorithme de contre-traductions pondérées.

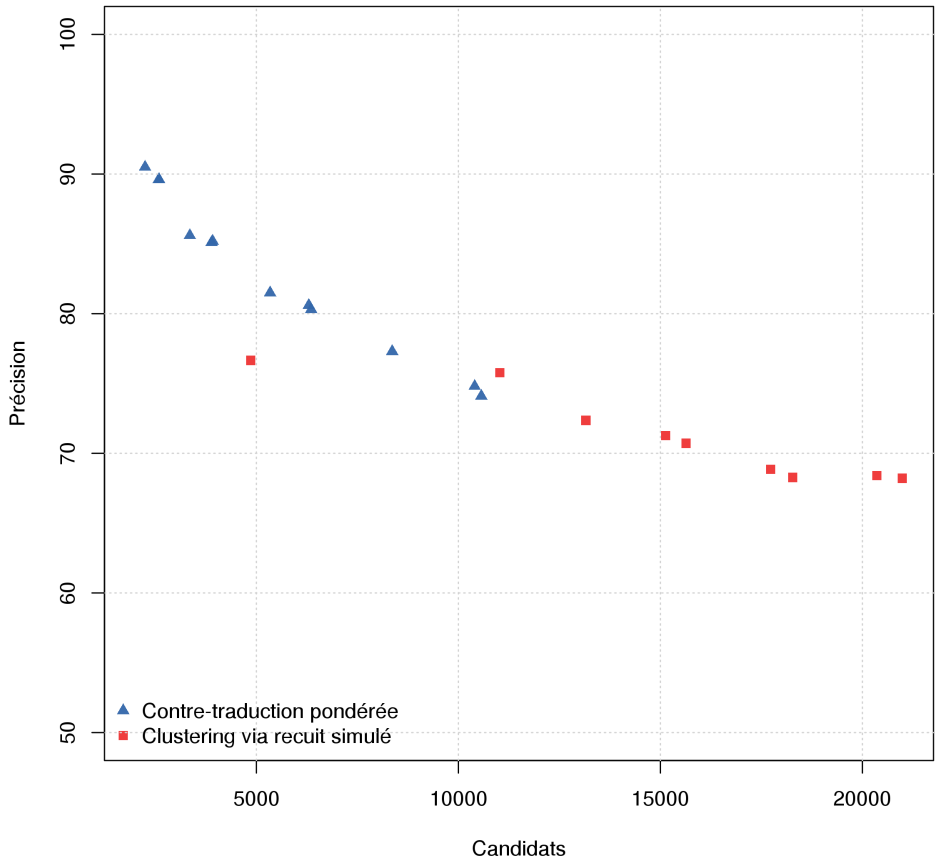
Le seul avantage que nous concevons pour l'application de cette technique pourrait être son utilisation dans la proposition de termes dans une langue « marginale » (très peu connectée à des termes dans d'autres langues) dans le graphe de traduction.

Hormis ce cas de figure, les algorithmes qui pourraient permettre une meilleure exploitation de YaMTG pour la proposition de candidats termes seraient donc ceux capables de contraindre l'espace de recherche des solutions dans le graphe à des zones topologiquement proches des nœuds *seed*. Ces alternatives pourraient prendre la forme, par exemple :

— de diffusion d'unités de signaux, comme proposé par Hu *et al.* (2008), mais avec un objectif restreint à quelques itérations et non à la découverte de clusters. En prenant pour nœuds source des signaux les nœuds *seed*, et en propageant l'émission des unités de signaux aux voisins à chaque étape (en gardant les comptes pour chaque nœud du nombre de signaux reçus), cette option formaliserait l'algorithme 1 de façon plus élégante et serait applicable à un graphe non dirigé.
— d'un calcul de confluence (Gaume, 2008 ; Gaillard *et al.*, 2011) entre des nœuds atteignables par une courte marche aléatoire (idéalement de longueur 3) depuis les nœuds *seed*.
— dans une moindre mesure, de marches aléatoires multi-agent (aussi appelées *spider walks*) très brèves (pour contraindre la localité) et reliées par une corde courte, dont les agents seraient originellement situés sur le ou les nœuds *seed* proches dans le graphe (Alamgir & Von Luxburg, 2010).

Enfin, un couplage avec les résultats de la partie II serait envisageable pour permettre de conforter certains candidats termes, de par le fait qu'ils ont été suggérés comme termes potentiels dans un corpus du domaine considéré. Cette piste sera évoquée de nouveau dans le chapitre de conclusion générale.

FIGURE 10.11 – Récapitulatif des couples précision/nombre de candidats estimés pour les différents paramètres scrutés lors de l'évaluation des deux algorithmes de complétion.



Quatrième partie

Synthèse

Chapitre 11

Conclusion générale

Sommaire

Le travail présenté dans cette thèse a débuté avec l'idée de développer deux axes relatifs à la construction et à la maintenance de terminologies multilingues. En premier lieu avec la construction d'un outil d'extraction terminologique pouvant traiter de façon acceptable des données textuelles relativement bruitées dans un grand nombre de langues, y compris des langues typologiquement éloignées. En second lieu avec le développement d'un module de proposition de nouveaux candidats termes, également indépendant de la langue.

11.1 Le multilinguisme est-il réellement une difficulté ?

Pour nourrir notre réflexion, nous avons abordé dans la partie I des questions théoriques et pratiques relatives à l'extraction terminologique. Nous avons également formulé un ensemble de questions typologiques soulevées par une approche multilingue.

La théorie terminologique a été construite sur la problématique, déjà présente au XVIII^{e} siècle, de la création de néologismes associée à une circulation rapide de l'information. D'abord normative et très cadrée, elle a été l'objet de remises en questions de grande ampleur afin de pouvoir s'adapter aux défis offerts par la production textuelle généralisée *via* l'usage d'outils numériques. La terminologie sert désormais aussi à comprendre et mettre en rapport les codes langagiers de différents groupes sociaux. Cette remise en question permet de prendre en considération la réalité d'usage des termes quelles que soient les circonstances (chapitre 1). Néanmoins, l'usage actuel fait de la terminologie dans la majorité des cas ne concerne que la création de ressources avec des méthodes monolingues ou adaptables à faible coût à des langues proches (chapitre 2). À notre connaissance, il n'existe pas de réflexion axée sur une ouverture réellement multilingue (c'est-à-dire à visée d'indépendance de la langue) des traitements d'extraction terminologique automatique, même si cette problématique a commencé à susciter récemment un intérêt (Daille & Blancafort, 2013). Or cette perspective exige de changer d'approche générale. Elle invite à une réflexion fondée sur des outils pouvant être appliqués d'une langue à l'autre de façon très générique. Cela nous a conduit à envisager une approche reposant sur une typologie des langues (chapitre 3), comme notamment préconisé par Bender (2011).

Nous avons testé notre système sur sept langues de familles linguistiques diverses (cf. tableau 4.1, p. 75). Les constats et réflexions développés dans la partie I ont orienté la suite de nos expérimentations autour de deux axes d'exploration concernant :

1. la nature des unités de traitement élémentaires à envisager pour l'extraction terminologique à travers les langues :
    a) obtenues soit par *tokenisation* (section 5.1.1), soit à l'aide d'un outil de segmentation (section 5.1.2) ;

   b) puis éventuellement amputées d'une partie de leurs unités morphologique (section 5.2), suivant différents degrés de « sous-spécification » (modérée ou franche).

2. l'entraînement, l'application et l'évaluation de modèles CRF génériques d'extraction terminologiques au sein d'une même langue et entre les langues :

   a) utilisant pour l'entraînement des modèles, des caractéristiques indépendantes de la langue (présentées dans la section 6.1.2, p. 101), puis normalisées et discrétisées (section 6.1.3, p. 103),

   b) et des données morphologiquement segmentées (correspondant à ce que nous avons appelé sous-spécification modérée), lorsqu'elles existent, pour l'adaptation des mesures de précision et de rappel terminologique à travers les langues (section 7.2).

Il ressort de cet ensemble d'expérimentations une infirmation de deux de nos hypothèses de départ. La première concerne les traits utilisés pour l'entraînement des modèles. Les caractéristiques comparant les fréquences de termes en corpus générique et en corpus de spécialité n'apparaissent que minoritairement dans les meilleurs modèles ; en terme de temps de traitement et de coût de développement, il n'est donc pas intéressant de procéder à une comparaison des fréquences entre corpus générique et de spécialité pour l'application de tels modèles CRF pour l'extraction de termes de domaines non-techniques. La fréquence en corpus de spécialité est quand à elle la seule caractéristique numérique présente dans l'ensemble des meilleurs modèles (toutes expériences confondues), quelle que soit la langue considérée. Toutefois, une utilisation isolée de cette caractéristique pour l'extraction ne mène pas à de bons résultats : il faut lui associer un ou plusieurs autres traits.

En second lieu, la sous-spécification telle que nous l'avons envisagée dégrade les résultats, quelles que soient les langues. Les scores de précision sont particulièrement affectés par cette opération. Toutefois, la sous-spécification dégrade moins les scores issus des modèles de l'arabe et du turc que ceux du polonais. En ce qui concerne la portabilité des modèles entre les langues, les modèles entraînés sur les langues exhibant des phénomènes morphologiques plus complexes (turc, et dans une moindre mesure, polonais, arabe et français) obtiennent souvent de bons scores terminologiques sur d'autres langues (section 8.2). Le résultat le plus surprenant concerne l'application des modèles de différentes langues aux donnée chinoises. Ces dernières tolèrent mieux des modèles non sous-spécifiés issus de l'allemand ou de l'arabe, du turc (modèle modérément sous-spécifié) et éventuellement du polonais : toutes ces langues exhibent non seulement des phénomènes morphologiques complexes et variés, mais qui plus est l'ordre canonique de leurs éléments dans la phrase n'ont pas une préférence pour la configuration SVO, à l'instar du chinois. L'ordre des composants dans la phrase donnée par les typologies n'influe pas sur la portabilité des modèles, mais l'information morphologique est

une source d'information importante pour leur entraînement, même pour une application sur des langues morphologiquement moins complexes.

Bien que ces expériences nous aient permis de réfuter ces hypothèses, il n'en reste pas moins que des modèles CRF produits en utilisant des caractéristiques indépendantes de la langue aboutissent à une extraction de termes de terrain de qualité satisfaisante. Les meilleurs scores obtenus (pour une évaluation monolingue des modèles) pour la majorité des langues se situent au dessus de l'iso-ligne de f-score $0,9$ (et même au dessus de $0,95$ pour l'arabe, le polonais et le turc). Seul le meilleur modèle chinois se situe entre les iso-lignes $0,8$ et $0,85$ (figure 8.2 p. 153). Or ces scores ont pu être améliorés pour quelques langues (allemand, français, polonais) grâce à l'application trans-lingue des meilleurs modèles de certaines langues (voir la section 8.2). Alors que l'on aurait pu penser le multilinguisme comme un frein à l'extraction terminologique automatique, il s'avère, avec cette approche tout du moins, que ce paramètre est potentiellement un excellent levier à des extractions terminologiques pour des langues ne disposant pas de leurs propres modèles.

La méthodologie générale de complétion multilingue abordée dans la partie III s'appuie d'une part sur la collecte et la mise en relation d'un grand nombre de traductions issues de ressources lexicales libres (YaMTG), et d'autre part sur une terminologie multilingue structurée [1]. Les expériences de complétion relatées dans le chapitre 10 ont indiqué qu'il est possible d'améliorer simplement des ressources terminologiques disposant de correspondances multilingues (WOLF, la terminologie RH de l'entreprise VERA).

Là encore, alors que le fort multilinguisme aurait pu constituer un inconvénient, il s'avère que ce paramètre peut au contraire constituer un réel atout sous certaines conditions. En premier lieu concernant le graphe de traduction, à défaut de disposer de toutes les traductions possibles dans l'ensemble des langues, il est nécessaire que les langues les mieux couvertes disposent de forts liens de traductions avec plusieurs autres langues, si possible issues de « pôles socio-culturels » disjoints (cf. section 9.3.4).

Les expériences décrites dans le chapitre 10 ont été menées principalement sur des langues européennes, très bien couvertes par YaMTG. Nous avons conscience qu'à cet égard, les résultats obtenus ne seraient sans doute pas aussi bon pour d'autres langues, en particulier pour celles disposant de peu de ressources, ou dans le cadre d'un vocabulaire très spécialisé. Il serait néanmoins possible d'augmenter la couverture de YaMTG dans ces langues ou domaines d'intérêt en y ajoutant des traductions issues de ressources supplémentaires.

1. Concernant l'organisation des termes, la section 10.1 a démontré que ces derniers peuvent être structurés de façon cohérente aussi bien en terme de regroupements conceptuels univoques que d'un point de vue applicatif, empirique.

Par ailleurs, ajouter de nouvelles ressources dans le graphe de traductions entraîne une autre amélioration, plus subtile. Même pour des couples de langues très bien traduits, il manquera toujours certains liens de traduction. Koehn & Knight (2001) ont par exemple déterminé, en ce qui concerne le couple de langue anglais-allemand, que seuls 68% des mots allemands (mais également 68% des noms allemands) possédaient une traduction anglaise pouvant être trouvé dans un dictionnaire bilingue adéquat. On peut imaginer qu'il soit possible d'améliorer sensiblement cette proportion dans un cadre fortement multilingue.

11.2 Contributions et productions

11.2.1 Contributions

La première contribution de ce travail de thèse a consisté en une prise en compte des paramètres relatifs à l'indépendance de la langue, *via* l'étude et l'expérimentation sur l'influence des typologies linguistiques dans le cadre d'une approche semi-supervisée pour l'extraction terminologique. Notre approche a investigué différentes hypothèses typologiques pour la sélection d'unités de traitement élémentaires préalablement au calcul des traits pour l'entraînement des modèles (segmentation en unités, « sous-spécification sémantique »).

La seconde contribution a concerné la réflexion autour des traits (ou caractéristiques) utilisés par les modèles. Nous avons investigué une façon de garder des traits vraiment indépendants de la langue et de l'influence des la taille des corpus. Pour ce faire, nous avons proposé d'appliquer des pré-traitements numériques connus à un ensemble de métriques (section 6.1.3). À aucun moment nous n'avons utilisé d'étiqueteur morpho-syntaxique ou d'analyseurs syntaxiques pour l'extraction de termes. Ces efforts ont permis de confirmer qu'il est possible d'entraîner un modèle CRF sur les données d'une langues pour l'appliquer avantageusement à une autre langue, même typologiquement éloignée.

Une troisième contribution a consisté à proposer un amendement à l'algorithme d'évaluation de Nazarenko *et al.* (2009), pour une meilleure adaptation à une utilisation multilingue (section 7.2). La différence réside principalement dans :

— la comparaison des UTE modérées [2] correspondant aux tokens des candidats termes, plutôt que les tokens directement, afin de procéder à l'identification des variantes de termes ;

2. Pour rappel, les UTE modérées ont été obtenues suite à une segmentation morphologique avec l'outil Morfessor, avec l'élimination des affixes présumés les plus productifs dans une langue (voir la section 5.2). Cette étape ne s'applique que pour les langues sur lesquelles il est intéressant d'appliquer cette simplification.

— la prise en compte du fait qu'un candidat terme puisse être lié à plusieurs éléments de la terminologie de référence (avec une une approche d'inspiration probabiliste), en abandonnant l'idée de partition proposée par Nazarenko *et al.* (2009) ;
— l'exclusion des principaux *stopwords* pour limiter l'influence de ces tokens dans le calcul des scores.

Enfin, nous avons montré qu'il était possible d'utiliser une méthode simple de complétion de terminologie structurée (ou d'ontologie legère généraliste) reposant sur un gros graphe de traduction, capable de mettre à profit des données fortement multilingues.

11.2.2 Productions

Ce travail de thèse a été l'occasion de produire deux types de données utiles à la communauté :

— différentes versions du graphe multilingue YAMTG, mises à disposition librement en plusieurs formats (GML [3] et TSV [4]) à l'adresse http://alpage.inria.fr/~hanoka/yamtg.html.
— la proposition de nouveaux termes (pour l'équivalent de 59% des synsets non vides du WOLF, parmi lesquels près de 16% ne contiennent que des termes proposés par notre approche), finalement inclus dans le WOLF version 1.0b4 (la dernière à ce jour). Cette inclusion valide notre proposition pour la complétion de terminologies multilingues structurées fondée sur le principe de la contre-traduction pondérée (section 10.2.1).

11.3 Perspectives

Les perspectives possibles pour l'amélioration de ces différents modules et leur assemblage sont nombreuses. Concernant le module d'extraction terminologique, il serait intéressant d'investiguer d'autres axes :

— *la question du déséquilibre des données lors de l'entraînement* (cf. section 7.1). Soit en utilisant un protocole permettant un meilleur rééquilibrage des données, soit en employant, au lieu des CRF, un classifieur mieux adapté aux données très déséquilibrées (Longadge & Dongre, 2013). S'adresser à cette problématique permettrait l'obtention de plus de termes complexes.
— *le choix de l'intervalle utilisé pour la discrétisation des caractéristiques numériques* (section 6.1.3.2). Notre choix de diviser toutes les plages numériques en $k = 5$ intervalles relevait d'un souci de rapidité de traitement étant donné le nombre d'expériences

3. Acronyme désignant le format *Graph Modelling Language*
4. Acronyme désignant le format *Tab-separated values*, dans lequel les valeurs sont séparées par des tabulations

à mener. Il serait néanmoins intéressant de faire quelques essais sur les meilleurs modèles en faisant varier k afin de déterminer l'influence de ce paramètre sur les résultats.

— *au regroupement de modèles (pooling).* Un avantage considérable offert par les modèles CRF consiste à pouvoir regrouper des modèles pré-existant en un seul. Il peut être intéressant de regarder dans quelle mesure le regroupement de modèles influe sur la qualité et la nature des termes extraits.

— *au filtre et au classement de candidats termes.* Nos expériences se sont limitées à l'extraction brute de termes et à leur évaluation. Nous n'avons pas proposé explicitement de méthode de filtrage, de classement ou de structuration des termes candidats. Or, il apparaît pertinent de mettre à profit l'heuristique d'évaluation proposée à la section 7.2 afin de filtrer et classer un premier échantillon de termes susceptibles d'être corrects. Une autre piste intéressante, mais non abordée dans cette thèse, consiste à tirer parti du fort multilinguisme de la terminologie ainsi extraite afin de regrouper et valider plus aisément des candidats termes, grâce notamment aux liens de synonymie et de traduction de YaMTG. En effet, des candidats termes présents dans plusieurs langues et possédant des termes sémantiquement proches également présents parmi les candidats termes d'autres langues langues ont plus de chance d'être corrects que des candidats termes peu représentés dans l'ensemble des langues.

En ce qui concerne le module de complétion multilingue, nous bornerons les perspectives d'améliorations à celles proposées dans la conclusion du chapitre 10 (p. 226). Nous y proposons d'examiner des algorithmes exploitant YaMTG en contraignant l'espace de recherche des solutions dans le graphe à des zones topologiquement proches des nœuds *seed*. Une seconde piste possible, que nous avons renoncé à aborder dans cette thèse de crainte de propager des erreurs dans le graphe, pourrait consister à induire des liens manquants dans YaMTG.

Cinquième partie

Annexes

Annexe A

Mesures d'associations : Courbes de densité

FIGURE A.1 – Courbes de densités de ZS, ODR, FAG, MD, CP et USUB en arabe et CP et USUB en allemand sans pré-traitement (corpus de spécialité).

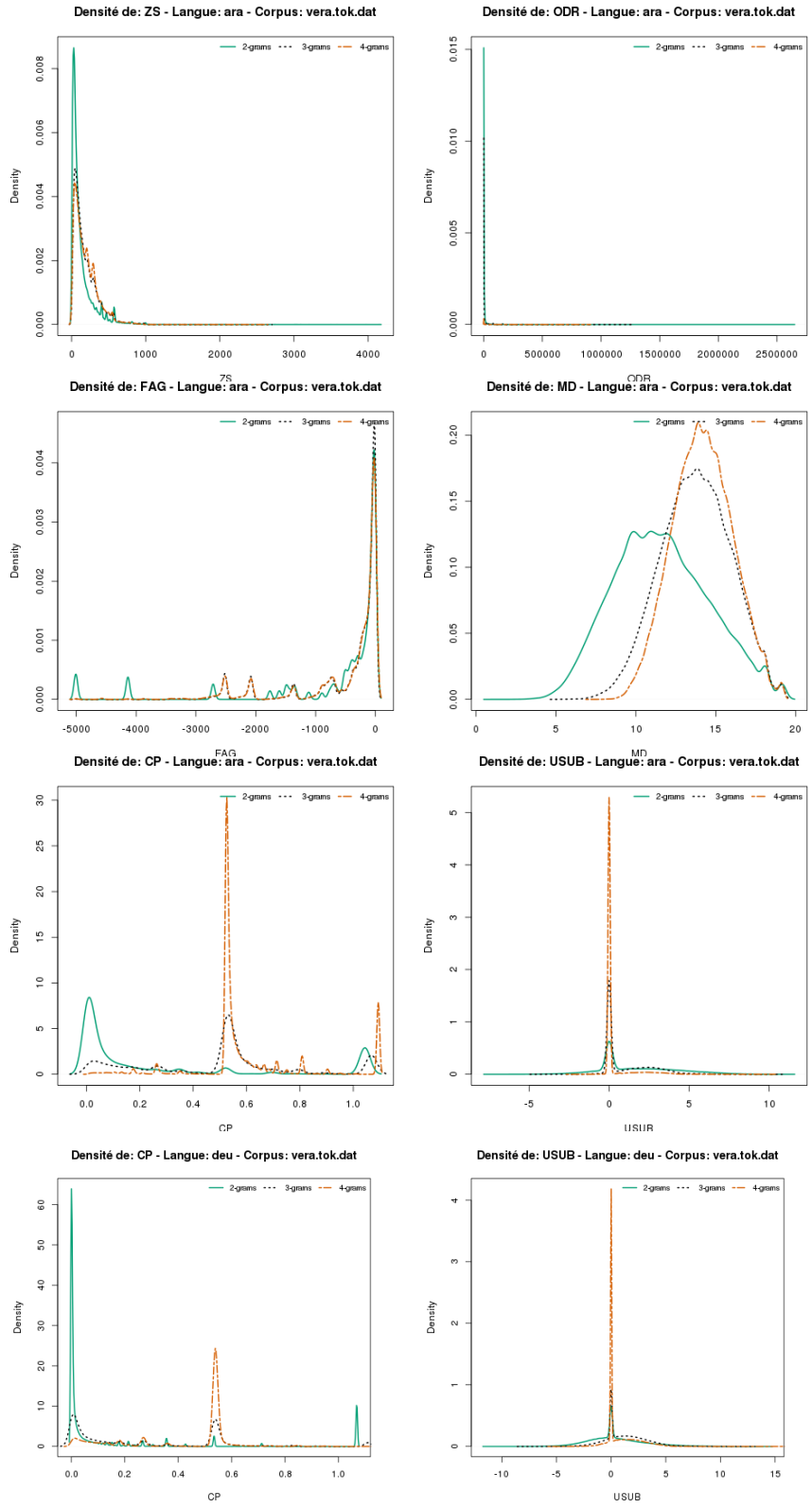


FIGURE A.2 – Courbes de densités de ZS, ODR, FAG, MD, CP et USUB en arabe et CP et USUB en allemand avec normalisation basée sur le z-score (corpus de spécialité).

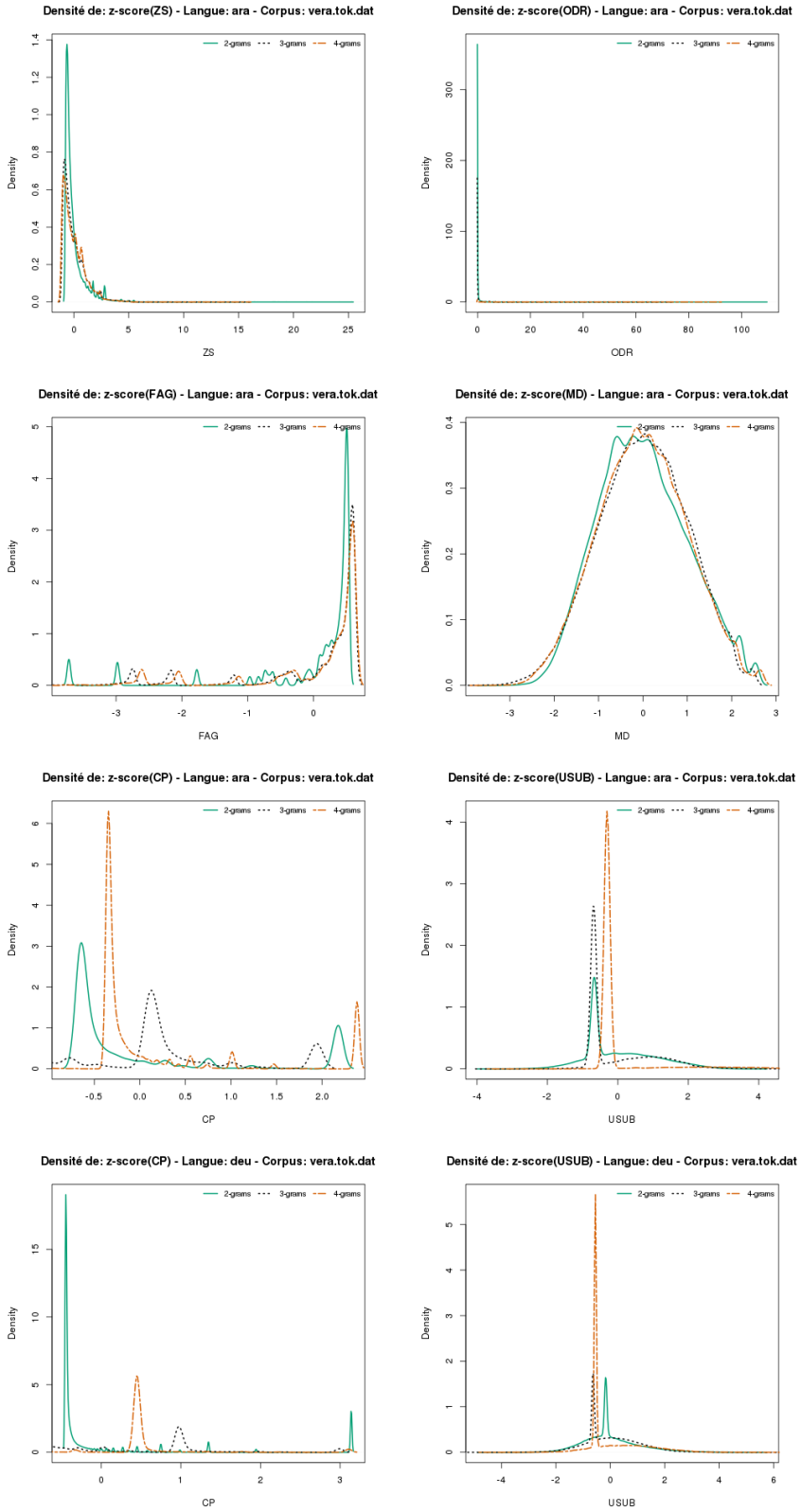


FIGURE A.3 – Courbes de densités de ZS, ODR, FAG, MD, CP et USUB en arabe et CP et USUB en allemand avec normalisation Min-Max (corpus de spécialité).

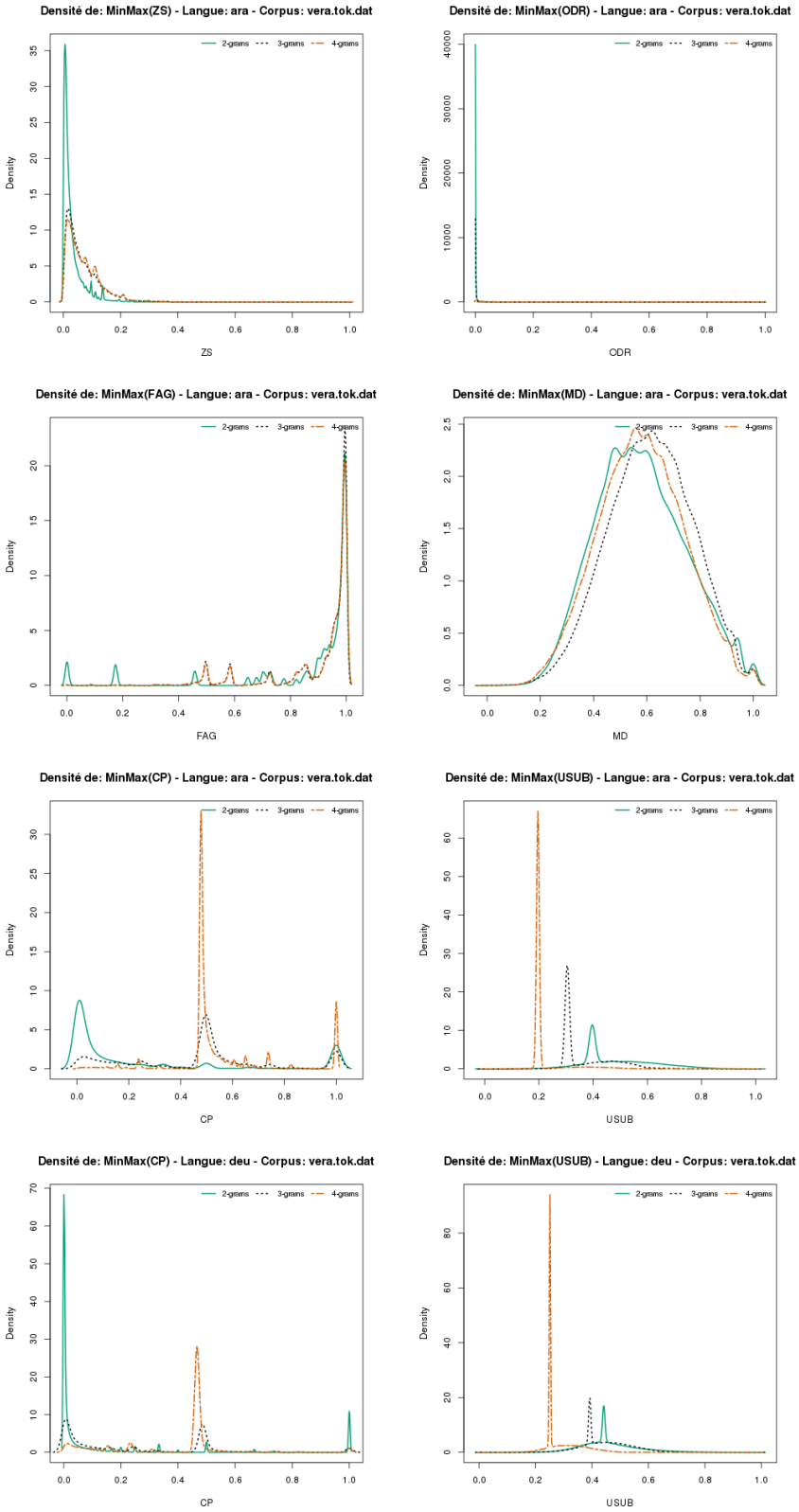


FIGURE A.4 – Courbes de densités de ZS, ODR, FAG, MD, CP et USUB en arabe et CP et USUB en allemand avec normalisation par mise à l'échelle décimale (corpus de spécialité).

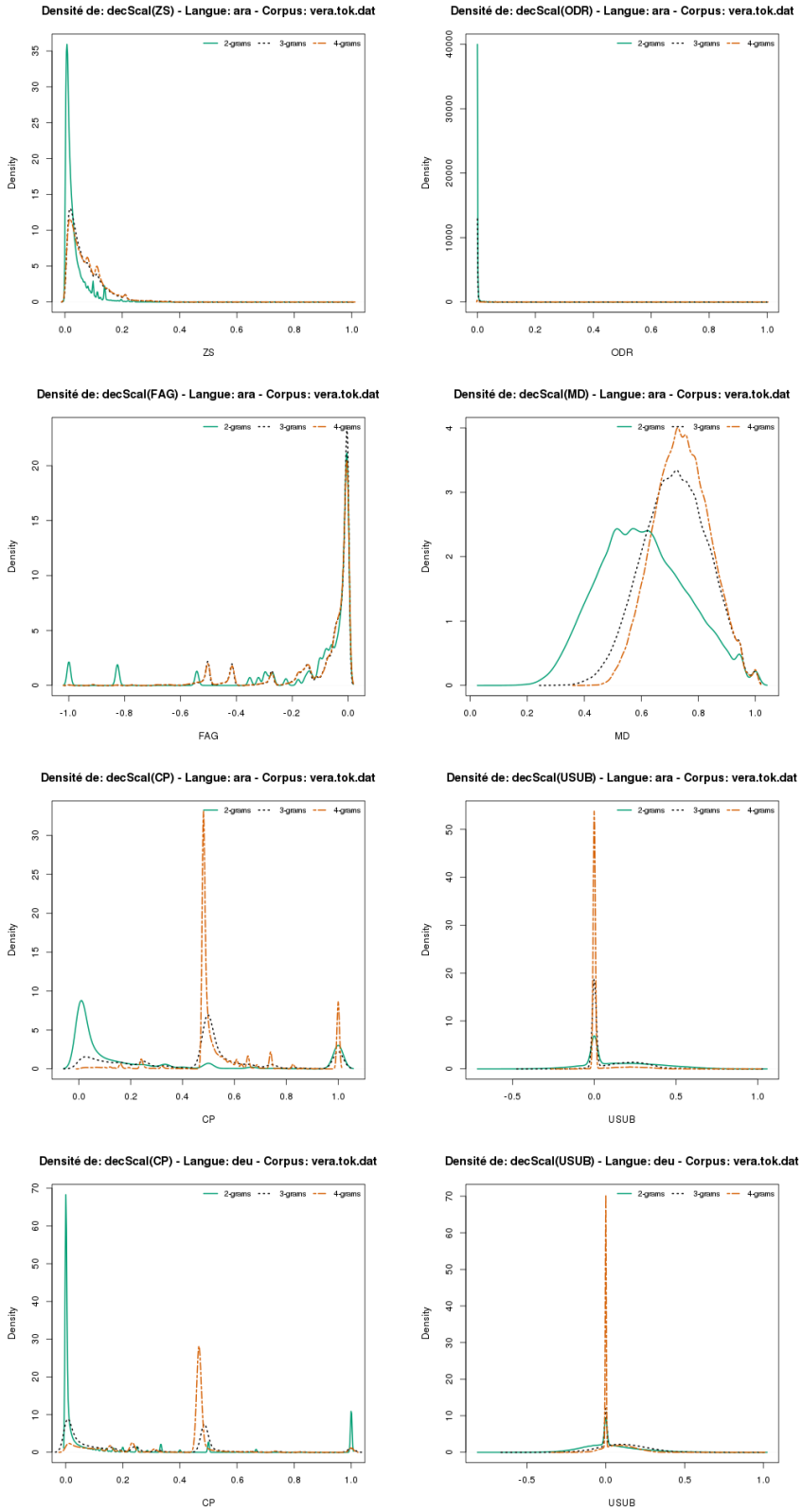


FIGURE A.5 – Points de découpages proposés par la normalisation EW pour les courbes de densités normalisées par mise à l'échelle décimale de ZS, ODR, FAG, MD, CP et USUB en arabe (corpus de spécialité).

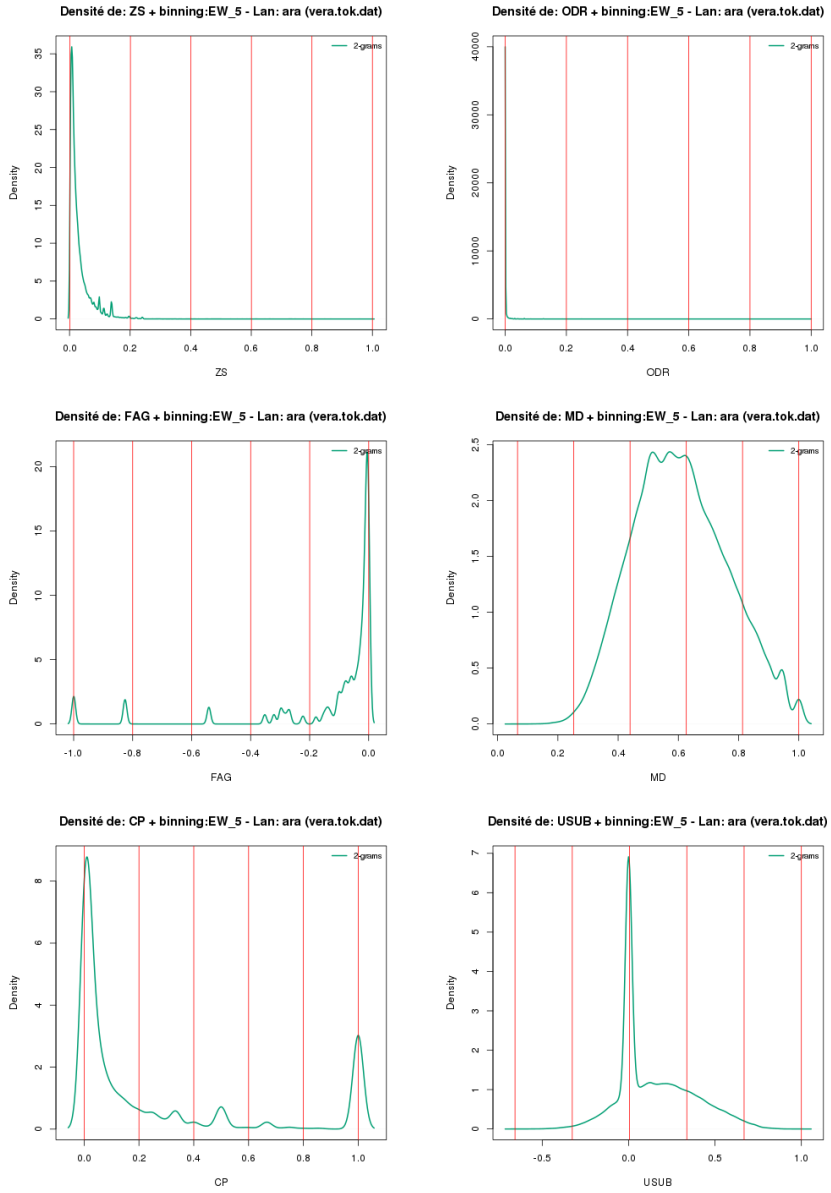


FIGURE A.6 – Points de découpages proposés par la normalisation EF pour les courbes de densités normalisées par mise à l'échelle décimale de ZS, ODR, FAG, MD, CP et USUB en arabe (corpus de spécialité).

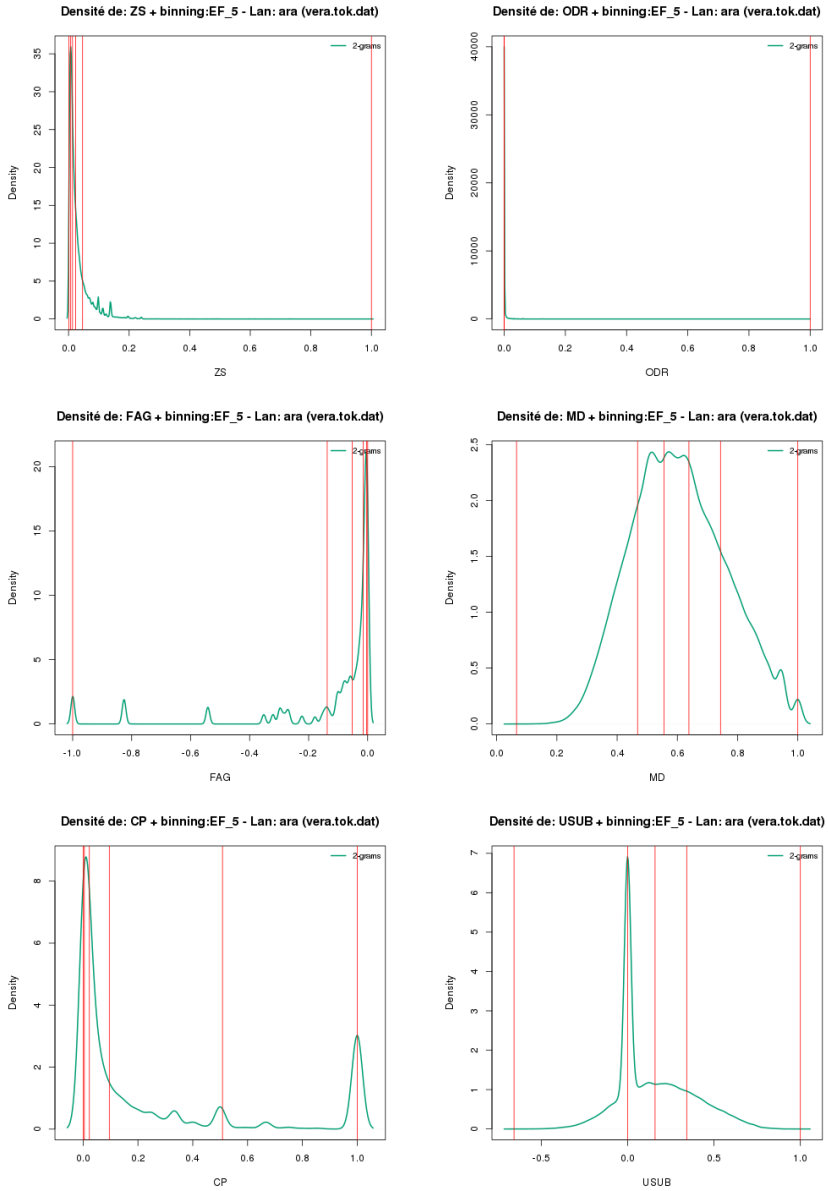
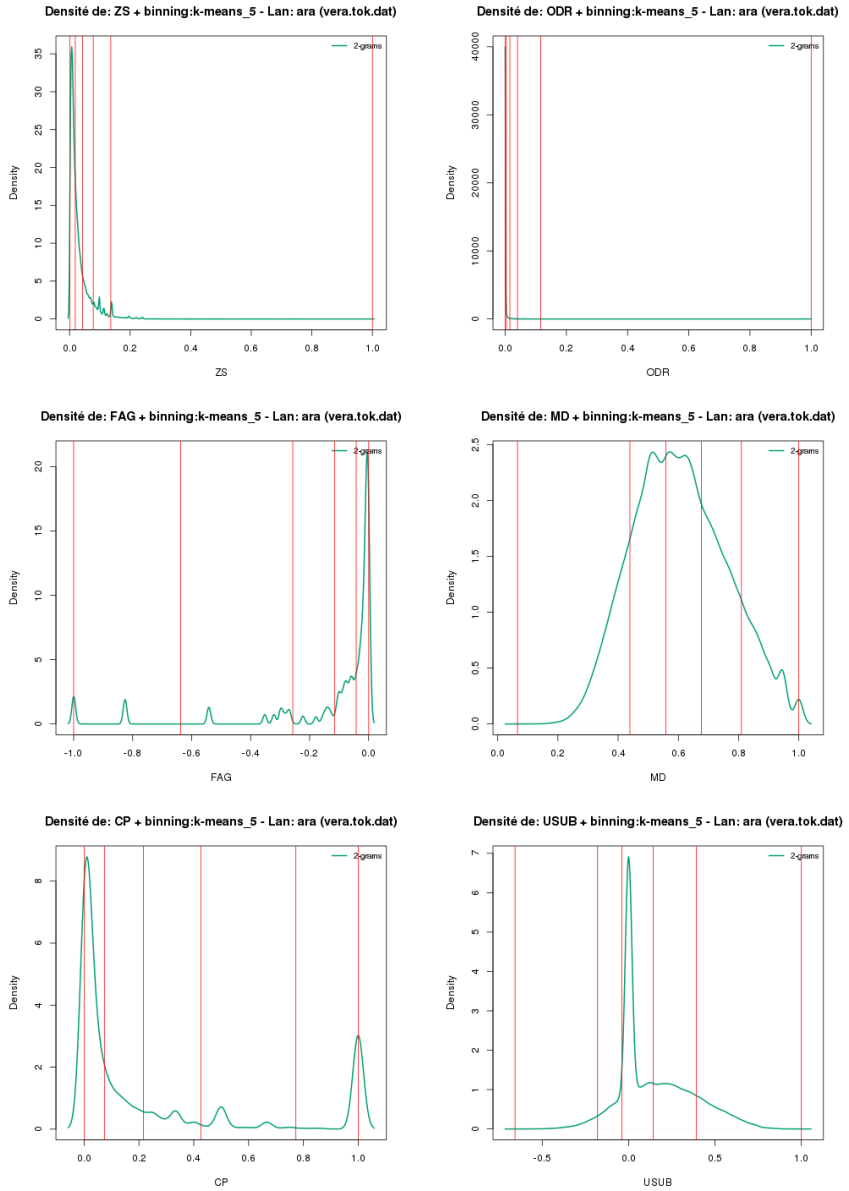


FIGURE A.7 – Points de découpages proposés par la normalisation k-means pour les courbes de densités normalisées par mise à l'échelle décimale de ZS, ODR, FAG, MD, CP et USUB en arabe (corpus de spécialité).



Annexe B

Scores de l'extraction terminologique

Dans les tableaux de résultats suivants, l'ensemble des traits ont été normalisés (voir la section 6.1.3.1) puis discrétisés (voir la section 6.1.3.2). Les codes utilisés pour les mesures d'association sont explicités dans le tableau 6.1.

Le trait « frequ » (fréquence), désignant la fréquence d'un n-gram donné, peut être préfixé de la lettre S (« Sfreq ») pour indiquer que le trait concerné est relatif aux fréquences du corpus de spécialité uniquement, ou de la lettre D (« D freq ») pour indiquer que les valeurs de ce trait sont la concaténation des valeurs normalisées et discrétisées des fréquences du corpus de spécialité et du corpus généraliste. Enfin, le trait « ttest » (à ne pas confondre avec la mesure d'association « TT ») est le résultat (normalisé et discrétisé) d'un t-test de Welch (1947) pour la comparaison des fréquences du corpus de spécialité et du corpus généraliste.

TABLE B.1 – Résultat de la première phase d'évaluation pour l'arabe (tokens informés).

Modèle	Précision	Rappel	F-score
Sfreq+SIM	0.955	0.947	0.951
Sfreq+FKZY	0.999	0.900	0.947
Sfreq+BB	0.999	0.899	0.947
CP+Sfreq	0.999	0.899	0.946
USUB+Sfreq	0.991	0.901	0.944
Sfreq+KLO	0.958	0.930	0.944
Sfreq+SAL	0.945	0.933	0.939
PRS+Sfreq	0.945	0.933	0.939
Sfreq+FAG	0.937	0.938	0.938
FSS+Sfreq	0.961	0.915	0.937
MTD+Sfreq	0.942	0.930	0.936
MD+Sfreq	0.930	0.940	0.935
JAC+Sfreq	0.932	0.934	0.933
ODR+Sfreq	0.925	0.940	0.933
TT+Sfreq	0.933	0.930	0.932
Sfreq+D freq	0.920	0.934	0.927
RCT+Sfreq	0.901	0.952	0.926
PMI+Sfreq	0.897	0.949	0.922
Sfreq+ttest	0.886	0.955	0.919
Sfreq	0.891	0.944	0.917
ZS+Sfreq	0.870	0.954	0.910
DRK+Sfreq	0.865	0.951	0.906
RCT+D freq	0.579	0.902	0.706
SAL+D freq	0.576	0.896	0.701
PMI+FAG	0.551	0.901	0.684
D freq+KLO	0.585	0.761	0.662
DRK+KLO	0.566	0.788	0.659
DRK+SAL	0.540	0.844	0.659
MD+D freq	0.575	0.769	0.658
ttest+SAL	0.543	0.808	0.649
PMI+SAL	0.516	0.860	0.645
DRK+FAG	0.548	0.779	0.644
PMI+KLO	0.588	0.707	0.642
RCT+ttest	0.565	0.702	0.626
CP+TT	0.536	0.749	0.625
PRS+TT	0.529	0.743	0.618
RCT+DRK	0.568	0.638	0.601
MD+PMI	0.555	0.558	0.557
ZS+SAL	0.527	0.464	0.494
RCT+PMI	0.573	0.425	0.488
PRS+CP	0.250	0.001	0.001
CP+BB	0.000	0.000	0.000
USUB+FKZY	∅	∅	∅
ODR+D freq	∅	∅	∅
FSS+CP	∅	∅	∅
ODR+FAG	∅	∅	∅
JAC+SIM	∅	∅	∅
ttest+SIM	∅	∅	∅
ZS+ttest	∅	∅	∅
USUB+FAG	∅	∅	∅
RCT+FAG	∅	∅	∅
DRK+D freq	∅	∅	∅
SIM	∅	∅	∅
USUB+MTD	∅	∅	∅
ZS+FAG	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
SAL+KLO	∅	∅	∅
PMI+BB	∅	∅	∅
ZS+JAC	∅	∅	∅
MTD+ttest	∅	∅	∅
FKZY+KLO	∅	∅	∅
JAC+KLO	∅	∅	∅
ODR	∅	∅	∅
MD+ODR	∅	∅	∅
ODR+TT	∅	∅	∅
FSS+KLO	∅	∅	∅
USUB+MD	∅	∅	∅
BB+D freq	∅	∅	∅
CP+RCT	∅	∅	∅
ZS+BB	∅	∅	∅
USUB+TT	∅	∅	∅
D freq	∅	∅	∅
PRS+PMI	∅	∅	∅
MTD+FKZY	∅	∅	∅
USUB+RCT	∅	∅	∅
MTD+KLO	∅	∅	∅
FSS+MTD	∅	∅	∅
BB+KLO	∅	∅	∅
FSS+FKZY	∅	∅	∅
ZS+ODR	∅	∅	∅
CP+FKZY	∅	∅	∅
DRK+ttest	∅	∅	∅
CP+SAL	∅	∅	∅
SIM+FAG	∅	∅	∅
FSS+RCT	∅	∅	∅
FKZY+FAG	∅	∅	∅
SIM+D freq	∅	∅	∅
CP+PMI	∅	∅	∅
ODR+RCT	∅	∅	∅
DRK	∅	∅	∅
FSS+D freq	∅	∅	∅
USUB+PMI	∅	∅	∅
ZS+ SIM	∅	∅	∅
FSS+SAL	∅	∅	∅
JAC+SAL	∅	∅	∅
RCT+FKZY	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
USUB+ODR	∅	∅	∅
ZS+FKZY	∅	∅	∅
TT+FAG	∅	∅	∅
PMI+JAC	∅	∅	∅
RCT+KLO	∅	∅	∅
ZS+D freq	∅	∅	∅
JAC+BB	∅	∅	∅
ZS+PMI	∅	∅	∅
USUB+SIM	∅	∅	∅
MD+MTD	∅	∅	∅
TT+JAC	∅	∅	∅
JAC+FAG	∅	∅	∅
FSS+PMI	∅	∅	∅
ttest+D freq	∅	∅	∅
RCT	∅	∅	∅
FSS+ZS	∅	∅	∅
ODR+KLO	∅	∅	∅
USUB	∅	∅	∅
CP+DRK	∅	∅	∅
USUB+SAL	∅	∅	∅
ODR+DRK	∅	∅	∅
MD+FAG	∅	∅	∅
ZS	∅	∅	∅
MD+TT	∅	∅	∅
MTD+SIM	∅	∅	∅
ttest+FAG	∅	∅	∅
MTD+PMI	∅	∅	∅
DRK+FKZY	∅	∅	∅
ZS+RCT	∅	∅	∅
FSS+JAC	∅	∅	∅
MD+DRK	∅	∅	∅
ttest+BB	∅	∅	∅
PRS+SAL	∅	∅	∅
FSS+TT	∅	∅	∅
TT+FKZY	∅	∅	∅
USUB+KLO	∅	∅	∅
BB+FAG	∅	∅	∅
BB+SAL	∅	∅	∅
PMI	∅	∅	∅
PRS+JAC	∅	∅	∅
SAL+FAG	∅	∅	∅
JAC+D freq	∅	∅	∅
PRS+SIM	∅	∅	∅
TT+SIM	∅	∅	∅
CP+ZS	∅	∅	∅
RCT+JAC	∅	∅	∅
FSS+MD	∅	∅	∅
TT+PMI	∅	∅	∅
CP	∅	∅	∅
PRS+DRK	∅	∅	∅
TT+ttest	∅	∅	∅
TT	∅	∅	∅
MD+JAC	∅	∅	∅
CP+JAC	∅	∅	∅
BB+SIM	∅	∅	∅
MD	∅	∅	∅
SAL+FKZY	∅	∅	∅
PRS+D freq	∅	∅	∅
ttest	∅	∅	∅
PRS+ttest	∅	∅	∅
ZS+DRK	∅	∅	∅
TT+KLO	∅	∅	∅
PMI+SIM	∅	∅	∅
JAC	∅	∅	∅
FSS+ODR	∅	∅	∅
CP+MD	∅	∅	∅
JAC+ttest	∅	∅	∅
SAL	∅	∅	∅
FSS+BB	∅	∅	∅
PRS+ZS	∅	∅	∅
SIM+SAL	∅	∅	∅
ODR+SIM	∅	∅	∅
CP+KLO	∅	∅	∅
SIM+KLO	∅	∅	∅
USUB+D freq	∅	∅	∅
RCT+SIM	∅	∅	∅
PMI+FKZY	∅	∅	∅
ODR+PMI	∅	∅	∅
FSS+DRK	∅	∅	∅
TT+D freq	∅	∅	∅
MTD	∅	∅	∅
DRK+SIM	∅	∅	∅
PRS+KLO	∅	∅	∅
FSS+SIM	∅	∅	∅
FSS	∅	∅	∅
PMI+ttest	∅	∅	∅
ZS+TT	∅	∅	∅
USUB+CP	∅	∅	∅
FSS+USUB	∅	∅	∅
DRK+BB	∅	∅	∅
ZS+KLO	∅	∅	∅
PRS+FSS	∅	∅	∅
BB	∅	∅	∅
USUB+JAC	∅	∅	∅
D freq+FAG	∅	∅	∅
PRS+ODR	∅	∅	∅
TT+BB	∅	∅	∅
CP+D freq	∅	∅	∅
PRS+FAG	∅	∅	∅
SIM+FKZY	∅	∅	∅
CP+ttest	∅	∅	∅
USUB+DRK	∅	∅	∅
PRS+MD	∅	∅	∅
BB+FKZY	∅	∅	∅
CP+MTD	∅	∅	∅
PRS+FKZY	∅	∅	∅
FSS+ttest	∅	∅	∅
ttest+KLO	∅	∅	∅
MTD+D freq	∅	∅	∅
MD+SAL	∅	∅	∅
D freq+FKZY	∅	∅	∅
PRS+MTD	∅	∅	∅
TT+SAL	∅	∅	∅
ODR+MTD	∅	∅	∅
MD+ZS	∅	∅	∅
ODR+ttest	∅	∅	∅
PMI+D freq	∅	∅	∅
RCT+MTD	∅	∅	∅
TT+DRK	∅	∅	∅
MD+RCT	∅	∅	∅
PRS+USUB	∅	∅	∅
RCT+TT	∅	∅	∅
CP+ODR	∅	∅	∅
MD+KLO	∅	∅	∅
CP+SIM	∅	∅	∅
PRS+RCT	∅	∅	∅
ODR+JAC	∅	∅	∅
FAG	∅	∅	∅
ZS+MTD	∅	∅	∅
USUB+ttest	∅	∅	∅
MD+ttest	∅	∅	∅
MTD+SAL	∅	∅	∅
RCT+BB	∅	∅	∅
MTD+DRK	∅	∅	∅
ODR+FKZY	∅	∅	∅
MTD+JAC	∅	∅	∅
FSS+FAG	∅	∅	∅
ODR+BB	∅	∅	∅
MD+BB	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
JAC+FKZY	∅	∅	∅
RCT+SAL	∅	∅	∅
MD+FKZY	∅	∅	∅
JAC+DRK	∅	∅	∅
MTD+BB	∅	∅	∅
ttest+FKZY	∅	∅	∅
ODR+SAL	∅	∅	∅
MD+SIM	∅	∅	∅
FKZY	∅	∅	∅
USUB+ZS	∅	∅	∅
TT+MTD	∅	∅	∅

TABLE B.2 – Résultat de la première phase d'évaluation pour l'allemand (tokens informés).

Modèle	Précision	Rappel	F-score
Sfreq+SIM	0.861	0.990	0.921
CP+Sfreq	0.718	0.987	0.831
Sfreq+ttest	0.711	0.991	0.828
Sfreq+D freq	0.690	0.992	0.814
USUB+Sfreq	0.597	0.987	0.744
ODR+Sfreq	0.596	0.986	0.743
JAC+Sfreq	0.591	0.986	0.739
TT+Sfreq	0.585	0.988	0.735
PRS+Sfreq	0.585	0.986	0.735
DRK+Sfreq	0.584	0.987	0.734
RCT+Sfreq	0.582	0.987	0.732
Sfreq+SAL	0.578	0.988	0.729
Sfreq+FAG	0.578	0.987	0.729
Sfreq+BB	0.576	0.989	0.728
FSS+Sfreq	0.575	0.989	0.727
PMI+Sfreq	0.573	0.987	0.725
Sfreq+KLO	0.569	0.988	0.722
ZS+Sfreq	0.568	0.987	0.721
Sfreq+FKZY	0.564	0.989	0.719
MTD+Sfreq	0.558	0.988	0.713
MD+Sfreq	0.556	0.988	0.712
Sfreq	0.555	0.988	0.711
PMI+D freq	0.501	0.860	0.633
D freq+KLO	0.495	0.870	0.631
D freq+FAG	0.484	0.898	0.629
MD+D freq	0.480	0.904	0.627
SAL+D freq	0.469	0.939	0.626
USUB+D freq	0.511	0.794	0.622
RCT+D freq	0.454	0.957	0.615
D freq	0.453	0.957	0.615
D freq+FKZY	0.452	0.954	0.613
MTD+D freq	0.454	0.946	0.613
CP+D freq	0.446	0.957	0.609
ttest+D freq	0.445	0.961	0.608
PRS+D freq	0.444	0.960	0.607
BB+D freq	0.443	0.960	0.606
JAC+D freq	0.442	0.961	0.606
ODR+ttest	0.444	0.894	0.594
TT+ttest	0.418	0.967	0.584
ttest+SIM	0.409	0.995	0.580
TT+PMI	0.456	0.796	0.580
PRS+ZS	0.434	0.840	0.572
TT+FKZY	0.436	0.831	0.572
ZS+JAC	0.442	0.803	0.571
TT+BB	0.434	0.832	0.570
PRS+DRK	0.434	0.830	0.570
PRS+MD	0.434	0.827	0.570
ZS+BB	0.434	0.825	0.569
RCT+PMI	0.439	0.807	0.568
JAC+DRK	0.430	0.837	0.568
DRK+ttest	0.400	0.974	0.567
PRS+KLO	0.431	0.830	0.567
ZS+FKZY	0.431	0.829	0.567
PMI+JAC	0.432	0.822	0.566
PMI+ttest	0.394	0.986	0.563
PRS+FAG	0.434	0.800	0.563
MTD+JAC	0.426	0.825	0.562
MD+JAC	0.426	0.824	0.562
FSS+USUB	0.426	0.819	0.561
JAC+KLO	0.431	0.801	0.560
USUB+ODR	0.423	0.828	0.560
ODR+FKZY	0.434	0.788	0.559
FSS+FKZY	0.426	0.812	0.559
MTD+DRK	0.426	0.813	0.559
PMI+KLO	0.402	0.915	0.558
RCT+ttest	0.387	0.993	0.557
FSS+ttest	0.385	0.993	0.555
CP+ttest	0.385	0.993	0.555
ZS+ttest	0.385	0.990	0.555
MTD+ttest	0.385	0.992	0.555
USUB+ttest	0.383	0.992	0.553
ttest+FAG	0.383	0.992	0.553
RCT+JAC	0.427	0.782	0.552
ttest	0.383	0.990	0.552
ODR+BB	0.435	0.755	0.552
ttest+KLO	0.383	0.990	0.552
JAC+FAG	0.435	0.753	0.551
CP+MD	0.395	0.914	0.551
PMI	0.447	0.719	0.551
ttest+FKZY	0.381	0.992	0.551
MD+ttest	0.381	0.994	0.551
ttest+BB	0.381	0.990	0.550
JAC+ttest	0.380	0.992	0.550
ttest+SAL	0.380	0.992	0.549
CP+MTD	0.394	0.901	0.548
PRS+ttest	0.379	0.991	0.548
ODR	0.385	0.912	0.541
CP+FAG	0.376	0.945	0.538
PRS+RCT	0.374	0.957	0.538
CP+TT	0.372	0.966	0.537
FSS+SIM	0.370	0.968	0.536
DRK+FAG	0.373	0.947	0.535
CP+RCT	0.366	0.973	0.531
PRS+PMI	0.446	0.650	0.529
BB+KLO	0.363	0.969	0.528
MD+DRK	0.361	0.969	0.527
USUB+ZS	0.352	0.978	0.518
CP+ODR	0.350	0.979	0.516
PRS+MTD	0.350	0.978	0.516
RCT+DRK	0.347	0.984	0.513
USUB+TT	0.340	0.985	0.506
DRK+BB	0.339	0.984	0.504
ODR+DRK	0.338	0.983	0.503
CP+KLO	0.338	0.988	0.503
RCT+TT	0.336	0.990	0.501
TT+FAG	0.331	0.989	0.496
PMI+BB	0.330	0.991	0.495
MD+RCT	0.330	0.996	0.495
MD+TT	0.330	0.991	0.495
TT+SIM	0.329	0.994	0.494
TT+DRK	0.328	0.992	0.493
CP+JAC	0.328	0.993	0.493
ZS+TT	0.328	0.993	0.493
RCT+FAG	0.327	0.994	0.492
FSS+TT	0.326	0.994	0.491
TT+MTD	0.326	0.992	0.491
TT+KLO	0.326	0.993	0.491
RCT+MTD	0.326	0.994	0.491
ODR+TT	0.326	0.994	0.491
ZS+PMI	0.326	0.993	0.491
ZS+RCT	0.325	0.996	0.490
MTD+KLO	0.325	0.994	0.490
DRK+KLO	0.325	0.995	0.490
FSS+PMI	0.325	0.993	0.489
MD+FAG	0.324	0.993	0.489
MTD+FAG	0.324	0.993	0.489
ZS+MTD	0.324	0.996	0.489
RCT+SIM	0.324	0.995	0.489
PMI+SIM	0.323	0.995	0.488
MD+MTD	0.324	0.993	0.488
ZS+SIM	0.323	0.995	0.488
SIM+KLO	0.323	0.995	0.488
RCT+KLO	0.323	0.996	0.488
JAC+SAL	0.323	0.994	0.487
ODR+RCT	0.323	0.993	0.487
USUB+PMI	0.323	0.992	0.487
FKZY+KLO	0.323	0.994	0.487
ZS+KLO	0.323	0.995	0.487
SIM+SAL	0.322	0.995	0.487
SIM+FAG	0.322	0.995	0.487
ODR+SIM	0.322	0.996	0.487
PMI+SAL	0.322	0.996	0.487
ZS+DRK	0.322	0.995	0.487
CP+ZS	0.322	0.997	0.487
TT+SAL	0.322	0.994	0.487
ODR+MTD	0.322	0.995	0.486
MD+ZS	0.322	0.997	0.486
BB+FAG	0.322	0.993	0.486
PMI+FKZY	0.322	0.993	0.486
ZS+FAG	0.322	0.995	0.486
MD+BB	0.322	0.995	0.486
ZS+SAL	0.321	0.995	0.486
MD+SIM	0.321	0.995	0.486
DRK+SIM	0.321	0.996	0.486
MD+KLO	0.321	0.996	0.486
CP+SIM	0.321	0.996	0.486
RCT+SAL	0.321	0.996	0.485
ODR+FAG	0.321	0.993	0.485
SIM	0.321	0.994	0.485
DRK+FKZY	0.321	0.995	0.485
DRK+SAL	0.320	0.996	0.485
FAG	0.320	0.995	0.485
RCT	0.320	0.995	0.485
FSS+CP	0.320	0.995	0.484
MTD+SIM	0.320	0.996	0.484
TT	0.320	0.993	0.484
ZS	0.320	0.994	0.484
FSS+ZS	0.319	0.996	0.484
RCT+BB	0.319	0.996	0.484
ZS+ODR	0.319	0.996	0.483
FSS+RCT	0.319	0.995	0.483
MD+SAL	0.319	0.996	0.483
SAL+FAG	0.319	0.995	0.483
MTD	0.319	0.994	0.483
ODR+KLO	0.319	0.995	0.483
USUB+RCT	0.319	0.996	0.483
USUB+KLO	0.319	0.994	0.483
CP+SAL	0.318	0.996	0.482
ODR+SAL	0.318	0.995	0.482
KLO	0.318	0.995	0.482
SAL+KLO	0.318	0.995	0.482
FSS+KLO	0.318	0.995	0.482
CP	0.318	0.995	0.482
RCT+FKZY	0.318	0.995	0.482
MTD+SAL	0.318	0.995	0.482
FSS+FAG	0.318	0.996	0.482
FSS+SAL	0.318	0.996	0.482
MD+ODR	0.318	0.996	0.482
CP+BB	0.318	0.995	0.482
FSS+MD	0.318	0.995	0.481
BB+SAL	0.318	0.995	0.481
MD	0.317	0.996	0.481
JAC	0.318	0.993	0.481
PRS+SAL	0.317	0.995	0.481
FSS+MTD	0.317	0.995	0.481
SAL	0.317	0.995	0.481
USUB+FAG	0.317	0.995	0.481
FSS+DRK	0.317	0.997	0.481
USUB+SAL	0.317	0.995	0.481
PRS	0.317	0.994	0.481
PRS+CP	0.317	0.996	0.481
SAL+FKZY	0.317	0.995	0.481
USUB+MD	0.317	0.996	0.481
FSS+ODR	0.317	0.996	0.480
FSS	0.316	0.995	0.480
MTD+FKZY	0.316	0.995	0.480
FKZY+FAG	0.316	0.995	0.480
MD+FKZY	0.316	0.996	0.480
CP+FKZY	0.315	0.996	0.479
FKZY	0.315	0.995	0.479
BB	0.315	0.995	0.479
USUB+MTD	0.315	0.995	0.479
MTD+BB	0.315	0.994	0.478
USUB	0.315	0.994	0.478
USUB+DRK	0.314	0.996	0.478
USUB+CP	0.314	0.996	0.478
ZS+D freq	0.678	0.030	0.057
DRK	0.458	0.016	0.030
SIM+D freq	0.595	0.006	0.012
FSS+D freq	0.362	0.001	0.001
DRK+D freq	1.000	0.000	0.000
TT+D freq	1.000	0.000	0.000
USUB+FKZY	∅	∅	∅
ODR+D freq	∅	∅	∅
JAC+SIM	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
CP+PMI	∅	∅	∅
JAC+BB	∅	∅	∅
USUB+SIM	∅	∅	∅
TT+JAC	∅	∅	∅
CP+DRK	∅	∅	∅
MTD+PMI	∅	∅	∅
FSS+JAC	∅	∅	∅
PMI+FAG	∅	∅	∅
PRS+JAC	∅	∅	∅
PRS+SIM	∅	∅	∅
BB+SIM	∅	∅	∅
FSS+BB	∅	∅	∅
ODR+PMI	∅	∅	∅
PRS+FSS	∅	∅	∅
USUB+JAC	∅	∅	∅
PRS+ODR	∅	∅	∅
SIM+FKZY	∅	∅	∅
BB+FKZY	∅	∅	∅
PRS+FKZY	∅	∅	∅
PRS+USUB	∅	∅	∅
ODR+JAC	∅	∅	∅
MD+PMI	∅	∅	∅
JAC+FKZY	∅	∅	∅
PRS+TT	∅	∅	∅

TABLE B.3 – Résultat de la première phase d'évaluation pour l'anglais (tokens informés).

Modèle	Précision	Rappel	F-score
Sfreq+SIM	0.863	0.929	0.895
Sfreq+D freq	0.831	0.900	0.864
CP+Sfreq	0.794	0.940	0.861
TT+Sfreq	0.802	0.926	0.860
Sfreq+ttest	0.787	0.942	0.857
MD+Sfreq	0.800	0.921	0.856
ODR+Sfreq	0.792	0.923	0.852
MTD+Sfreq	0.785	0.929	0.851
Sfreq+FAG	0.778	0.921	0.844
Sfreq+SAL	0.777	0.920	0.843
PMI+Sfreq	0.768	0.932	0.842
DRK+Sfreq	0.775	0.919	0.841
ZS+Sfreq	0.755	0.947	0.840
RCT+Sfreq	0.756	0.936	0.836
Sfreq+KLO	0.765	0.921	0.836
FSS+Sfreq	0.754	0.932	0.834
Sfreq+FKZY	0.743	0.934	0.828
JAC+Sfreq	0.752	0.918	0.827
Sfreq	0.761	0.903	0.826
Sfreq+BB	0.739	0.930	0.824
PRS+Sfreq	0.727	0.942	0.821
USUB+Sfreq	0.727	0.938	0.819
SIM+D freq	0.571	0.904	0.700
D freq+FKZY	0.560	0.921	0.697
ttest+D freq	0.552	0.941	0.695
D freq+FAG	0.554	0.930	0.694
MD+D freq	0.554	0.929	0.694
PRS+D freq	0.553	0.926	0.693
MTD+D freq	0.555	0.918	0.692
ZS+D freq	0.550	0.931	0.691
RCT+D freq	0.545	0.939	0.690
SAL+D freq	0.545	0.919	0.685
ODR+D freq	0.543	0.923	0.684
D freq+KLO	0.539	0.932	0.683
D freq	0.542	0.914	0.680
PMI+D freq	0.539	0.920	0.680
TT+D freq	0.538	0.923	0.680
DRK+D freq	0.525	0.948	0.676
BB+D freq	0.532	0.925	0.676
FSS+D freq	0.533	0.915	0.673
CP+D freq	0.524	0.940	0.673
USUB+D freq	0.532	0.911	0.671
PMI+ttest	0.541	0.884	0.671
JAC+D freq	0.530	0.913	0.671
DRK+ttest	0.514	0.908	0.657
PRS+BB	0.552	0.805	0.655
ZS+ttest	0.492	0.975	0.654
MTD+ttest	0.495	0.961	0.654
ODR+ttest	0.493	0.966	0.653
TT+ttest	0.489	0.963	0.649
ttest+SIM	0.488	0.965	0.648
RCT+ttest	0.483	0.977	0.646
PRS+ttest	0.484	0.969	0.646
MD+ttest	0.480	0.971	0.642
ttest+FAG	0.478	0.973	0.641
MD+PMI	0.483	0.951	0.641
FSS+ttest	0.478	0.971	0.640
ttest+FKZY	0.473	0.964	0.634
ttest+KLO	0.469	0.967	0.632
ttest+SAL	0.469	0.968	0.632
ttest	0.468	0.967	0.631
CP+ttest	0.466	0.967	0.629
ttest+BB	0.463	0.970	0.627
USUB+ttest	0.464	0.964	0.627
JAC+ttest	0.463	0.966	0.626
ZS	0.461	0.960	0.623
RCT	0.457	0.971	0.621
TT+FAG	0.451	0.974	0.617
MD+SAL	0.449	0.980	0.615
PRS+FAG	0.444	0.974	0.610
JAC+SAL	0.443	0.976	0.610
MD+TT	0.444	0.972	0.610
SAL+KLO	0.442	0.975	0.608
MTD+FAG	0.442	0.972	0.607
FSS+SAL	0.440	0.974	0.607
ZS+PMI	0.441	0.966	0.606
SAL+FAG	0.439	0.976	0.606
FKZY+FAG	0.438	0.976	0.605
FSS+DRK	0.438	0.977	0.604
MD+SIM	0.437	0.974	0.603
SAL	0.438	0.972	0.603
DRK+SAL	0.437	0.975	0.603
PRS+SAL	0.437	0.971	0.603
PMI+FAG	0.439	0.960	0.603
MTD+SAL	0.437	0.973	0.603
TT+SIM	0.439	0.963	0.603
DRK+FAG	0.437	0.973	0.603
PRS+ZS	0.435	0.980	0.602
CP+TT	0.437	0.968	0.602
MTD+SIM	0.436	0.972	0.602
MTD+DRK	0.436	0.971	0.602
FSS+FAG	0.436	0.973	0.602
TT+KLO	0.437	0.963	0.602
RCT+MTD	0.435	0.977	0.602
ODR+MTD	0.435	0.974	0.601
TT+SAL	0.436	0.968	0.601
ODR+SAL	0.435	0.975	0.601
ZS+MTD	0.433	0.979	0.601
TT+MTD	0.436	0.968	0.601
CP+FAG	0.434	0.973	0.601
MD+MTD	0.435	0.970	0.600
PRS+PMI	0.433	0.980	0.600
SAL+FKZY	0.434	0.973	0.600
ODR+FAG	0.434	0.973	0.600
RCT+DRK	0.432	0.980	0.600
SIM+SAL	0.433	0.972	0.599
ODR+DRK	0.432	0.979	0.599
PMI+SAL	0.432	0.976	0.599
PRS+TT	0.433	0.971	0.599
FSS+TT	0.433	0.969	0.599
ZS+FAG	0.431	0.979	0.599
MD+JAC	0.432	0.972	0.598
JAC+FAG	0.432	0.973	0.598
ZS+RCT	0.430	0.980	0.598
MD+DRK	0.432	0.973	0.598
MD+ODR	0.432	0.973	0.598
ZS+TT	0.431	0.972	0.598
CP+RCT	0.430	0.979	0.597
FSS+MTD	0.431	0.974	0.597
SIM+KLO	0.431	0.971	0.597
ZS+FKZY	0.429	0.980	0.597
RCT+TT	0.431	0.968	0.596
FSS+CP	0.429	0.976	0.596
SIM+FAG	0.430	0.971	0.596
ODR+KLO	0.429	0.976	0.596
RCT+SIM	0.429	0.977	0.596
ODR+PMI	0.430	0.968	0.596
FSS+ZS	0.429	0.978	0.596
TT+DRK	0.432	0.960	0.596
USUB+SAL	0.429	0.973	0.596
FSS+MD	0.429	0.972	0.595
USUB+RCT	0.427	0.981	0.595
ZS+SIM	0.428	0.975	0.595
ZS+DRK	0.427	0.978	0.595
RCT+PMI	0.428	0.972	0.595
RCT+KLO	0.428	0.974	0.595
CP+SIM	0.427	0.975	0.594
ZS+SAL	0.427	0.975	0.594
MD+KLO	0.429	0.966	0.594
TT+PMI	0.429	0.965	0.594
ODR+RCT	0.426	0.978	0.594
PRS+CP	0.426	0.979	0.594
PMI+SIM	0.426	0.976	0.593
BB+SIM	0.427	0.972	0.593
MTD+KLO	0.428	0.966	0.593
CP+ODR	0.426	0.976	0.593
FKZY+KLO	0.427	0.971	0.593
ZS+ODR	0.426	0.976	0.593
CP+FKZY	0.426	0.974	0.593
MD+FKZY	0.427	0.970	0.593
RCT+SAL	0.427	0.972	0.593
PRS+RCT	0.426	0.975	0.593
ZS+KLO	0.425	0.979	0.593
MD+FAG	0.427	0.971	0.593
FSS+RCT	0.426	0.973	0.592
ODR+TT	0.427	0.969	0.592
TT+FKZY	0.427	0.967	0.592
FSS+ODR	0.425	0.974	0.592
CP+MTD	0.426	0.970	0.592
RCT+FAG	0.425	0.971	0.592
DRK+KLO	0.425	0.972	0.592
PRS+MD	0.426	0.967	0.591
SIM+FKZY	0.426	0.969	0.591
MTD+JAC	0.425	0.970	0.591
PRS+KLO	0.425	0.969	0.591
USUB+FAG	0.425	0.973	0.591
PRS+SIM	0.426	0.967	0.591
BB+FAG	0.425	0.969	0.591
CP+KLO	0.425	0.968	0.591
MD+BB	0.425	0.968	0.591
CP+ZS	0.423	0.977	0.590
RCT+FKZY	0.423	0.977	0.590
MD+RCT	0.423	0.975	0.590
CP+MD	0.424	0.971	0.590
DRK+FKZY	0.423	0.972	0.590
FSS+PMI	0.423	0.974	0.590
PMI+FKZY	0.422	0.976	0.589
ZS+JAC	0.422	0.975	0.589
USUB+ODR	0.422	0.975	0.589
CP+SAL	0.423	0.966	0.589
JAC+SIM	0.423	0.968	0.588
USUB+PMI	0.421	0.978	0.588
TT+BB	0.424	0.961	0.588
RCT+JAC	0.421	0.973	0.588
MD+ZS	0.421	0.974	0.588
MTD+BB	0.422	0.967	0.588
ODR+SIM	0.422	0.970	0.588
ZS+BB	0.420	0.977	0.588
RCT+BB	0.421	0.972	0.588
FSS+SIM	0.421	0.969	0.587
USUB+ZS	0.420	0.977	0.587
USUB+MD	0.421	0.970	0.587
PRS+DRK	0.421	0.971	0.587
MTD+FKZY	0.421	0.968	0.587
BB+FKZY	0.420	0.971	0.586
DRK+SIM	0.421	0.966	0.586
PRS+ODR	0.419	0.972	0.586
PRS+FKZY	0.419	0.970	0.586
PRS+MTD	0.421	0.963	0.586
FSS+FKZY	0.419	0.970	0.585
ODR+BB	0.419	0.972	0.585
USUB+SIM	0.419	0.970	0.585
USUB+TT	0.421	0.958	0.585
CP+PMI	0.418	0.973	0.585
USUB+FKZY	0.417	0.972	0.584
PMI+KLO	0.418	0.966	0.584
BB+SAL	0.419	0.962	0.584
ODR+FKZY	0.418	0.969	0.584
ODR+JAC	0.417	0.970	0.583
USUB+CP	0.416	0.976	0.583
JAC+FKZY	0.417	0.969	0.583
FSS+KLO	0.417	0.966	0.583
PRS+FSS	0.417	0.964	0.582
USUB+MTD	0.417	0.967	0.582
TT+JAC	0.418	0.960	0.582
USUB+DRK	0.415	0.970	0.582
JAC+KLO	0.416	0.966	0.581
PMI+BB	0.414	0.972	0.581
CP+DRK	0.414	0.971	0.580
PMI+JAC	0.413	0.972	0.580
FSS+BB	0.414	0.965	0.580
FSS+USUB	0.414	0.966	0.580
FSS+JAC	0.415	0.962	0.580
USUB+KLO	0.414	0.964	0.579
CP+JAC	0.412	0.965	0.578
BB+KLO	0.413	0.958	0.577
CP+BB	0.411	0.962	0.576
USUB+JAC	0.410	0.961	0.575
USUB+BB	0.408	0.962	0.573
PRS+USUB	0.407	0.962	0.572
SIM	0.594	0.007	0.014
TT	0.740	0.001	0.002
CP	1.000	0.000	0.001
PMI+DRK	∅	∅	∅
ODR	∅	∅	∅
DRK	∅	∅	∅
JAC+BB	∅	∅	∅
USUB	∅	∅	∅
MTD+PMI	∅	∅	∅
PMI	∅	∅	∅
PRS+JAC	∅	∅	∅
MD	∅	∅	∅
JAC	∅	∅	∅
MTD	∅	∅	∅
FSS	∅	∅	∅
DRK+BB	∅	∅	∅
BB	∅	∅	∅
FAG	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
JAC+DRK	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.4 – Résultat de la première phase d'évaluation pour le français (tokens informés).

Modèle	Précision	Rappel	F-score
Sfreq+KLO	0.906	0.931	0.918
CP+Sfreq	0.885	0.953	0.918
MD+Sfreq	0.891	0.944	0.917
Sfreq+FAG	0.892	0.941	0.916
Sfreq+ttest	0.883	0.950	0.915
MTD+Sfreq	0.881	0.949	0.914
ODR+Sfreq	0.885	0.943	0.913
PMI+Sfreq	0.885	0.940	0.912
Sfreq+D freq	0.881	0.941	0.910
ZS+Sfreq	0.879	0.942	0.910
PRS+Sfreq	0.872	0.950	0.910
Sfreq+SAL	0.871	0.945	0.907
FSS+Sfreq	0.877	0.938	0.906
DRK+Sfreq	0.871	0.944	0.906
Sfreq+BB	0.872	0.941	0.905
JAC+Sfreq	0.859	0.951	0.903
USUB+Sfreq	0.860	0.945	0.901
RCT+Sfreq	0.856	0.950	0.900
TT+Sfreq	0.845	0.956	0.897
Sfreq	0.856	0.924	0.889
MD+D freq	0.620	0.970	0.756
SIM+D freq	0.615	0.968	0.752
D freq+FAG	0.616	0.966	0.752
FSS+D freq	0.617	0.963	0.752
ttest+D freq	0.611	0.969	0.749
D freq+FKZY	0.611	0.968	0.749
SAL+D freq	0.613	0.963	0.749
FSS+ttest	0.618	0.949	0.749
MTD+D freq	0.608	0.969	0.747
USUB+D freq	0.608	0.969	0.747
CP+D freq	0.607	0.968	0.746
BB+D freq	0.610	0.957	0.746
ODR+D freq	0.606	0.965	0.744
TT+D freq	0.604	0.967	0.743
ZS+D freq	0.606	0.961	0.743
D freq+KLO	0.604	0.965	0.743
ttest+FKZY	0.625	0.915	0.743
ttest+SIM	0.602	0.969	0.742
RCT+D freq	0.603	0.960	0.740
D freq	0.599	0.960	0.738
USUB+PMI	0.618	0.908	0.735
MTD+ttest	0.586	0.982	0.734
MTD+SAL	0.595	0.950	0.732
TT+ttest	0.584	0.981	0.732
ODR+MTD	0.594	0.943	0.729
USUB+ttest	0.581	0.969	0.727
RCT+TT	0.594	0.934	0.726
RCT+ttest	0.576	0.982	0.726
ttest+FAG	0.576	0.979	0.725
PMI+SIM	0.643	0.829	0.724
FSS+MD	0.587	0.945	0.724
ODR+DRK	0.644	0.824	0.723
PRS+ttest	0.574	0.974	0.722
TT+KLO	0.586	0.940	0.722
PMI+FKZY	0.604	0.896	0.722
MD+ttest	0.569	0.983	0.721
CP+TT	0.629	0.842	0.720
ttest+SAL	0.566	0.981	0.718
TT	0.590	0.915	0.717
FSS+KLO	0.583	0.931	0.717
CP+ttest	0.571	0.963	0.717
SAL+KLO	0.577	0.943	0.716
TT+FKZY	0.568	0.968	0.716
ZS+ttest	0.563	0.982	0.715
DRK+SIM	0.591	0.902	0.714
FSS+RCT	0.565	0.968	0.714
PMI+SAL	0.567	0.961	0.713
DRK+ttest	0.559	0.982	0.713
MTD+FKZY	0.558	0.986	0.713
ODR+ttest	0.560	0.980	0.713
ttest+KLO	0.560	0.978	0.713
DRK+FKZY	0.574	0.939	0.712
ODR+PMI	0.636	0.809	0.712
CP+ZS	0.571	0.941	0.711
PMI+ttest	0.556	0.983	0.710

Modèle	Précision	Rappel	F-score
KLO	0.579	0.908	0.707
MD+SIM	0.555	0.968	0.706
MTD+SIM	0.554	0.971	0.705
RCT+SAL	0.553	0.971	0.705
CP+SAL	0.554	0.966	0.704
ttest+BB	0.551	0.976	0.704
MD+SAL	0.548	0.983	0.704
PMI	0.609	0.831	0.703
FSS+CP	0.627	0.799	0.703
TT+SIM	0.554	0.959	0.702
JAC+ttest	0.548	0.975	0.701
DRK+SAL	0.546	0.974	0.700
MD	0.547	0.970	0.700
SAL	0.542	0.983	0.699
SAL+FKZY	0.541	0.985	0.698
PRS+SAL	0.539	0.989	0.698
FSS+SAL	0.542	0.978	0.697
USUB+DRK	0.545	0.962	0.696
ODR+SAL	0.540	0.974	0.695
FKZY+FAG	0.538	0.981	0.695
ODR+FKZY	0.539	0.974	0.694
DRK	0.544	0.952	0.692
TT+SAL	0.531	0.985	0.690
ZS+SIM	0.532	0.979	0.689
ZS+FKZY	0.530	0.983	0.689
MD+TT	0.530	0.982	0.689
SIM+SAL	0.530	0.983	0.689
ZS+MTD	0.529	0.984	0.688
FSS+TT	0.530	0.979	0.688
USUB+CP	0.529	0.979	0.687
MD+ODR	0.528	0.980	0.687
PRS+TT	0.526	0.985	0.686
SIM+KLO	0.529	0.977	0.686
DRK+BB	0.530	0.968	0.685
TT+FAG	0.526	0.978	0.684
ZS+KLO	0.524	0.980	0.683
FSS+ZS	0.524	0.980	0.683
TT+BB	0.523	0.979	0.682
RCT+FKZY	0.520	0.984	0.681
TT+PMI	0.520	0.981	0.680
PRS+FAG	0.519	0.986	0.680
TT+DRK	0.521	0.978	0.680
PRS+ZS	0.518	0.986	0.679
PRS+FSS	0.518	0.983	0.678
PRS+MTD	0.518	0.985	0.678
PRS+MD	0.517	0.986	0.678
ZS+RCT	0.518	0.982	0.678
JAC+BB	0.521	0.971	0.678
TT+JAC	0.517	0.984	0.678
ODR+TT	0.518	0.981	0.678
PRS+RCT	0.517	0.983	0.677
RCT	0.516	0.983	0.677
MTD+BB	0.516	0.982	0.676
USUB+RCT	0.515	0.983	0.676
PRS+FKZY	0.515	0.984	0.676
BB+SAL	0.514	0.985	0.676
FKZY+KLO	0.514	0.983	0.675
ODR	0.516	0.973	0.675
ZS+TT	0.513	0.982	0.674
MD+ZS	0.511	0.986	0.674
RCT+SIM	0.512	0.982	0.673
BB+FKZY	0.514	0.976	0.673
JAC+SAL	0.512	0.982	0.673
ZS+SAL	0.511	0.983	0.673
USUB+FAG	0.512	0.981	0.673
JAC+SIM	0.511	0.983	0.673
MTD+JAC	0.511	0.981	0.672
CP+SIM	0.511	0.978	0.672
JAC+FKZY	0.510	0.983	0.671
JAC+FAG	0.510	0.982	0.671
PRS+DRK	0.510	0.981	0.671
FSS+FKZY	0.510	0.981	0.671
MD+BB	0.509	0.983	0.670
USUB+TT	0.509	0.982	0.670
USUB+FKZY	0.507	0.987	0.670

Modèle	Précision	Rappel	F-score
BB+SIM	0.508	0.981	0.669
ZS+FAG	0.507	0.983	0.669
BB+FAG	0.508	0.981	0.669
BB+KLO	0.507	0.982	0.669
RCT+BB	0.507	0.983	0.669
CP+BB	0.506	0.984	0.669
JAC+DRK	0.507	0.983	0.669
ODR+JAC	0.507	0.981	0.669
ZS+BB	0.507	0.981	0.669
USUB+SAL	0.506	0.984	0.668
PRS+KLO	0.507	0.982	0.668
PRS+BB	0.506	0.984	0.668
PMI+JAC	0.506	0.984	0.668
PRS+ODR	0.505	0.983	0.667
ZS+ODR	0.505	0.984	0.667
USUB+ZS	0.505	0.982	0.667
JAC+KLO	0.506	0.981	0.667
USUB+MD	0.505	0.985	0.667
PRS+PMI	0.505	0.983	0.667
ZS+DRK	0.504	0.984	0.666
FSS+JAC	0.505	0.980	0.666
RCT+JAC	0.504	0.981	0.666
MD+JAC	0.505	0.979	0.666
ODR+BB	0.504	0.980	0.666
FSS+BB	0.504	0.979	0.665
FSS+USUB	0.502	0.983	0.665
USUB+ODR	0.502	0.984	0.665
PRS+CP	0.502	0.982	0.665
PRS+JAC	0.501	0.981	0.663
PRS+USUB	0.499	0.985	0.663
CP+JAC	0.500	0.977	0.661
USUB+BB	0.499	0.979	0.661
ZS	0.498	0.982	0.661
USUB+JAC	0.496	0.979	0.658
FAG	0.725	0.109	0.189
MD+RCT	0.690	0.034	0.064
RCT+MTD	0.728	0.023	0.045
RCT+KLO	0.682	0.018	0.035
SIM	0.623	0.009	0.017
MD+KLO	0.726	0.008	0.016
MD+MTD	0.732	0.008	0.016
MD+DRK	0.802	0.008	0.015
ODR+FAG	0.665	0.008	0.015
MTD+KLO	0.713	0.006	0.012
MD+PMI	0.822	0.006	0.012
DRK+KLO	0.838	0.006	0.012
MD+FAG	0.676	0.005	0.010
RCT+PMI	0.708	0.005	0.010
MTD+FAG	0.695	0.004	0.009
FSS	0.641	0.004	0.008
PMI+KLO	0.687	0.004	0.008
ODR+RCT	0.728	0.004	0.008
MTD+PMI	0.724	0.003	0.007
CP	0.559	0.002	0.004
MTD+DRK	0.729	0.002	0.003
ODR+KLO	0.714	0.001	0.003
CP+PMI	0.833	0.001	0.001
DRK+FAG	0.708	0.001	0.001
CP+RCT	0.435	0.001	0.001
FSS+FAG	1.000	0.000	0.001
PMI+DRK	1.000	0.000	0.001
FSS+MTD	0.496	0.000	0.001
CP+KLO	0.658	0.000	0.001
FSS+PMI	1.000	0.000	0.001
CP+FAG	0.500	0.000	0.000
PMI+FAG	1.000	0.000	0.000
CP+MD	0.333	0.000	0.000
CP+MTD	0.316	0.000	0.000
FSS+DRK	1.000	0.000	0.000
USUB	∅	∅	∅
JAC	∅	∅	∅
BB	∅	∅	∅
CP+ODR	∅	∅	∅
PRS	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.5 – Résultat de la première phase d'évaluation pour le polonais (tokens informés).

Modèle	Précision	Rappel	F-score
Sfreq+SIM	0.942	0.973	0.958
RCT+Sfreq	0.891	0.947	0.919
Sfreq+SAL	0.893	0.945	0.918
ODR+Sfreq	0.878	0.953	0.914
ZS+Sfreq	0.872	0.952	0.910
Sfreq+FKZY	0.866	0.954	0.908
MTD+Sfreq	0.868	0.951	0.908
MD+Sfreq	0.867	0.951	0.907
TT+Sfreq	0.865	0.952	0.907
Sfreq	0.861	0.950	0.903
USUB+Sfreq	0.859	0.952	0.903
DRK+Sfreq	0.860	0.950	0.902
JAC+Sfreq	0.854	0.952	0.900
PRS+Sfreq	0.851	0.954	0.900
Sfreq+KLO	0.853	0.951	0.899
FSS+Sfreq	0.846	0.958	0.899
Sfreq+FAG	0.849	0.954	0.898
Sfreq+BB	0.847	0.955	0.898
PMI+Sfreq	0.841	0.958	0.895
CP+Sfreq	0.805	0.968	0.879
Sfreq+ttest	0.791	0.973	0.873
Sfreq+D freq	0.781	0.969	0.865
SIM+D freq	0.590	0.883	0.707
SAL+D freq	0.552	0.920	0.690
ttest+D freq	0.541	0.951	0.690
RCT+D freq	0.556	0.903	0.688
D freq	0.544	0.929	0.686
D freq+FAG	0.564	0.874	0.685
D freq+KLO	0.551	0.901	0.684
CP+D freq	0.538	0.939	0.684
ZS+D freq	0.558	0.879	0.683
ODR+D freq	0.560	0.874	0.683
MD+D freq	0.562	0.868	0.682
FSS+D freq	0.561	0.870	0.682
TT+D freq	0.558	0.876	0.682
MTD+D freq	0.558	0.867	0.679
DRK+D freq	0.563	0.837	0.673
USUB+D freq	0.570	0.817	0.672
PMI+D freq	0.564	0.829	0.671
ttest+SIM	0.481	0.985	0.647
TT+SIM	0.504	0.896	0.645
SIM+FAG	0.503	0.879	0.640
BB+D freq	0.576	0.716	0.638
PRS+D freq	0.601	0.677	0.637
SIM+SAL	0.490	0.909	0.637
JAC+D freq	0.586	0.695	0.636
PMI+SIM	0.497	0.874	0.634
SIM+KLO	0.492	0.887	0.633
ZS+SIM	0.484	0.913	0.633
PMI+ttest	0.467	0.979	0.633
MD+SIM	0.488	0.897	0.632
RCT+SIM	0.491	0.885	0.631
TT+ttest	0.465	0.981	0.631
DRK+ttest	0.465	0.979	0.630
ODR+SIM	0.513	0.814	0.630
ttest	0.464	0.978	0.629
ttest+BB	0.463	0.980	0.629
ttest+KLO	0.462	0.982	0.629
JAC+ttest	0.463	0.977	0.628
CP+ttest	0.461	0.984	0.628
MTD+ttest	0.462	0.981	0.628
FSS+ttest	0.461	0.982	0.628
ZS+ttest	0.462	0.980	0.628
PRS+ttest	0.463	0.974	0.628
RCT+ttest	0.462	0.978	0.627
ODR+ttest	0.461	0.978	0.627
FSS+SIM	0.508	0.811	0.624
ttest+FKZY	0.457	0.981	0.624
MD+TT	0.469	0.932	0.624
TT+FAG	0.483	0.878	0.623
USUB+SIM	0.502	0.821	0.623
USUB+ttest	0.456	0.982	0.623
RCT+TT	0.482	0.878	0.622
ttest+FAG	0.455	0.981	0.622
DRK+SIM	0.506	0.805	0.621
TT+SAL	0.476	0.890	0.621
ZS+TT	0.466	0.928	0.621
MTD+SIM	0.467	0.916	0.619
ttest+SAL	0.452	0.979	0.618
FSS+ZS	0.470	0.904	0.618
DRK+SAL	0.471	0.900	0.618
MD	0.472	0.895	0.618
DRK+FAG	0.481	0.860	0.617
ZS+SAL	0.459	0.937	0.617
FSS+TT	0.498	0.809	0.616
ZS+ODR	0.474	0.880	0.616
ZS+FAG	0.455	0.954	0.616
DRK+KLO	0.475	0.876	0.616
PMI+FAG	0.468	0.900	0.616
MD+ttest	0.448	0.985	0.616
ODR+SAL	0.470	0.892	0.616
PMI+DRK	0.480	0.859	0.616
ZS+RCT	0.457	0.945	0.616
ODR+FAG	0.474	0.875	0.615
FSS+SAL	0.466	0.903	0.615
ZS+KLO	0.468	0.895	0.614
PMI+SAL	0.469	0.891	0.614
FSS+FAG	0.476	0.863	0.614
ZS+BB	0.481	0.848	0.614
ZS+PMI	0.460	0.921	0.614
MD+DRK	0.469	0.887	0.614
RCT+DRK	0.471	0.877	0.613
USUB+FAG	0.475	0.866	0.613
MD+PMI	0.472	0.875	0.613
FSS+KLO	0.474	0.865	0.613
TT+BB	0.491	0.813	0.612
PRS+SAL	0.489	0.819	0.612
FSS+RCT	0.468	0.884	0.612
USUB+TT	0.495	0.801	0.612
MD+ODR	0.468	0.884	0.612
TT+FKZY	0.491	0.809	0.611
MTD+PMI	0.454	0.934	0.611
MTD+DRK	0.463	0.898	0.611
SAL+KLO	0.457	0.924	0.611
ZS+DRK	0.471	0.869	0.611
TT+MTD	0.453	0.935	0.611
TT+KLO	0.488	0.816	0.611
RCT+SAL	0.448	0.960	0.611
ODR+RCT	0.466	0.883	0.610
USUB+CP	0.459	0.909	0.610
ODR+TT	0.495	0.794	0.610
TT+PMI	0.498	0.785	0.610
MD+KLO	0.455	0.922	0.610
MD+FAG	0.455	0.923	0.609
TT+DRK	0.486	0.817	0.609
USUB+RCT	0.467	0.878	0.609
RCT+KLO	0.450	0.942	0.609
ODR+MTD	0.459	0.907	0.609
CP+FKZY	0.464	0.885	0.609
FSS+MD	0.461	0.896	0.609
FKZY+KLO	0.487	0.811	0.608
SAL+FAG	0.451	0.932	0.608
SAL+FKZY	0.475	0.842	0.607
PRS+MD	0.486	0.808	0.607
MD+FKZY	0.478	0.831	0.607
ZS+FKZY	0.481	0.821	0.606
MD+SAL	0.449	0.931	0.606
FSS+MTD	0.458	0.895	0.606
BB+SAL	0.476	0.831	0.606
RCT+FKZY	0.486	0.803	0.606
MD+ZS	0.443	0.954	0.605
FKZY+FAG	0.492	0.786	0.605
RCT+FAG	0.443	0.954	0.605
MD+BB	0.484	0.806	0.605
USUB+MTD	0.457	0.893	0.605
FSS+USUB	0.485	0.803	0.605
BB+FAG	0.487	0.792	0.603
USUB+ZS	0.480	0.812	0.603
FSS+DRK	0.483	0.800	0.602
USUB+SAL	0.478	0.815	0.602
CP+ODR	0.440	0.954	0.602
MTD+FKZY	0.469	0.841	0.602
PMI+FKZY	0.481	0.806	0.602
RCT+PMI	0.464	0.858	0.602
USUB+DRK	0.483	0.799	0.602
PRS+MTD	0.470	0.837	0.602
USUB+MD	0.477	0.814	0.602
BB+KLO	0.486	0.789	0.602
MTD+BB	0.469	0.837	0.601
CP+BB	0.471	0.829	0.600
PRS+CP	0.479	0.804	0.600
FSS+CP	0.435	0.966	0.600
JAC+SAL	0.480	0.799	0.600
ZS+JAC	0.476	0.810	0.599
PRS+FAG	0.502	0.743	0.599
PRS+KLO	0.485	0.784	0.599
MTD+JAC	0.470	0.826	0.599
PMI+BB	0.481	0.795	0.599
CP+SIM	0.433	0.972	0.599
MD+RCT	0.434	0.966	0.599
MTD+SAL	0.435	0.962	0.599
MTD	0.449	0.901	0.599
ODR+KLO	0.485	0.781	0.599
MTD+KLO	0.437	0.949	0.599
PRS+RCT	0.489	0.770	0.599
ODR+DRK	0.479	0.797	0.598
CP+JAC	0.476	0.803	0.598
PMI+KLO	0.481	0.791	0.598
USUB+KLO	0.483	0.784	0.598
PRS+ZS	0.483	0.782	0.597
FSS+PMI	0.484	0.778	0.597
SAL	0.437	0.939	0.597
RCT+BB	0.482	0.780	0.596
MD+JAC	0.480	0.786	0.596
MTD+FAG	0.432	0.959	0.596
JAC+KLO	0.485	0.771	0.595
ZS+MTD	0.428	0.968	0.594
FSS+ODR	0.484	0.768	0.594
USUB+ODR	0.481	0.775	0.594
RCT+JAC	0.481	0.773	0.593
CP+TT	0.424	0.977	0.591
MD+MTD	0.425	0.968	0.591
CP+KLO	0.425	0.968	0.591
USUB+PMI	0.478	0.772	0.590
ODR+PMI	0.481	0.750	0.586
CP+PMI	0.420	0.970	0.586
RCT+MTD	0.418	0.976	0.585
CP+DRK	0.416	0.979	0.584
DRK+BB	0.509	0.684	0.584
PRS+TT	0.535	0.641	0.583
TT+JAC	0.532	0.644	0.583
DRK+FKZY	0.513	0.668	0.580
CP+ZS	0.411	0.983	0.579
CP+FAG	0.412	0.976	0.579
JAC+FAG	0.519	0.654	0.578
PRS+DRK	0.515	0.657	0.578
CP+RCT	0.405	0.982	0.573
CP+MD	0.403	0.981	0.571
CP+SAL	0.402	0.982	0.571
CP+MTD	0.399	0.980	0.567
PRS+PMI	0.519	0.623	0.566
PRS+SIM	0.555	0.553	0.554
PMI+JAC	0.528	0.572	0.549
PRS+BB	0.540	0.517	0.528
ZS	0.765	0.020	0.039
SIM	0.563	0.007	0.014
CP	0.676	0.006	0.013
TT	0.769	0.003	0.005
KLO	0.778	0.002	0.004
FAG	0.833	0.001	0.003
RCT	0.667	0.001	0.001
FSS	1.000	0.001	0.001
ODR	1.000	0.000	0.001
USUB+FKZY	∅	∅	∅
JAC+SIM	∅	∅	∅
USUB+BB	∅	∅	∅
FSS+FKZY	∅	∅	∅
DRK	∅	∅	∅
JAC+BB	∅	∅	∅
USUB	∅	∅	∅
FSS+JAC	∅	∅	∅
PMI	∅	∅	∅
PRS+JAC	∅	∅	∅
BB+SIM	∅	∅	∅
JAC	∅	∅	∅
FSS+BB	∅	∅	∅
PRS+FSS	∅	∅	∅
BB	∅	∅	∅
USUB+JAC	∅	∅	∅
PRS+ODR	∅	∅	∅
SIM+FKZY	∅	∅	∅
BB+FKZY	∅	∅	∅
PRS+FKZY	∅	∅	∅
D freq+FKZY	∅	∅	∅
PRS+USUB	∅	∅	∅
ODR+JAC	∅	∅	∅
ODR+FKZY	∅	∅	∅
ODR+BB	∅	∅	∅
PRS	∅	∅	∅
JAC+FKZY	∅	∅	∅
JAC+DRK	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.6 – Résultat de la première phase d'évaluation pour le turc (tokens informés).

Modèle	Précision	Rappel	F-score
Sfreq+SIM	0.991	0.948	0.969
Sfreq+SAL	0.954	0.944	0.949
FSS+Sfreq	0.951	0.943	0.947
PMI+Sfreq	0.953	0.939	0.946
ZS+Sfreq	0.950	0.941	0.945
Sfreq+FAG	0.948	0.942	0.945
DRK+Sfreq	0.944	0.944	0.944
USUB+Sfreq	0.941	0.946	0.943
TT+Sfreq	0.936	0.951	0.943
ODR+Sfreq	0.940	0.945	0.943
Sfreq+KLO	0.939	0.946	0.942
MD+Sfreq	0.930	0.953	0.941
JAC+Sfreq	0.934	0.948	0.941
Sfreq+ttest	0.929	0.952	0.940
Sfreq+BB	0.931	0.949	0.940
Sfreq+FKZY	0.921	0.951	0.936
RCT+Sfreq	0.917	0.954	0.935
Sfreq	0.911	0.956	0.933
PRS+Sfreq	0.914	0.952	0.932
MTD+Sfreq	0.911	0.954	0.932
CP+Sfreq	0.865	0.969	0.914
Sfreq+D freq	0.844	0.966	0.901
MD+D freq	0.671	0.860	0.754
SAL+D freq	0.670	0.859	0.753
FSS+D freq	0.660	0.850	0.743
DRK+D freq	0.659	0.851	0.743
PRS+D freq	0.644	0.869	0.740
JAC+D freq	0.641	0.870	0.738
BB+D freq	0.645	0.863	0.738
D freq+FKZY	0.643	0.866	0.738
SIM+D freq	0.665	0.828	0.737
ODR+D freq	0.665	0.819	0.734
ZS+D freq	0.653	0.835	0.733
MTD+D freq	0.647	0.842	0.732
CP+ttest	0.597	0.935	0.729
TT+D freq	0.652	0.817	0.725
USUB+D freq	0.652	0.816	0.725
JAC+SAL	0.647	0.822	0.724
PRS+SAL	0.630	0.850	0.723
MD+JAC	0.609	0.859	0.713
FSS+JAC	0.600	0.843	0.701
JAC+FKZY	0.598	0.827	0.694
JAC+BB	0.589	0.834	0.690
MTD+JAC	0.578	0.857	0.690
ODR+JAC	0.575	0.853	0.687
DRK+BB	0.597	0.808	0.686
ZS+JAC	0.613	0.778	0.686
USUB+JAC	0.578	0.832	0.682
PRS+MTD	0.576	0.834	0.682
JAC+KLO	0.606	0.775	0.680
FSS+BB	0.588	0.804	0.679
PRS+BB	0.583	0.809	0.678
MTD+FKZY	0.594	0.788	0.677
PRS+ZS	0.600	0.775	0.677
JAC+SIM	0.607	0.764	0.676
BB+SIM	0.596	0.782	0.676
PRS+TT	0.600	0.772	0.675
BB+FKZY	0.573	0.816	0.673
PRS+JAC	0.572	0.817	0.673
PRS+FKZY	0.577	0.807	0.673
PRS+ODR	0.580	0.796	0.671
PRS+SIM	0.607	0.747	0.670
MTD+BB	0.588	0.776	0.669
USUB+FKZY	0.592	0.763	0.667
PRS+DRK	0.587	0.767	0.665
PMI+JAC	0.579	0.766	0.660
JAC+DRK	0.598	0.735	0.660
TT+JAC	0.585	0.755	0.659
PRS+FSS	0.592	0.739	0.658
PRS+USUB	0.573	0.762	0.654
SIM+FKZY	0.609	0.705	0.653
ODR+BB	0.586	0.734	0.652
USUB+BB	0.581	0.738	0.650
ODR+FKZY	0.580	0.727	0.645
FSS+FKZY	0.591	0.710	0.645
PMI+D freq	0.790	0.005	0.010
D freq+KLO	0.788	0.005	0.010
ttest	0.847	0.004	0.009
PRS+MD	0.714	0.001	0.003
SAL+FKZY	0.416	0.001	0.003
ZS+FKZY	0.714	0.001	0.003
BB+SAL	0.382	0.001	0.003
ZS+BB	0.399	0.001	0.001
MTD+SIM	1.000	0.000	0.001
DRK+FKZY	1.000	0.000	0.001
PRS+PMI	0.000	0.000	0.000
ZS+SAL	0.000	0.000	0.000
JAC+FAG	0.000	0.000	0.000
SIM+SAL	0.000	0.000	0.000
TT+BB	0.000	0.000	0.000
MTD+SAL	0.000	0.000	0.000
FSS+CP	∅	∅	∅
ODR+FAG	∅	∅	∅
RCT+DRK	∅	∅	∅
ttest+SIM	∅	∅	∅
PMI+SAL	∅	∅	∅
ZS+ttest	∅	∅	∅
USUB+FAG	∅	∅	∅
RCT+FAG	∅	∅	∅
SIM	∅	∅	∅
USUB+MTD	∅	∅	∅
ZS+FAG	∅	∅	∅
PMI+DRK	∅	∅	∅
CP+TT	∅	∅	∅
SAL+KLO	∅	∅	∅
PMI+BB	∅	∅	∅
RCT+D freq	∅	∅	∅
MTD+ttest	∅	∅	∅
FKZY+KLO	∅	∅	∅
ODR	∅	∅	∅
MD+ODR	∅	∅	∅
ODR+TT	∅	∅	∅
FSS+KLO	∅	∅	∅
USUB+MD	∅	∅	∅
CP+RCT	∅	∅	∅
USUB+TT	∅	∅	∅
D freq	∅	∅	∅
USUB+RCT	∅	∅	∅
MTD+KLO	∅	∅	∅
FSS+MTD	∅	∅	∅
DRK+KLO	∅	∅	∅
BB+KLO	∅	∅	∅
ZS+ODR	∅	∅	∅
CP+FKZY	∅	∅	∅
DRK+ttest	∅	∅	∅
CP+SAL	∅	∅	∅
SIM+FAG	∅	∅	∅
FSS+RCT	∅	∅	∅
FKZY+FAG	∅	∅	∅
CP+PMI	∅	∅	∅
ODR+RCT	∅	∅	∅
DRK+FAG	∅	∅	∅
DRK	∅	∅	∅
USUB+PMI	∅	∅	∅
ZS+SIM	∅	∅	∅
FSS+SAL	∅	∅	∅
RCT+FKZY	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
USUB+ODR	∅	∅	∅
TT+FAG	∅	∅	∅
ttest+SAL	∅	∅	∅
RCT+KLO	∅	∅	∅
RCT+ttest	∅	∅	∅
ZS+PMI	∅	∅	∅
USUB+SIM	∅	∅	∅
MD+MTD	∅	∅	∅
FSS+PMI	∅	∅	∅
ttest+D freq	∅	∅	∅
RCT	∅	∅	∅
FSS+ZS	∅	∅	∅
ODR+KLO	∅	∅	∅
USUB	∅	∅	∅
CP+DRK	∅	∅	∅
USUB+SAL	∅	∅	∅
ODR+DRK	∅	∅	∅
MD+FAG	∅	∅	∅
ZS	∅	∅	∅
MD+TT	∅	∅	∅
ttest+FAG	∅	∅	∅
MTD+PMI	∅	∅	∅
PMI+KLO	∅	∅	∅
ZS+RCT	∅	∅	∅
CP+BB	∅	∅	∅
MD+DRK	∅	∅	∅
ttest+BB	∅	∅	∅
FSS+TT	∅	∅	∅
PMI+FAG	∅	∅	∅
TT+FKZY	∅	∅	∅
USUB+KLO	∅	∅	∅
BB+FAG	∅	∅	∅
PMI	∅	∅	∅
SAL+FAG	∅	∅	∅
TT+SIM	∅	∅	∅
CP+ZS	∅	∅	∅
RCT+JAC	∅	∅	∅
FSS+MD	∅	∅	∅
TT+PMI	∅	∅	∅
CP	∅	∅	∅
TT+ttest	∅	∅	∅
TT	∅	∅	∅
CP+JAC	∅	∅	∅
MD	∅	∅	∅
PRS+ttest	∅	∅	∅
ZS+DRK	∅	∅	∅
TT+KLO	∅	∅	∅
PMI+SIM	∅	∅	∅
JAC	∅	∅	∅
FSS+ODR	∅	∅	∅
CP+MD	∅	∅	∅
JAC+ttest	∅	∅	∅
RCT+PMI	∅	∅	∅
SAL	∅	∅	∅
ODR+SIM	∅	∅	∅
CP+KLO	∅	∅	∅
SIM+KLO	∅	∅	∅
RCT+SIM	∅	∅	∅
PMI+FKZY	∅	∅	∅
ODR+PMI	∅	∅	∅
FSS+DRK	∅	∅	∅
MTD	∅	∅	∅
DRK+SIM	∅	∅	∅
PRS+KLO	∅	∅	∅
FSS+SIM	∅	∅	∅
FSS	∅	∅	∅
PMI+ttest	∅	∅	∅
ZS+TT	∅	∅	∅
USUB+CP	∅	∅	∅
FSS+USUB	∅	∅	∅
ZS+KLO	∅	∅	∅
BB	∅	∅	∅
D freq+FAG	∅	∅	∅
CP+D freq	∅	∅	∅
PRS+FAG	∅	∅	∅
USUB+DRK	∅	∅	∅
CP+MTD	∅	∅	∅
FSS+ttest	∅	∅	∅
ttest+KLO	∅	∅	∅
MD+SAL	∅	∅	∅
TT+SAL	∅	∅	∅
ODR+MTD	∅	∅	∅
MD+ZS	∅	∅	∅
ODR+ttest	∅	∅	∅
RCT+MTD	∅	∅	∅
TT+DRK	∅	∅	∅
MD+RCT	∅	∅	∅
RCT+TT	∅	∅	∅
CP+ODR	∅	∅	∅
MD+KLO	∅	∅	∅
CP+SIM	∅	∅	∅
PRS+RCT	∅	∅	∅
FAG	∅	∅	∅
ZS+MTD	∅	∅	∅
USUB+ttest	∅	∅	∅
MD+ttest	∅	∅	∅
PRS+CP	∅	∅	∅
RCT+BB	∅	∅	∅
MTD+DRK	∅	∅	∅
FSS+FAG	∅	∅	∅
MD+PMI	∅	∅	∅
MD+BB	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
RCT+SAL	∅	∅	∅
DRK+SAL	∅	∅	∅
MD+FKZY	∅	∅	∅
ttest+FKZY	∅	∅	∅
ODR+SAL	∅	∅	∅
MD+SIM	∅	∅	∅
FKZY	∅	∅	∅
USUB+ZS	∅	∅	∅
TT+MTD	∅	∅	∅

TABLE B.7 – Résultat de la première phase d'évaluation pour le chinois (tokens informés).

Modèle	Précision	Rappel	F-score
ZS+Sfreq	0.904	0.755	0.823
DRK+Sfreq	0.923	0.741	0.822
TT+Sfreq	0.918	0.738	0.818
MTD+Sfreq	0.904	0.746	0.817
RCT+Sfreq	0.913	0.728	0.810
MD+Sfreq	0.901	0.734	0.809
USUB+Sfreq	0.928	0.714	0.807
Sfreq+KLO	0.925	0.716	0.807
Sfreq+FKZY	0.911	0.722	0.805
Sfreq+ttest	0.906	0.723	0.805
CP+Sfreq	0.892	0.732	0.804
ODR+Sfreq	0.894	0.724	0.800
PRS+Sfreq	0.923	0.705	0.799
Sfreq+SAL	0.900	0.714	0.797
Sfreq+D freq	0.923	0.691	0.790
JAC+Sfreq	0.890	0.709	0.789
PMI+Sfreq	0.895	0.699	0.785
Sfreq+FAG	0.893	0.699	0.784
FSS+Sfreq	0.909	0.685	0.781
Sfreq+BB	0.907	0.675	0.774
Sfreq+SIM	0.943	0.647	0.768
Sfreq	0.933	0.616	0.742
SAL	0.437	0.908	0.590
RCT	0.388	0.879	0.539
MD	0.408	0.789	0.538
MTD	0.376	0.866	0.524
ZS	0.371	0.875	0.521
ttest	0.350	0.818	0.491
DRK	0.438	0.448	0.443
PMI	0.771	0.006	0.011
TT	0.750	0.001	0.002
FSS+KLO	∅	∅	∅
RCT+PMI	∅	∅	∅
ttest+SAL	∅	∅	∅
PRS+SAL	∅	∅	∅
MD+RCT	∅	∅	∅
USUB+BB	∅	∅	∅
RCT+D freq	∅	∅	∅
BB+FKZY	∅	∅	∅
ODR+ttest	∅	∅	∅
PRS+JAC	∅	∅	∅
DRK+SAL	∅	∅	∅
KLO	∅	∅	∅
FAG	∅	∅	∅
RCT+DRK	∅	∅	∅
PRS+FKZY	∅	∅	∅
TT+BB	∅	∅	∅
PRS+FAG	∅	∅	∅
USUB+JAC	∅	∅	∅
TT+PMI	∅	∅	∅
MTD+FKZY	∅	∅	∅
FSS+TT	∅	∅	∅
PRS+RCT	∅	∅	∅
PMI+FKZY	∅	∅	∅
CP+JAC	∅	∅	∅
ZS+MTD	∅	∅	∅
FSS+RCT	∅	∅	∅
JAC+BB	∅	∅	∅
JAC+FAG	∅	∅	∅
SIM	∅	∅	∅
PMI+JAC	∅	∅	∅
JAC+SAL	∅	∅	∅
RCT+KLO	∅	∅	∅
FSS+FKZY	∅	∅	∅
D freq+KLO	∅	∅	∅
DRK+FAG	∅	∅	∅
MD+JAC	∅	∅	∅
DRK+KLO	∅	∅	∅
D freq	∅	∅	∅
FSS+ttest	∅	∅	∅
USUB+KLO	∅	∅	∅
JAC+KLO	∅	∅	∅
PRS+TT	∅	∅	∅
ODR+D freq	∅	∅	∅
ODR	∅	∅	∅
FSS+MD	∅	∅	∅
FKZY+KLO	∅	∅	∅
BB+FAG	∅	∅	∅
ZS+SIM	∅	∅	∅
CP+MTD	∅	∅	∅
CP+BB	∅	∅	∅
FSS+PMI	∅	∅	∅
ODR+FKZY	∅	∅	∅
MD+FAG	∅	∅	∅
DRK+FKZY	∅	∅	∅
ODR+FAG	∅	∅	∅
FSS+MTD	∅	∅	∅
CP+RCT	∅	∅	∅
ODR+KLO	∅	∅	∅
ZS+TT	∅	∅	∅
MD+BB	∅	∅	∅
CP	∅	∅	∅
CP+DRK	∅	∅	∅
PRS+USUB	∅	∅	∅
FKZY+FAG	∅	∅	∅
JAC+ttest	∅	∅	∅
BB+SIM	∅	∅	∅
FKZY	∅	∅	∅
ZS+ttest	∅	∅	∅
TT+SAL	∅	∅	∅
SIM+SAL	∅	∅	∅
MTD+JAC	∅	∅	∅
CP+ODR	∅	∅	∅
TT+DRK	∅	∅	∅
PRS+PMI	∅	∅	∅
RCT+BB	∅	∅	∅
ttest+D freq	∅	∅	∅
PRS+ODR	∅	∅	∅
MD+MTD	∅	∅	∅
PRS+MTD	∅	∅	∅
USUB+CP	∅	∅	∅
FSS+DRK	∅	∅	∅
RCT+ttest	∅	∅	∅
ZS+ODR	∅	∅	∅
CP+MD	∅	∅	∅
FSS+D freq	∅	∅	∅
BB+D freq	∅	∅	∅
MD+D freq	∅	∅	∅
ZS+FKZY	∅	∅	∅
ODR+DRK	∅	∅	∅
PMI+ttest	∅	∅	∅
PRS+CP	∅	∅	∅
ODR+MTD	∅	∅	∅
JAC+DRK	∅	∅	∅
DRK+BB	∅	∅	∅
BB+SAL	∅	∅	∅
DRK+D freq	∅	∅	∅
FSS+ODR	∅	∅	∅
ZS+DRK	∅	∅	∅
JAC+SIM	∅	∅	∅
MTD+KLO	∅	∅	∅
TT+FKZY	∅	∅	∅
FSS+SAL	∅	∅	∅
ODR+SAL	∅	∅	∅
USUB+SAL	∅	∅	∅
MTD+SIM	∅	∅	∅
MD+KLO	∅	∅	∅
SAL+FKZY	∅	∅	∅
MTD+DRK	∅	∅	∅
MTD+SAL	∅	∅	∅
SAL+KLO	∅	∅	∅
MD+SIM	∅	∅	∅
ODR+BB	∅	∅	∅
DRK+SIM	∅	∅	∅
CP+TT	∅	∅	∅
CP+ZS	∅	∅	∅
PRS+ttest	∅	∅	∅
ZS+FAG	∅	∅	∅
SAL+D freq	∅	∅	∅
JAC+D freq	∅	∅	∅
PRS+BB	∅	∅	∅
MD+DRK	∅	∅	∅
MD+ZS	∅	∅	∅
ZS+SAL	∅	∅	∅
FSS+USUB	∅	∅	∅
RCT+FKZY	∅	∅	∅
USUB+ODR	∅	∅	∅
ZS+KLO	∅	∅	∅
BB	∅	∅	∅
ZS+PMI	∅	∅	∅
ZS+JAC	∅	∅	∅
FSS+FAG	∅	∅	∅
MTD+BB	∅	∅	∅
SIM+KLO	∅	∅	∅
RCT+FAG	∅	∅	∅
ZS+D freq	∅	∅	∅
RCT+MTD	∅	∅	∅
TT+KLO	∅	∅	∅
FSS+ZS	∅	∅	∅
USUB+DRK	∅	∅	∅
USUB+FAG	∅	∅	∅
RCT+SAL	∅	∅	∅
PMI+BB	∅	∅	∅
ttest+FAG	∅	∅	∅
FSS+BB	∅	∅	∅
FSS+CP	∅	∅	∅
MTD+D freq	∅	∅	∅
JAC	∅	∅	∅
ODR+PMI	∅	∅	∅
FSS	∅	∅	∅
MTD+ttest	∅	∅	∅
PMI+D freq	∅	∅	∅
TT+JAC	∅	∅	∅
FSS+SIM	∅	∅	∅
TT+FAG	∅	∅	∅
USUB+FKZY	∅	∅	∅
USUB+RCT	∅	∅	∅
MD+FKZY	∅	∅	∅
JAC+FKZY	∅	∅	∅
USUB+MTD	∅	∅	∅
PMI+SIM	∅	∅	∅
SIM+D freq	∅	∅	∅
PRS+ZS	∅	∅	∅
PMI+FAG	∅	∅	∅
MD+PMI	∅	∅	∅
CP+SIM	∅	∅	∅
ODR+JAC	∅	∅	∅
MTD+PMI	∅	∅	∅
MD+SAL	∅	∅	∅
CP+FAG	∅	∅	∅
CP+D freq	∅	∅	∅
MTD+FAG	∅	∅	∅
PRS+KLO	∅	∅	∅
DRK+ttest	∅	∅	∅
ODR+TT	∅	∅	∅
BB+KLO	∅	∅	∅
MD+TT	∅	∅	∅
USUB+D freq	∅	∅	∅
D freq+FKZY	∅	∅	∅
ZS+RCT	∅	∅	∅
PRS+MD	∅	∅	∅
ZS+BB	∅	∅	∅
USUB+ZS	∅	∅	∅
USUB+TT	∅	∅	∅
MD+ttest	∅	∅	∅
USUB	∅	∅	∅
ttest+FKZY	∅	∅	∅
PRS	∅	∅	∅
PMI+DRK	∅	∅	∅
TT+SIM	∅	∅	∅
PMI+KLO	∅	∅	∅
PRS+DRK	∅	∅	∅
PRS+SIM	∅	∅	∅
TT+ttest	∅	∅	∅
CP+ttest	∅	∅	∅
PRS+D freq	∅	∅	∅
TT+D freq	∅	∅	∅
TT+MTD	∅	∅	∅
CP+SAL	∅	∅	∅
SAL+FAG	∅	∅	∅
USUB+PMI	∅	∅	∅
ttest+BB	∅	∅	∅
ttest+KLO	∅	∅	∅
CP+PMI	∅	∅	∅
USUB+SIM	∅	∅	∅
ODR+SIM	∅	∅	∅
USUB+ttest	∅	∅	∅
PRS+FSS	∅	∅	∅
MD+ODR	∅	∅	∅
RCT+SIM	∅	∅	∅
ODR+RCT	∅	∅	∅
SIM+FKZY	∅	∅	∅
USUB+MD	∅	∅	∅
FSS+JAC	∅	∅	∅
RCT+TT	∅	∅	∅
CP+KLO	∅	∅	∅
SIM+FAG	∅	∅	∅
RCT+JAC	∅	∅	∅
CP+FKZY	∅	∅	∅
PMI+SAL	∅	∅	∅
ttest+SIM	∅	∅	∅
D freq+FAG	∅	∅	∅

TABLE B.8 – Résultat de la première phase d'évaluation pour l'arabe (UTE modérées).

Modèle	Précision	Rappel	F-score
Sfreq+ttest	0.777	0.934	0.848
TT+Sfreq	0.783	0.925	0.848
ZS+Sfreq	0.769	0.939	0.846
MD+Sfreq	0.773	0.933	0.845
CP+Sfreq	0.761	0.941	0.842
PMI+Sfreq	0.756	0.940	0.838
Sfreq+KLO	0.760	0.931	0.837
RCT+Sfreq	0.727	0.949	0.823
FSS+Sfreq	0.721	0.947	0.818
Sfreq+FKZY	0.723	0.942	0.818
PRS+Sfreq	0.712	0.947	0.813
Sfreq+BB	0.705	0.945	0.807
Sfreq+SAL	0.698	0.953	0.806
ODR+Sfreq	0.694	0.947	0.801
USUB+Sfreq	0.680	0.954	0.794
DRK+Sfreq	0.668	0.964	0.789
JAC+Sfreq	0.672	0.952	0.788
Sfreq+FAG	0.669	0.954	0.786
Sfreq	0.670	0.949	0.786
Sfreq+SIM	0.667	0.954	0.785
MTD+Sfreq	0.662	0.955	0.782
Sfreq+D freq	0.634	0.962	0.764
DRK+ttest	0.577	0.892	0.701
DRK+SAL	0.551	0.881	0.678
MTD+SAL	0.540	0.889	0.672
RCT+MTD	0.542	0.859	0.664
PMI+ttest	0.539	0.853	0.660
SAL+D freq	0.575	0.774	0.660
PMI+KLO	0.586	0.728	0.649
ZS+FAG	0.541	0.810	0.649
RCT+DRK	0.578	0.735	0.647
MD+KLO	0.581	0.728	0.646
MD+DRK	0.554	0.770	0.644
ZS+PMI	0.561	0.705	0.625
MD+FAG	0.552	0.712	0.622
ZS+DRK	0.572	0.663	0.614
TT+FKZY	0.494	0.809	0.613
PMI+FAG	0.557	0.638	0.595
MD+PMI	0.516	0.667	0.582
MD+ZS	0.529	0.625	0.573
RCT+PMI	0.579	0.311	0.405
ZS+KLO	0.618	0.287	0.392
BB+FKZY	0.535	0.065	0.115
TT+BB	0.333	0.001	0.001
JAC+BB	0.000	0.000	0.000
USUB+FKZY	∅	∅	∅
ODR+D freq	∅	∅	∅
FSS+CP	∅	∅	∅
ODR+FAG	∅	∅	∅
JAC+SIM	∅	∅	∅
ttest+SIM	∅	∅	∅
PMI+SAL	∅	∅	∅
ZS+ttest	∅	∅	∅
USUB+FAG	∅	∅	∅
RCT+FAG	∅	∅	∅
DRK+D freq	∅	∅	∅
SIM	∅	∅	∅
USUB+MTD	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
CP+TT	∅	∅	∅
SAL+KLO	∅	∅	∅
PMI+BB	∅	∅	∅
ZS+JAC	∅	∅	∅
RCT+D freq	∅	∅	∅
MTD+ttest	∅	∅	∅
FKZY+KLO	∅	∅	∅
JAC+KLO	∅	∅	∅
ODR	∅	∅	∅
MD+ODR	∅	∅	∅
ODR+TT	∅	∅	∅
FSS+KLO	∅	∅	∅
USUB+MD	∅	∅	∅
BB+D freq	∅	∅	∅
CP+RCT	∅	∅	∅
ZS+BB	∅	∅	∅
USUB+TT	∅	∅	∅
D freq	∅	∅	∅
PRS+PMI	∅	∅	∅
MTD+FKZY	∅	∅	∅
USUB+RCT	∅	∅	∅
MTD+KLO	∅	∅	∅
FSS+MTD	∅	∅	∅

Modèle	Précision	Rappel	F-score
DRK+KLO	∅	∅	∅
BB+KLO	∅	∅	∅
FSS+FKZY	∅	∅	∅
ZS+ODR	∅	∅	∅
CP+FKZY	∅	∅	∅
CP+SAL	∅	∅	∅
SIM+FAG	∅	∅	∅
FSS+RCT	∅	∅	∅
FKZY+FAG	∅	∅	∅
SIM+D freq	∅	∅	∅
MD+D freq	∅	∅	∅
CP+PMI	∅	∅	∅
ODR+RCT	∅	∅	∅
DRK+FAG	∅	∅	∅
DRK	∅	∅	∅
FSS+D freq	∅	∅	∅
USUB+PMI	∅	∅	∅
ZS+SIM	∅	∅	∅
FSS+SAL	∅	∅	∅
ZS+SAL	∅	∅	∅
JAC+SAL	∅	∅	∅
RCT+FKZY	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
USUB+ODR	∅	∅	∅
ZS+FKZY	∅	∅	∅
TT+FAG	∅	∅	∅
ttest+SAL	∅	∅	∅
PMI+JAC	∅	∅	∅
RCT+KLO	∅	∅	∅
ZS+D freq	∅	∅	∅
RCT+ttest	∅	∅	∅
USUB+SIM	∅	∅	∅
MD+MTD	∅	∅	∅
TT+JAC	∅	∅	∅
JAC+FAG	∅	∅	∅
FSS+PMI	∅	∅	∅
ttest+D freq	∅	∅	∅
RCT	∅	∅	∅
FSS+ZS	∅	∅	∅
ODR+KLO	∅	∅	∅
USUB	∅	∅	∅
CP+DRK	∅	∅	∅
USUB+SAL	∅	∅	∅
ODR+DRK	∅	∅	∅
ZS	∅	∅	∅
MD+TT	∅	∅	∅
MTD+SIM	∅	∅	∅
ttest+FAG	∅	∅	∅
MTD+PMI	∅	∅	∅
DRK+FKZY	∅	∅	∅
ZS+RCT	∅	∅	∅
CP+BB	∅	∅	∅
FSS+JAC	∅	∅	∅
ttest+BB	∅	∅	∅
PRS+SAL	∅	∅	∅
FSS+TT	∅	∅	∅
USUB+KLO	∅	∅	∅
BB+FAG	∅	∅	∅
BB+SAL	∅	∅	∅
PMI	∅	∅	∅
PRS+JAC	∅	∅	∅
SAL+FAG	∅	∅	∅
JAC+D freq	∅	∅	∅
PRS+SIM	∅	∅	∅
TT+SIM	∅	∅	∅
CP+ZS	∅	∅	∅
RCT+JAC	∅	∅	∅
FSS+MD	∅	∅	∅
TT+PMI	∅	∅	∅
CP	∅	∅	∅
PRS+DRK	∅	∅	∅
TT+ttest	∅	∅	∅
TT	∅	∅	∅
MD+JAC	∅	∅	∅
CP+JAC	∅	∅	∅
BB+SIM	∅	∅	∅
MD	∅	∅	∅
SAL+FKZY	∅	∅	∅
PRS+D freq	∅	∅	∅
ttest	∅	∅	∅
PRS+ttest	∅	∅	∅
TT+KLO	∅	∅	∅
PMI+SIM	∅	∅	∅

Modèle	Précision	Rappel	F-score
JAC	∅	∅	∅
FSS+ODR	∅	∅	∅
CP+MD	∅	∅	∅
JAC+ttest	∅	∅	∅
SAL	∅	∅	∅
FSS+BB	∅	∅	∅
PRS+ZS	∅	∅	∅
SIM+SAL	∅	∅	∅
ODR+SIM	∅	∅	∅
CP+KLO	∅	∅	∅
SIM+KLO	∅	∅	∅
USUB+D freq	∅	∅	∅
RCT+SIM	∅	∅	∅
PMI+FKZY	∅	∅	∅
ODR+PMI	∅	∅	∅
FSS+DRK	∅	∅	∅
TT+D freq	∅	∅	∅
MTD	∅	∅	∅
DRK+SIM	∅	∅	∅
PRS+KLO	∅	∅	∅
FSS+SIM	∅	∅	∅
FSS	∅	∅	∅
ZS+TT	∅	∅	∅
USUB+CP	∅	∅	∅
FSS+USUB	∅	∅	∅
DRK+BB	∅	∅	∅
PRS+FSS	∅	∅	∅
BB	∅	∅	∅
USUB+JAC	∅	∅	∅
D freq+FAG	∅	∅	∅
PRS+ODR	∅	∅	∅
CP+D freq	∅	∅	∅
PRS+FAG	∅	∅	∅
SIM+FKZY	∅	∅	∅
CP+ttest	∅	∅	∅
USUB+DRK	∅	∅	∅
PRS+MD	∅	∅	∅
CP+MTD	∅	∅	∅
PRS+FKZY	∅	∅	∅
FSS+ttest	∅	∅	∅
ttest+KLO	∅	∅	∅
MTD+D freq	∅	∅	∅
MD+SAL	∅	∅	∅
D freq+FKZY	∅	∅	∅
PRS+MTD	∅	∅	∅
TT+SAL	∅	∅	∅
ODR+MTD	∅	∅	∅
ODR+ttest	∅	∅	∅
PMI+D freq	∅	∅	∅
TT+DRK	∅	∅	∅
MD+RCT	∅	∅	∅
PRS+USUB	∅	∅	∅
RCT+TT	∅	∅	∅
CP+ODR	∅	∅	∅
CP+SIM	∅	∅	∅
PRS+RCT	∅	∅	∅
ODR+JAC	∅	∅	∅
FAG	∅	∅	∅
ZS+MTD	∅	∅	∅
USUB+ttest	∅	∅	∅
MD+ttest	∅	∅	∅
PRS+CP	∅	∅	∅
RCT+BB	∅	∅	∅
MTD+DRK	∅	∅	∅
ODR+FKZY	∅	∅	∅
MTD+JAC	∅	∅	∅
FSS+FAG	∅	∅	∅
ODR+BB	∅	∅	∅
MD+BB	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
JAC+FKZY	∅	∅	∅
RCT+SAL	∅	∅	∅
MD+FKZY	∅	∅	∅
JAC+DRK	∅	∅	∅
MTD+BB	∅	∅	∅
ttest+FKZY	∅	∅	∅
ODR+SAL	∅	∅	∅
D freq+KLO	∅	∅	∅
MD+SIM	∅	∅	∅
PRS+TT	∅	∅	∅
FKZY	∅	∅	∅
USUB+ZS	∅	∅	∅
TT+MTD	∅	∅	∅

TABLE B.9 – Résultat de la première phase d'évaluation pour l'allemand (UTE modérées).

Modèle	Précision	Rappel	F-score
Sfreq+SIM	0.584	0.978	0.731
PMI+Sfreq	0.576	0.973	0.724
ODR+Sfreq	0.575	0.972	0.723
Sfreq+ttest	0.569	0.976	0.719
Sfreq+FAG	0.570	0.971	0.719
TT+Sfreq	0.568	0.974	0.718
ZS+Sfreq	0.564	0.978	0.715
CP+Sfreq	0.564	0.978	0.715
FSS+Sfreq	0.563	0.975	0.714
DRK+Sfreq	0.563	0.974	0.714
JAC+Sfreq	0.562	0.971	0.712
Sfreq+KLO	0.561	0.974	0.712
Sfreq+D freq	0.559	0.980	0.711
USUB+Sfreq	0.560	0.974	0.711
RCT+Sfreq	0.556	0.976	0.708
Sfreq	0.554	0.972	0.706
PRS+Sfreq	0.553	0.975	0.705
MD+Sfreq	0.550	0.977	0.704
MTD+Sfreq	0.550	0.977	0.704
Sfreq+BB	0.549	0.978	0.703
Sfreq+FKZY	0.549	0.977	0.703
Sfreq+SAL	0.543	0.977	0.698
ttest+D freq	0.491	0.968	0.651
D freq	0.491	0.969	0.651
PRS+D freq	0.492	0.964	0.651
JAC+D freq	0.484	0.971	0.646
BB+D freq	0.430	0.976	0.597
MD+ZS	0.433	0.828	0.568
PRS+TT	0.436	0.814	0.568
TT+ttest	0.398	0.987	0.567
TT+BB	0.434	0.815	0.566
PRS+SAL	0.432	0.823	0.566
ttest+SIM	0.396	0.991	0.566
JAC+SAL	0.430	0.826	0.565
RCT+PMI	0.434	0.804	0.564
RCT+ttest	0.395	0.988	0.564
PMI+ttest	0.393	0.987	0.562
JAC+DRK	0.428	0.821	0.562
BB+SAL	0.428	0.819	0.562
ZS+ttest	0.390	0.987	0.559
DRK+BB	0.421	0.831	0.559
ODR+ttest	0.389	0.988	0.559
PRS+ttest	0.389	0.988	0.558
DRK+ttest	0.389	0.985	0.558
CP+ttest	0.388	0.990	0.558
USUB+ttest	0.389	0.987	0.558
ZS+MTD	0.401	0.916	0.557
FSS+ttest	0.388	0.990	0.557
MTD+JAC	0.434	0.776	0.557
ttest+BB	0.388	0.987	0.557
RCT+MTD	0.401	0.911	0.557
ttest+KLO	0.387	0.988	0.557
PRS+ODR	0.423	0.812	0.556
MTD+ttest	0.387	0.988	0.556
ttest+FKZY	0.387	0.988	0.556
ttest+FAG	0.387	0.989	0.556
JAC+ttest	0.387	0.987	0.556
ttest	0.387	0.986	0.556
ZS+FAG	0.397	0.924	0.555
MD+ttest	0.385	0.991	0.555
ODR+BB	0.427	0.793	0.555
MTD+BB	0.430	0.780	0.554
TT+FAG	0.450	0.719	0.554
RCT+FAG	0.398	0.910	0.554
JAC+KLO	0.422	0.802	0.553
DRK+KLO	0.418	0.814	0.553
BB+KLO	0.427	0.779	0.552
ttest+SAL	0.382	0.990	0.551
USUB+SIM	0.392	0.923	0.551
MD+RCT	0.391	0.927	0.550
PRS+MTD	0.415	0.813	0.550
MD+FAG	0.418	0.797	0.549
ODR+MTD	0.388	0.916	0.545
FSS+FKZY	0.418	0.779	0.544
FSS+USUB	0.390	0.895	0.544
MD+JAC	0.389	0.901	0.544
MTD	0.418	0.770	0.542
TT+MTD	0.379	0.948	0.542
PRS+DRK	0.378	0.952	0.541
ODR+TT	0.376	0.960	0.540
MD+TT	0.377	0.942	0.539
RCT+TT	0.369	0.974	0.535
PRS+KLO	0.431	0.693	0.531
TT+FKZY	0.363	0.971	0.529

Modèle	Précision	Rappel	F-score
ZS+ODR	0.363	0.974	0.528
PMI+JAC	0.441	0.658	0.528
ZS+TT	0.360	0.982	0.527
ODR+RCT	0.361	0.972	0.527
PRS+PMI	0.355	0.971	0.520
USUB+PMI	0.355	0.965	0.519
JAC+FAG	0.355	0.959	0.519
FSS+PMI	0.352	0.974	0.517
PMI+BB	0.348	0.975	0.513
ODR+FKZY	0.345	0.975	0.510
RCT+JAC	0.341	0.984	0.507
TT+KLO	0.340	0.986	0.505
ZS+KLO	0.337	0.988	0.503
ZS+RCT	0.334	0.991	0.500
TT+SIM	0.332	0.991	0.497
FSS+TT	0.331	0.989	0.496
PMI+SIM	0.331	0.989	0.496
RCT+KLO	0.330	0.992	0.496
FSS+SIM	0.331	0.988	0.495
ZS+JAC	0.328	0.994	0.493
FSS+CP	0.327	0.990	0.492
PRS+MD	0.327	0.992	0.491
FSS+ZS	0.326	0.994	0.491
MTD+KLO	0.326	0.991	0.491
RCT+SIM	0.326	0.994	0.491
USUB+TT	0.326	0.992	0.490
ZS+SIM	0.325	0.995	0.490
PRS+FAG	0.325	0.991	0.489
PRS+RCT	0.324	0.994	0.489
MD+KLO	0.324	0.992	0.488
ZS	0.323	0.995	0.488
ZS+SAL	0.323	0.996	0.488
PRS+ZS	0.322	0.996	0.487
MTD+SIM	0.322	0.993	0.487
MTD+SAL	0.322	0.996	0.487
SIM+SAL	0.322	0.994	0.486
TT+SAL	0.322	0.994	0.486
PMI+FKZY	0.321	0.993	0.485
RCT+SAL	0.321	0.996	0.485
PMI+SAL	0.320	0.995	0.485
SAL+FKZY	0.320	0.994	0.484
FSS+DRK	0.320	0.992	0.484
SAL+KLO	0.320	0.995	0.484
RCT+FKZY	0.320	0.995	0.484
SIM+KLO	0.320	0.993	0.484
FSS+RCT	0.320	0.996	0.484
SIM+FAG	0.320	0.992	0.484
ODR+KLO	0.320	0.991	0.483
FSS+FAG	0.320	0.990	0.483
ODR+SIM	0.319	0.993	0.483
CP+SIM	0.319	0.995	0.483
USUB+RCT	0.319	0.995	0.483
ZS+FKZY	0.319	0.996	0.483
DRK+SAL	0.318	0.995	0.483
MD+BB	0.318	0.994	0.482
ZS+BB	0.318	0.996	0.482
RCT+BB	0.318	0.995	0.481
USUB+ZS	0.317	0.995	0.481
ODR+SAL	0.318	0.995	0.481
MD+SIM	0.317	0.995	0.481
DRK+FKZY	0.317	0.993	0.481
SIM	0.317	0.992	0.481
FSS+KLO	0.317	0.993	0.481
SAL+FAG	0.317	0.995	0.481
MD+SAL	0.317	0.995	0.481
USUB+CP	0.317	0.993	0.481
CP+SAL	0.317	0.995	0.480
FSS+SAL	0.317	0.994	0.480
FSS+MTD	0.316	0.995	0.480
DRK+SIM	0.316	0.994	0.480
SAL	0.316	0.994	0.479
RCT	0.315	0.995	0.479
CP+FKZY	0.316	0.995	0.479
FSS	0.316	0.990	0.479
FAG	0.315	0.992	0.479
CP+BB	0.315	0.995	0.479
USUB+DRK	0.315	0.993	0.479
MTD+FKZY	0.315	0.994	0.478
CP+JAC	0.315	0.996	0.478
USUB+SAL	0.315	0.995	0.478
FKZY	0.315	0.994	0.478
USUB+KLO	0.314	0.993	0.477
PRS+CP	0.314	0.995	0.477
FSS+MD	0.314	0.995	0.477

Modèle	Précision	Rappel	F-score
FKZY+KLO	0.314	0.994	0.477
MD	0.314	0.994	0.477
USUB+FAG	0.314	0.993	0.477
CP	0.313	0.994	0.476
KLO	0.313	0.994	0.476
FSS+ODR	0.313	0.994	0.476
FKZY+FAG	0.313	0.993	0.476
PRS	0.313	0.992	0.476
MD+FKZY	0.312	0.995	0.475
USUB+ODR	0.312	0.995	0.475
JAC	0.312	0.993	0.474
USUB+MTD	0.311	0.994	0.474
USUB+MD	0.311	0.994	0.474
BB+FAG	0.311	0.994	0.473
BB	0.310	0.994	0.473
USUB	0.310	0.994	0.473
TT	0.526	0.054	0.099
PMI	0.574	0.049	0.090
DRK	0.420	0.007	0.015
D freq+FKZY	0.726	0.007	0.014
SAL+D freq	0.611	0.007	0.014
ODR	0.355	0.007	0.014
SIM+D freq	0.167	0.000	0.000
USUB+FKZY	∅	∅	∅
ODR+D freq	∅	∅	∅
ODR+FAG	∅	∅	∅
RCT+DRK	∅	∅	∅
JAC+SIM	∅	∅	∅
DRK+D freq	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
CP+TT	∅	∅	∅
RCT+D freq	∅	∅	∅
MD+ODR	∅	∅	∅
CP+RCT	∅	∅	∅
MD+D freq	∅	∅	∅
CP+PMI	∅	∅	∅
DRK+FAG	∅	∅	∅
FSS+D freq	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
ZS+D freq	∅	∅	∅
JAC+BB	∅	∅	∅
ZS+PMI	∅	∅	∅
MD+MTD	∅	∅	∅
TT+JAC	∅	∅	∅
CP+DRK	∅	∅	∅
ODR+DRK	∅	∅	∅
MTD+PMI	∅	∅	∅
PMI+KLO	∅	∅	∅
FSS+JAC	∅	∅	∅
MD+DRK	∅	∅	∅
PMI+FAG	∅	∅	∅
PRS+JAC	∅	∅	∅
PRS+SIM	∅	∅	∅
CP+ZS	∅	∅	∅
TT+PMI	∅	∅	∅
BB+SIM	∅	∅	∅
ZS+DRK	∅	∅	∅
CP+MD	∅	∅	∅
FSS+BB	∅	∅	∅
CP+KLO	∅	∅	∅
USUB+D freq	∅	∅	∅
ODR+PMI	∅	∅	∅
TT+D freq	∅	∅	∅
PRS+FSS	∅	∅	∅
USUB+JAC	∅	∅	∅
D freq+FAG	∅	∅	∅
CP+D freq	∅	∅	∅
SIM+FKZY	∅	∅	∅
BB+FKZY	∅	∅	∅
CP+MTD	∅	∅	∅
PRS+FKZY	∅	∅	∅
MTD+D freq	∅	∅	∅
PMI+D freq	∅	∅	∅
TT+DRK	∅	∅	∅
PRS+USUB	∅	∅	∅
CP+ODR	∅	∅	∅
ODR+JAC	∅	∅	∅
MTD+DRK	∅	∅	∅
MD+PMI	∅	∅	∅
JAC+FKZY	∅	∅	∅
D freq+KLO	∅	∅	∅

TABLE B.10 – Résultat de la première phase d'évaluation pour l'anglais (UTE modérées).

Modèle	Précision	Rappel	F-score
ZS+Sfreq	0.750	0.917	0.825
ODR+Sfreq	0.733	0.929	0.819
Sfreq+ttest	0.724	0.938	0.817
PRS+Sfreq	0.731	0.919	0.814
MD+Sfreq	0.728	0.917	0.812
MTD+Sfreq	0.722	0.928	0.812
PMI+Sfreq	0.722	0.917	0.808
CP+Sfreq	0.712	0.933	0.808
Sfreq+FAG	0.719	0.917	0.806
Sfreq+D freq	0.716	0.920	0.806
TT+Sfreq	0.706	0.931	0.803
RCT+Sfreq	0.707	0.926	0.801
Sfreq+SIM	0.699	0.936	0.800
Sfreq+KLO	0.713	0.907	0.798
Sfreq	0.718	0.884	0.793
Sfreq+SAL	0.696	0.916	0.791
Sfreq+FKZY	0.672	0.947	0.786
USUB+Sfreq	0.682	0.919	0.783
FSS+Sfreq	0.670	0.931	0.779
DRK+Sfreq	0.672	0.926	0.779
JAC+Sfreq	0.679	0.911	0.778
Sfreq+BB	0.669	0.928	0.777
SIM+D freq	0.585	0.923	0.716
D freq+FKZY	0.589	0.911	0.715
ZS+D freq	0.579	0.934	0.715
ttest+D freq	0.573	0.933	0.710
MD+D freq	0.578	0.919	0.710
D freq+FAG	0.575	0.920	0.708
MTD+D freq	0.569	0.935	0.707
RCT+D freq	0.568	0.935	0.707
ODR+D freq	0.568	0.930	0.705
SAL+D freq	0.562	0.928	0.700
USUB+D freq	0.574	0.888	0.698
D freq+KLO	0.558	0.928	0.697
PMI+D freq	0.548	0.938	0.691
BB+D freq	0.551	0.928	0.691
TT+D freq	0.548	0.935	0.691
DRK+D freq	0.555	0.911	0.690
PRS+D freq	0.555	0.902	0.688
MTD+ttest	0.553	0.878	0.678
FSS+D freq	0.530	0.941	0.678
CP+D freq	0.533	0.925	0.676
ODR+ttest	0.536	0.907	0.674
ZS+ttest	0.515	0.961	0.671
DRK+ttest	0.531	0.911	0.671
RCT+ttest	0.513	0.965	0.670
MD+ttest	0.512	0.958	0.667
ttest+FAG	0.507	0.966	0.665
ttest+FKZY	0.503	0.970	0.663
ttest	0.502	0.970	0.662
PRS+ttest	0.502	0.965	0.660
FSS+ttest	0.501	0.966	0.660
TT+ttest	0.505	0.951	0.660
ttest+SIM	0.502	0.961	0.659
MTD+KLO	0.531	0.860	0.657
PMI+ttest	0.581	0.749	0.654
MD+DRK	0.502	0.936	0.654
ttest+SAL	0.495	0.964	0.654
TT+PMI	0.506	0.914	0.652
BB+SIM	0.554	0.792	0.652
ODR+JAC	0.551	0.791	0.649
ttest+KLO	0.491	0.958	0.649
PRS+FSS	0.561	0.752	0.643
CP+ttest	0.481	0.966	0.643
USUB+ttest	0.481	0.963	0.641
JAC+ttest	0.482	0.957	0.641
ttest+BB	0.480	0.956	0.639
ODR+DRK	0.465	0.950	0.624
RCT+DRK	0.456	0.971	0.620
MD+TT	0.450	0.970	0.615
ZS+DRK	0.451	0.954	0.612
CP+MTD	0.448	0.964	0.612
CP+SAL	0.444	0.978	0.611
PRS+FAG	0.445	0.972	0.610
SAL+FAG	0.443	0.975	0.610
TT+SIM	0.445	0.966	0.609
MTD+SAL	0.443	0.975	0.609
TT+MTD	0.444	0.968	0.609
FSS+TT	0.443	0.972	0.608
ZS+TT	0.442	0.976	0.608
TT+DRK	0.446	0.956	0.608
PRS+MD	0.442	0.974	0.608
MD+SAL	0.442	0.974	0.608
ZS+ODR	0.441	0.976	0.608
MD+BB	0.441	0.972	0.607
FSS+FAG	0.441	0.975	0.607
MTD+SIM	0.441	0.972	0.607
TT+SAL	0.441	0.968	0.606
FSS+MD	0.439	0.976	0.606
FSS+MTD	0.438	0.975	0.605
MTD+FKZY	0.438	0.974	0.605
SIM+SAL	0.438	0.974	0.604
SAL	0.438	0.973	0.604
JAC+SAL	0.438	0.972	0.604
TT+FAG	0.439	0.965	0.604
ODR+TT	0.438	0.970	0.604
MD+KLO	0.437	0.970	0.603
TT+KLO	0.438	0.961	0.602
ODR+SAL	0.436	0.973	0.602
MD+FAG	0.436	0.971	0.602
FSS+SAL	0.435	0.974	0.602
MTD+JAC	0.435	0.974	0.602
MD+ZS	0.435	0.974	0.602
ZS+KLO	0.434	0.978	0.601
USUB+SAL	0.435	0.975	0.601
ZS+RCT	0.434	0.978	0.601
DRK+FAG	0.436	0.966	0.601
PMI+SAL	0.435	0.970	0.601
FSS+ZS	0.433	0.979	0.601
ZS+SIM	0.433	0.978	0.601
MD+SIM	0.435	0.970	0.601
ODR+FAG	0.434	0.973	0.600
FSS+PMI	0.436	0.965	0.600
SIM+FAG	0.434	0.971	0.600
PRS+TT	0.435	0.960	0.599
ODR+KLO	0.432	0.974	0.598
RCT+TT	0.433	0.966	0.598
CP+RCT	0.430	0.981	0.597
DRK+SIM	0.430	0.977	0.597
FSS+DRK	0.431	0.973	0.597
CP+TT	0.433	0.963	0.597
RCT+SAL	0.431	0.971	0.597
DRK+SAL	0.431	0.969	0.597
ZS+FKZY	0.429	0.976	0.596
SAL+FKZY	0.430	0.971	0.596
FSS+RCT	0.429	0.976	0.596
PRS+ZS	0.429	0.976	0.596
PRS+SAL	0.431	0.967	0.596
USUB+TT	0.431	0.965	0.596
JAC+KLO	0.429	0.974	0.595
USUB+MTD	0.429	0.972	0.595
DRK+KLO	0.430	0.965	0.595
FKZY+FAG	0.429	0.969	0.595
RCT+SIM	0.429	0.973	0.595
PRS+SIM	0.429	0.968	0.595
PMI+SIM	0.428	0.975	0.595
USUB+RCT	0.428	0.974	0.595
MD+RCT	0.428	0.972	0.595
ZS+JAC	0.427	0.979	0.594
RCT+KLO	0.429	0.968	0.594
FSS+SIM	0.428	0.973	0.594
SIM+KLO	0.428	0.969	0.594
RCT+JAC	0.427	0.975	0.594
CP+KLO	0.428	0.971	0.594
USUB+ZS	0.426	0.978	0.594
ODR+RCT	0.427	0.973	0.594
PRS+PMI	0.426	0.977	0.594
MD+ODR	0.427	0.973	0.594
MD+JAC	0.428	0.969	0.593
FSS+CP	0.426	0.975	0.593
SAL+KLO	0.428	0.964	0.593
TT+JAC	0.431	0.953	0.593
MD+FKZY	0.427	0.969	0.593
TT+FKZY	0.428	0.965	0.593
ZS+FAG	0.426	0.973	0.593
CP+FAG	0.427	0.969	0.593
PRS+ODR	0.426	0.976	0.593
CP+ODR	0.426	0.972	0.593
PRS+MTD	0.427	0.968	0.592
ZS+SAL	0.426	0.970	0.592
RCT+FKZY	0.426	0.970	0.592
RCT+FAG	0.426	0.969	0.592
USUB+MD	0.426	0.971	0.592
ZS+BB	0.425	0.974	0.592
FSS+KLO	0.426	0.970	0.592
FKZY+KLO	0.426	0.969	0.591
JAC+FAG	0.426	0.966	0.591
RCT+BB	0.423	0.977	0.591
MTD+BB	0.424	0.971	0.590
USUB+FAG	0.423	0.970	0.589
DRK+FKZY	0.423	0.970	0.589
TT+BB	0.425	0.957	0.589
CP+SIM	0.422	0.972	0.588
USUB+SIM	0.422	0.970	0.588
CP+FKZY	0.422	0.971	0.588
PRS+FKZY	0.421	0.969	0.587
USUB+PMI	0.420	0.973	0.587
CP+ZS	0.420	0.974	0.587
PRS+RCT	0.421	0.970	0.587
PRS+KLO	0.422	0.965	0.587
ODR+SIM	0.421	0.966	0.587
PMI+FKZY	0.421	0.968	0.586
BB+FAG	0.421	0.966	0.586
CP+MD	0.420	0.967	0.586
BB+SAL	0.420	0.964	0.585
ODR+FKZY	0.418	0.969	0.584
JAC+SIM	0.418	0.970	0.584
FSS+JAC	0.420	0.960	0.584
USUB+ODR	0.418	0.967	0.584
FSS+ODR	0.418	0.965	0.583
PMI+JAC	0.416	0.973	0.582
CP+BB	0.415	0.969	0.581
USUB+KLO	0.415	0.966	0.581
CP+JAC	0.415	0.967	0.581
PMI+BB	0.414	0.971	0.580
JAC+FKZY	0.415	0.963	0.580
CP+DRK	0.413	0.967	0.579
PRS+CP	0.413	0.965	0.578
FSS+FKZY	0.412	0.965	0.578
USUB+CP	0.412	0.967	0.578
USUB+DRK	0.411	0.968	0.577
BB+KLO	0.413	0.958	0.577
FSS+USUB	0.410	0.966	0.576
ODR+BB	0.409	0.959	0.574
USUB+JAC	0.406	0.964	0.572
ZS	0.592	0.090	0.156
RCT	0.417	0.018	0.034
D freq	0.472	0.011	0.021
SIM	0.083	0.000	0.000
TT	0.000	0.000	0.000
SIM+FKZY	0.000	0.000	0.000
USUB+FKZY	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
ODR	∅	∅	∅
CP+PMI	∅	∅	∅
DRK	∅	∅	∅
MTD+FAG	∅	∅	∅
JAC+BB	∅	∅	∅
ZS+PMI	∅	∅	∅
MD+MTD	∅	∅	∅
USUB	∅	∅	∅
MTD+PMI	∅	∅	∅
PMI+KLO	∅	∅	∅
PMI+FAG	∅	∅	∅
PMI	∅	∅	∅
PRS+JAC	∅	∅	∅
JAC+D freq	∅	∅	∅
CP	∅	∅	∅
PRS+DRK	∅	∅	∅
MD	∅	∅	∅
JAC	∅	∅	∅
RCT+PMI	∅	∅	∅
FSS+BB	∅	∅	∅
ODR+PMI	∅	∅	∅
MTD	∅	∅	∅
FSS	∅	∅	∅
DRK+BB	∅	∅	∅
BB	∅	∅	∅
BB+FKZY	∅	∅	∅
ODR+MTD	∅	∅	∅
RCT+MTD	∅	∅	∅
PRS+USUB	∅	∅	∅
FAG	∅	∅	∅
ZS+MTD	∅	∅	∅
MTD+DRK	∅	∅	∅
MD+PMI	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
JAC+DRK	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.11 – Résultat de la première phase d'évaluation pour le français (UTE modérées).

Modèle	Précision	Rappel	F-score
TT+Sfreq	0.752	0.961	0.844
Sfreq+ttest	0.748	0.962	0.841
PRS+Sfreq	0.740	0.969	0.839
MTD+Sfreq	0.742	0.965	0.839
Sfreq+D freq	0.742	0.963	0.838
Sfreq+SIM	0.735	0.968	0.835
MD+Sfreq	0.735	0.962	0.833
Sfreq+FAG	0.734	0.961	0.832
ZS+Sfreq	0.730	0.962	0.830
Sfreq	0.736	0.951	0.830
Sfreq+SAL	0.728	0.958	0.828
RCT+Sfreq	0.723	0.966	0.827
FSS+Sfreq	0.727	0.957	0.826
ODR+Sfreq	0.723	0.963	0.826
CP+Sfreq	0.722	0.965	0.826
Sfreq+KLO	0.723	0.960	0.825
Sfreq+FKZY	0.718	0.967	0.824
PMI+Sfreq	0.719	0.961	0.823
Sfreq+BB	0.720	0.958	0.822
JAC+Sfreq	0.721	0.956	0.822
DRK+Sfreq	0.717	0.957	0.820
USUB+Sfreq	0.701	0.959	0.810
MTD+D freq	0.645	0.963	0.772
ttest+D freq	0.639	0.961	0.767
PRS+D freq	0.637	0.962	0.766
MD+D freq	0.629	0.968	0.762
SAL+D freq	0.629	0.965	0.761
TT+D freq	0.628	0.966	0.761
SIM+D freq	0.631	0.958	0.761
ZS+D freq	0.623	0.966	0.757
ODR+D freq	0.622	0.966	0.757
RCT+D freq	0.621	0.962	0.755
FSS+D freq	0.619	0.965	0.754
D freq+FKZY	0.618	0.962	0.753
D freq+KLO	0.617	0.965	0.753
PMI+D freq	0.618	0.960	0.752
D freq+FAG	0.617	0.962	0.752
CP+D freq	0.612	0.969	0.750
ttest+FKZY	0.615	0.955	0.748
BB+D freq	0.610	0.963	0.747
DRK+D freq	0.611	0.960	0.747
JAC+D freq	0.611	0.961	0.747
ttest+FAG	0.628	0.915	0.745
D freq	0.606	0.958	0.742
USUB+D freq	0.600	0.972	0.742
FSS+ttest	0.678	0.812	0.739
USUB+ttest	0.604	0.938	0.735
ttest+KLO	0.612	0.919	0.734
TT+ttest	0.587	0.979	0.734
PRS+ttest	0.586	0.976	0.733
MTD+ttest	0.583	0.978	0.731
TT+FKZY	0.602	0.928	0.730
ttest	0.580	0.981	0.729
ttest+SIM	0.662	0.811	0.729
MD+ODR	0.609	0.899	0.726
TT+BB	0.612	0.893	0.726
USUB+RCT	0.600	0.916	0.726
USUB+ZS	0.601	0.915	0.725
FKZY+FAG	0.607	0.898	0.724
SIM+SAL	0.585	0.951	0.724
MD+ttest	0.575	0.979	0.724
CP+TT	0.608	0.894	0.724
ODR+ttest	0.573	0.979	0.723
CP+ttest	0.587	0.941	0.723
RCT+ttest	0.572	0.978	0.722
ttest+SAL	0.571	0.979	0.721
TT+KLO	0.590	0.924	0.720
TT+MTD	0.661	0.790	0.720
FSS+ZS	0.603	0.891	0.719
ZS+ttest	0.569	0.978	0.719
USUB+ODR	0.598	0.900	0.718
MD+ZS	0.575	0.954	0.718
PMI+ttest	0.562	0.977	0.714
PMI+SAL	0.645	0.798	0.713
SIM+KLO	0.574	0.941	0.713
BB+FAG	0.598	0.881	0.713
ttest+BB	0.565	0.965	0.712
DRK+ttest	0.560	0.971	0.710
FSS+ODR	0.572	0.937	0.710
RCT+SAL	0.578	0.917	0.709
PRS+TT	0.563	0.959	0.709
ZS+RCT	0.564	0.952	0.708
MD+FKZY	0.583	0.900	0.708
PRS+DRK	0.649	0.777	0.707
JAC+ttest	0.553	0.977	0.706
ZS+BB	0.568	0.931	0.705
SAL+KLO	0.611	0.830	0.704
RCT+MTD	0.612	0.829	0.704
PRS+ODR	0.559	0.950	0.704
CP+SIM	0.555	0.963	0.704
RCT+TT	0.558	0.952	0.703
FSS+SAL	0.552	0.966	0.702
SAL+FKZY	0.546	0.983	0.702
CP+FKZY	0.618	0.810	0.702
ODR+SAL	0.550	0.969	0.701
PRS+PMI	0.578	0.892	0.701
TT	0.561	0.933	0.701
TT+FAG	0.552	0.957	0.700
ODR+FKZY	0.619	0.806	0.700
ODR+DRK	0.611	0.818	0.699
USUB+FAG	0.549	0.961	0.699
MTD+FKZY	0.544	0.975	0.699
ZS+MTD	0.541	0.981	0.698
SIM+FAG	0.547	0.961	0.697
ZS+FKZY	0.545	0.965	0.697
ODR+BB	0.615	0.803	0.697
ZS+PMI	0.543	0.969	0.696
BB+SIM	0.549	0.949	0.696
ODR+SIM	0.545	0.960	0.696
USUB+BB	0.608	0.808	0.694
TT+DRK	0.540	0.969	0.694
MTD	0.543	0.961	0.693
DRK+FKZY	0.605	0.810	0.693
USUB+KLO	0.542	0.958	0.692
TT+PMI	0.538	0.967	0.691
PMI+JAC	0.539	0.963	0.691
PMI	0.548	0.935	0.691
FSS+TT	0.530	0.980	0.688
BB+KLO	0.534	0.966	0.687
JAC+BB	0.602	0.801	0.687
MD+TT	0.530	0.976	0.687
PRS+SAL	0.527	0.986	0.687
PMI+SIM	0.642	0.739	0.687
FSS+USUB	0.530	0.974	0.687
USUB+TT	0.529	0.975	0.686
FKZY+KLO	0.529	0.975	0.686
ZS	0.528	0.977	0.685
TT+SAL	0.525	0.984	0.685
PRS+ZS	0.527	0.977	0.685
RCT+SIM	0.529	0.969	0.685
SAL	0.526	0.976	0.684
RCT+FKZY	0.524	0.984	0.684
ZS+SIM	0.527	0.974	0.684
MTD+BB	0.536	0.944	0.684
ZS+TT	0.523	0.984	0.683
PRS+RCT	0.520	0.983	0.680
ODR+TT	0.521	0.979	0.680
ZS+DRK	0.521	0.976	0.679
PRS+MTD	0.518	0.983	0.679
USUB+SAL	0.518	0.981	0.678
PRS+FKZY	0.521	0.968	0.678
PRS+FAG	0.516	0.984	0.677
TT+SIM	0.521	0.968	0.677
PRS+BB	0.524	0.958	0.677
USUB+FKZY	0.516	0.983	0.677
PRS+FSS	0.517	0.980	0.677
JAC+FKZY	0.519	0.968	0.676
BB+SAL	0.517	0.976	0.676
PRS+JAC	0.522	0.955	0.675
SIM+FKZY	0.516	0.975	0.675
PRS+MD	0.514	0.978	0.674
JAC+DRK	0.514	0.974	0.673
PRS+USUB	0.512	0.981	0.673
JAC+SIM	0.512	0.980	0.673
USUB+SIM	0.512	0.980	0.672
FSS+SIM	0.513	0.974	0.672
FSS+FKZY	0.511	0.978	0.672
TT+JAC	0.511	0.974	0.671
JAC+SAL	0.509	0.982	0.670
RCT	0.509	0.980	0.670
PRS+KLO	0.508	0.979	0.669
FSS+JAC	0.509	0.976	0.669
ODR	0.508	0.974	0.668
USUB+JAC	0.513	0.954	0.667
MTD+JAC	0.505	0.979	0.667
FSS+BB	0.506	0.972	0.665
PRS+CP	0.504	0.977	0.665
ZS+SAL	0.503	0.981	0.665
ZS+ODR	0.503	0.981	0.665
PRS+SIM	0.504	0.975	0.665
ZS+JAC	0.501	0.981	0.664
RCT+JAC	0.501	0.980	0.663
MD+JAC	0.500	0.981	0.662
ODR+JAC	0.500	0.978	0.662
RCT+BB	0.500	0.979	0.662
JAC+KLO	0.499	0.978	0.661
CP+JAC	0.492	0.978	0.655
JAC+FAG	0.492	0.972	0.653
MD	0.759	0.098	0.173
DRK	0.792	0.051	0.096
ZS+KLO	0.673	0.020	0.038
MTD+SAL	0.569	0.019	0.036
ZS+FAG	0.694	0.014	0.028
MD+SAL	0.660	0.014	0.027
KLO	0.598	0.012	0.024
FAG	0.634	0.012	0.023
DRK+SAL	0.667	0.008	0.016
SAL+FAG	0.565	0.006	0.011
ODR+RCT	0.654	0.004	0.008
PMI+DRK	0.813	0.003	0.006
CP+ZS	0.596	0.003	0.005
ODR+FAG	0.798	0.002	0.005
RCT+PMI	0.887	0.002	0.005
SIM	0.440	0.002	0.004
MD+RCT	0.676	0.002	0.004
MD+FAG	0.684	0.002	0.004
CP+SAL	0.537	0.002	0.003
PMI+KLO	0.844	0.002	0.003
RCT+KLO	0.523	0.001	0.003
MD+DRK	0.818	0.001	0.003
PMI+FAG	0.700	0.001	0.002
CP	0.655	0.001	0.002
ODR+MTD	0.714	0.001	0.001
FSS	0.571	0.001	0.001
ODR+PMI	1.000	0.000	0.001
RCT+DRK	0.600	0.000	0.001
MD+MTD	0.735	0.000	0.001
MTD+DRK	0.500	0.000	0.001
MTD+PMI	0.481	0.000	0.001
CP+PMI	0.500	0.000	0.000
RCT+FAG	0.200	0.000	0.000
MTD+KLO	0.500	0.000	0.000
ODR+KLO	1.000	0.000	0.000
CP+ODR	0.222	0.000	0.000
FSS+KLO	0.000	0.000	0.000
CP+RCT	0.000	0.000	0.000
FSS+CP	∅	∅	∅
USUB+MTD	∅	∅	∅
PMI+BB	∅	∅	∅
USUB+MD	∅	∅	∅
FSS+MTD	∅	∅	∅
DRK+KLO	∅	∅	∅
FSS+RCT	∅	∅	∅
DRK+FAG	∅	∅	∅
USUB+PMI	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
FSS+PMI	∅	∅	∅
USUB	∅	∅	∅
CP+DRK	∅	∅	∅
MTD+SIM	∅	∅	∅
CP+BB	∅	∅	∅
FSS+MD	∅	∅	∅
JAC	∅	∅	∅
CP+MD	∅	∅	∅
CP+KLO	∅	∅	∅
PMI+FKZY	∅	∅	∅
FSS+DRK	∅	∅	∅
DRK+SIM	∅	∅	∅
USUB+CP	∅	∅	∅
DRK+BB	∅	∅	∅
BB	∅	∅	∅
USUB+DRK	∅	∅	∅
BB+FKZY	∅	∅	∅
CP+MTD	∅	∅	∅
MD+KLO	∅	∅	∅
FSS+FAG	∅	∅	∅
MD+PMI	∅	∅	∅
MD+BB	∅	∅	∅
PRS	∅	∅	∅
MD+SIM	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.12 – Résultat de la première phase d'évaluation pour le polonais (UTE modérées).

Modèle	Précision	Rappel	F-score
Sfreq+SIM	0.702	0.969	0.814
ODR+Sfreq	0.709	0.953	0.813
PRS+Sfreq	0.706	0.950	0.810
Sfreq+FAG	0.700	0.956	0.808
Sfreq+FKZY	0.698	0.958	0.808
USUB+Sfreq	0.695	0.960	0.806
RCT+Sfreq	0.699	0.951	0.806
Sfreq+SAL	0.694	0.959	0.805
MTD+Sfreq	0.693	0.959	0.805
Sfreq+KLO	0.696	0.953	0.804
JAC+Sfreq	0.694	0.954	0.803
Sfreq+BB	0.691	0.958	0.803
FSS+Sfreq	0.691	0.959	0.803
ZS+Sfreq	0.690	0.957	0.802
Sfreq	0.691	0.956	0.802
TT+Sfreq	0.687	0.958	0.800
DRK+Sfreq	0.685	0.959	0.799
Sfreq+ttest	0.688	0.952	0.799
PMI+Sfreq	0.684	0.958	0.798
MD+Sfreq	0.678	0.961	0.795
Sfreq+D freq	0.677	0.962	0.795
CP+Sfreq	0.668	0.969	0.790
SIM+D freq	0.595	0.897	0.715
ttest+D freq	0.562	0.960	0.709
RCT+D freq	0.573	0.918	0.706
SAL+D freq	0.572	0.905	0.701
D freq+FAG	0.576	0.895	0.701
CP+D freq	0.556	0.945	0.700
PMI+D freq	0.574	0.890	0.698
MTD+D freq	0.578	0.881	0.698
BB+D freq	0.591	0.848	0.696
MD+D freq	0.565	0.897	0.693
D freq+KLO	0.569	0.886	0.693
ZS+D freq	0.584	0.835	0.687
ODR+D freq	0.581	0.841	0.687
TT+D freq	0.585	0.829	0.686
DRK+D freq	0.572	0.855	0.685
FSS+D freq	0.575	0.841	0.683
TT+ttest	0.502	0.969	0.661
ttest+SIM	0.495	0.975	0.656
D freq	0.601	0.708	0.650
USUB+D freq	0.597	0.707	0.648
ttest+KLO	0.484	0.976	0.647
MTD+ttest	0.484	0.968	0.646
FSS+ttest	0.482	0.974	0.644
DRK+ttest	0.481	0.975	0.644
TT+SIM	0.485	0.956	0.643
ttest+BB	0.479	0.975	0.642
JAC+ttest	0.480	0.969	0.642
ODR+ttest	0.479	0.972	0.642
RCT+ttest	0.478	0.974	0.642
PMI+ttest	0.478	0.974	0.642
ttest+FKZY	0.477	0.973	0.640
CP+ttest	0.476	0.979	0.640
PRS+ttest	0.478	0.966	0.639
ttest+SAL	0.474	0.976	0.638
ZS+ttest	0.475	0.972	0.638
ttest	0.473	0.980	0.638
FSS+TT	0.481	0.942	0.637
ODR+TT	0.482	0.934	0.636
ttest+FAG	0.471	0.974	0.635
PRS+TT	0.527	0.799	0.635
PRS+D freq	0.600	0.674	0.635
USUB+TT	0.493	0.891	0.634
PRS+FAG	0.506	0.849	0.634
USUB+ttest	0.469	0.975	0.633
D freq+FKZY	0.590	0.683	0.633
JAC+D freq	0.593	0.675	0.631
PMI+SIM	0.473	0.940	0.629
MD+ttest	0.465	0.974	0.629
MTD+SIM	0.493	0.869	0.629
DRK+SIM	0.477	0.913	0.627
ZS+SIM	0.465	0.957	0.626
TT+JAC	0.514	0.795	0.624
PRS+RCT	0.493	0.847	0.623
TT	0.504	0.815	0.623
TT+MTD	0.457	0.972	0.622
TT+FKZY	0.506	0.806	0.622
MTD+DRK	0.460	0.957	0.622
PRS+MD	0.481	0.875	0.621
ODR+MTD	0.482	0.870	0.620
TT+BB	0.505	0.803	0.620
USUB+MTD	0.481	0.861	0.617
USUB+ZS	0.474	0.883	0.617

Modèle	Précision	Rappel	F-score
PRS+SAL	0.499	0.808	0.617
MTD+PMI	0.454	0.960	0.617
SIM+KLO	0.455	0.958	0.617
PRS+MTD	0.498	0.808	0.617
RCT+FKZY	0.473	0.883	0.616
FSS+MTD	0.473	0.880	0.616
PRS+CP	0.471	0.887	0.615
ODR+DRK	0.462	0.920	0.615
ZS+JAC	0.499	0.801	0.615
SAL	0.460	0.927	0.615
BB+SAL	0.469	0.892	0.615
PRS+ZS	0.495	0.809	0.614
RCT+TT	0.448	0.976	0.614
SAL+FKZY	0.466	0.898	0.614
ODR+PMI	0.454	0.945	0.614
USUB+SIM	0.505	0.780	0.613
FSS+PMI	0.453	0.948	0.613
TT+PMI	0.448	0.970	0.613
MTD+KLO	0.448	0.967	0.613
RCT+BB	0.470	0.878	0.612
FSS+DRK	0.458	0.921	0.612
USUB+KLO	0.462	0.907	0.612
MTD+FAG	0.448	0.961	0.611
TT+FAG	0.445	0.971	0.610
USUB+PMI	0.471	0.864	0.610
FSS+SIM	0.504	0.772	0.610
ZS+TT	0.444	0.975	0.610
ODR+SIM	0.501	0.778	0.610
FKZY+FAG	0.472	0.860	0.609
RCT+MTD	0.445	0.963	0.609
ZS+ODR	0.449	0.946	0.609
MD+BB	0.466	0.877	0.609
FKZY+KLO	0.472	0.857	0.609
TT+DRK	0.442	0.975	0.608
SIM+FAG	0.444	0.966	0.608
ZS+MTD	0.442	0.969	0.607
MD+FKZY	0.466	0.873	0.607
FSS+ODR	0.475	0.841	0.607
RCT+SIM	0.441	0.971	0.606
PRS+PMI	0.519	0.728	0.606
CP+JAC	0.462	0.876	0.605
PMI+DRK	0.444	0.948	0.605
JAC+SAL	0.486	0.799	0.604
SIM+SAL	0.437	0.977	0.604
ZS+BB	0.473	0.834	0.604
FSS+KLO	0.441	0.955	0.604
TT+KLO	0.438	0.969	0.603
ZS+FKZY	0.478	0.815	0.603
CP+FKZY	0.460	0.872	0.602
PRS+DRK	0.494	0.771	0.602
BB+FAG	0.486	0.792	0.602
FSS+ZS	0.437	0.967	0.602
USUB+SAL	0.438	0.962	0.602
TT+SAL	0.434	0.976	0.601
ZS+DRK	0.436	0.967	0.601
RCT+JAC	0.492	0.772	0.601
CP+SIM	0.435	0.969	0.601
RCT+DRK	0.435	0.971	0.600
USUB+DRK	0.485	0.788	0.600
MTD+BB	0.484	0.790	0.600
DRK+BB	0.489	0.775	0.600
MD+SIM	0.434	0.967	0.599
ODR+KLO	0.437	0.952	0.599
USUB+FAG	0.434	0.966	0.599
JAC+DRK	0.480	0.796	0.599
ODR+FAG	0.435	0.959	0.598
DRK+FKZY	0.489	0.769	0.598
CP	0.471	0.818	0.598
FSS+FAG	0.432	0.967	0.597
DRK+SAL	0.432	0.967	0.597
PRS+KLO	0.488	0.767	0.597
CP+BB	0.474	0.805	0.596
CP+TT	0.428	0.983	0.596
MD+JAC	0.479	0.788	0.596
MTD+FKZY	0.479	0.787	0.596
MD+TT	0.429	0.973	0.596
FSS+MD	0.430	0.966	0.595
ZS+RCT	0.429	0.971	0.595
MTD+SAL	0.427	0.980	0.595
USUB+RCT	0.428	0.974	0.595
MD+MTD	0.428	0.970	0.594
MD+ODR	0.427	0.974	0.594
PMI+BB	0.482	0.771	0.593
JAC+KLO	0.479	0.778	0.593

Modèle	Précision	Rappel	F-score
RCT+KLO	0.425	0.978	0.593
JAC+FAG	0.484	0.763	0.593
ODR+SAL	0.425	0.977	0.592
DRK+KLO	0.426	0.971	0.592
ZS+PMI	0.427	0.964	0.592
FSS+USUB	0.487	0.754	0.592
PMI+JAC	0.486	0.756	0.592
PMI+FAG	0.426	0.971	0.592
FSS+SAL	0.425	0.975	0.592
ZS+KLO	0.424	0.976	0.591
DRK+FAG	0.426	0.966	0.591
ODR+RCT	0.424	0.973	0.591
USUB+ODR	0.487	0.749	0.591
MD+DRK	0.423	0.978	0.591
ZS+SAL	0.423	0.976	0.590
BB+KLO	0.482	0.761	0.590
RCT+FAG	0.423	0.974	0.590
ZS+FAG	0.422	0.978	0.590
PMI+SAL	0.422	0.976	0.590
ODR+FKZY	0.483	0.754	0.589
FSS+FKZY	0.477	0.770	0.589
CP+MTD	0.422	0.974	0.589
FSS+RCT	0.422	0.972	0.589
SAL+KLO	0.420	0.979	0.587
SAL+FAG	0.419	0.976	0.587
PMI+KLO	0.419	0.975	0.586
RCT+PMI	0.420	0.969	0.586
USUB+MD	0.420	0.965	0.586
USUB+CP	0.421	0.959	0.586
PMI+FKZY	0.483	0.741	0.585
RCT+SAL	0.417	0.981	0.585
MD+KLO	0.417	0.976	0.584
CP+DRK	0.414	0.980	0.583
MD+FAG	0.414	0.972	0.581
CP+KLO	0.412	0.983	0.580
MD+RCT	0.412	0.981	0.580
MD+ZS	0.410	0.980	0.578
FSS+CP	0.409	0.982	0.578
MD+PMI	0.410	0.976	0.577
CP+PMI	0.407	0.980	0.575
CP+ZS	0.405	0.986	0.574
CP+RCT	0.405	0.984	0.574
MD+SAL	0.407	0.975	0.574
BB+SIM	0.537	0.615	0.573
SIM+FKZY	0.524	0.632	0.573
CP+ODR	0.405	0.978	0.573
ZS	0.530	0.622	0.572
PRS+FSS	0.521	0.635	0.572
CP+FAG	0.401	0.978	0.569
CP+SAL	0.399	0.984	0.568
ODR+BB	0.513	0.630	0.566
CP+MD	0.392	0.985	0.561
FSS+BB	0.513	0.618	0.561
PRS+ODR	0.523	0.577	0.548
PRS+BB	0.544	0.536	0.540
JAC+BB	0.541	0.471	0.504
SIM	0.475	0.003	0.007
MD	0.909	0.003	0.005
RCT	0.666	0.003	0.005
FAG	0.778	0.002	0.004
KLO	0.857	0.002	0.003
DRK	0.714	0.001	0.003
MTD	0.625	0.001	0.003
ODR	1.000	0.001	0.001
FSS	1.000	0.001	0.001
PRS+SIM	0.000	0.000	0.000
PRS	0.000	0.000	0.000
USUB+FKZY	∅	∅	∅
JAC+SIM	∅	∅	∅
USUB+BB	∅	∅	∅
USUB	∅	∅	∅
FSS+JAC	∅	∅	∅
PMI	∅	∅	∅
PRS+JAC	∅	∅	∅
JAC	∅	∅	∅
BB	∅	∅	∅
USUB+JAC	∅	∅	∅
BB+FKZY	∅	∅	∅
PRS+FKZY	∅	∅	∅
PRS+USUB	∅	∅	∅
ODR+JAC	∅	∅	∅
MTD+JAC	∅	∅	∅
JAC+FKZY	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.13 – Résultat de la première phase d'évaluation pour le turc (UTE modérées).

Modèle	Précision	Rappel	F-score
DRK+Sfreq	0.884	0.973	0.927
Sfreq+KLO	0.887	0.970	0.926
JAC+Sfreq	0.883	0.972	0.926
FSS+Sfreq	0.876	0.978	0.924
Sfreq+SIM	0.875	0.977	0.923
MD+Sfreq	0.876	0.975	0.923
USUB+Sfreq	0.874	0.975	0.922
PMI+Sfreq	0.872	0.977	0.922
Sfreq	0.874	0.972	0.920
Sfreq+FAG	0.868	0.978	0.920
RCT+Sfreq	0.868	0.978	0.919
Sfreq+FKZY	0.867	0.977	0.919
Sfreq+SAL	0.865	0.979	0.919
ODR+Sfreq	0.867	0.978	0.919
ZS+Sfreq	0.865	0.978	0.918
Sfreq+D freq	0.868	0.975	0.918
Sfreq+ttest	0.868	0.972	0.917
PRS+Sfreq	0.861	0.980	0.917
Sfreq+BB	0.861	0.979	0.916
MTD+Sfreq	0.861	0.977	0.916
TT+Sfreq	0.858	0.979	0.915
CP+Sfreq	0.849	0.980	0.910
ttest+D freq	0.693	0.902	0.784
RCT+ttest	0.636	0.944	0.760
ttest+SIM	0.628	0.959	0.759
ttest+BB	0.676	0.863	0.759
ODR+ttest	0.613	0.978	0.753
PRS+ttest	0.648	0.898	0.753
ttest+FAG	0.616	0.966	0.753
USUB+ttest	0.620	0.955	0.752
TT+ttest	0.610	0.980	0.752
ttest+KLO	0.613	0.972	0.752
FSS+ttest	0.607	0.979	0.749
ttest+FKZY	0.627	0.921	0.746
PMI+ttest	0.606	0.964	0.744
DRK+ttest	0.599	0.982	0.744
ttest+SAL	0.596	0.985	0.743
MTD+ttest	0.596	0.983	0.742
MD+ttest	0.604	0.953	0.740
CP+ttest	0.608	0.940	0.738
ZS+ttest	0.589	0.982	0.736
PRS+D freq	0.680	0.786	0.729
JAC+D freq	0.663	0.785	0.719
PRS+JAC	0.600	0.758	0.670
ttest	0.791	0.045	0.084
JAC+BB	0.000	0.000	0.000
USUB+FKZY	∅	∅	∅
ODR+D freq	∅	∅	∅
FSS+CP	∅	∅	∅
ODR+FAG	∅	∅	∅
RCT+DRK	∅	∅	∅
JAC+SIM	∅	∅	∅
PMI+SAL	∅	∅	∅
USUB+FAG	∅	∅	∅
RCT+FAG	∅	∅	∅
DRK+D freq	∅	∅	∅
SIM	∅	∅	∅
USUB+MTD	∅	∅	∅
ZS+FAG	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
CP+TT	∅	∅	∅
SAL+KLO	∅	∅	∅
PMI+BB	∅	∅	∅
ZS+JAC	∅	∅	∅
RCT+D freq	∅	∅	∅
FKZY+KLO	∅	∅	∅
JAC+KLO	∅	∅	∅
ODR	∅	∅	∅
MD+ODR	∅	∅	∅
ODR+TT	∅	∅	∅
FSS+KLO	∅	∅	∅
USUB+MD	∅	∅	∅
BB+D freq	∅	∅	∅
CP+RCT	∅	∅	∅
ZS+BB	∅	∅	∅
USUB+TT	∅	∅	∅
D freq	∅	∅	∅
PRS+PMI	∅	∅	∅
MTD+FKZY	∅	∅	∅
USUB+RCT	∅	∅	∅
MTD+KLO	∅	∅	∅
FSS+MTD	∅	∅	∅
DRK+KLO	∅	∅	∅
BB+KLO	∅	∅	∅
FSS+FKZY	∅	∅	∅
ZS+ODR	∅	∅	∅
CP+FKZY	∅	∅	∅
CP+SAL	∅	∅	∅
SIM+FAG	∅	∅	∅
FSS+RCT	∅	∅	∅
FKZY+FAG	∅	∅	∅
SIM+D freq	∅	∅	∅
MD+D freq	∅	∅	∅
CP+PMI	∅	∅	∅
ODR+RCT	∅	∅	∅
DRK+FAG	∅	∅	∅
DRK	∅	∅	∅
FSS+D freq	∅	∅	∅
USUB+PMI	∅	∅	∅
ZS+SIM	∅	∅	∅
FSS+SAL	∅	∅	∅
ZS+SAL	∅	∅	∅
JAC+SAL	∅	∅	∅
RCT+FKZY	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
USUB+ODR	∅	∅	∅
ZS+FKZY	∅	∅	∅
TT+FAG	∅	∅	∅
PMI+JAC	∅	∅	∅
RCT+KLO	∅	∅	∅
ZS+D freq	∅	∅	∅
ZS+PMI	∅	∅	∅
USUB+SIM	∅	∅	∅
MD+MTD	∅	∅	∅
TT+JAC	∅	∅	∅
JAC+FAG	∅	∅	∅
FSS+PMI	∅	∅	∅
RCT	∅	∅	∅
FSS+ZS	∅	∅	∅
ODR+KLO	∅	∅	∅
USUB	∅	∅	∅
CP+DRK	∅	∅	∅
USUB+SAL	∅	∅	∅
ODR+DRK	∅	∅	∅
MD+FAG	∅	∅	∅
ZS	∅	∅	∅
MD+TT	∅	∅	∅
MTD+SIM	∅	∅	∅
MTD+PMI	∅	∅	∅
PMI+KLO	∅	∅	∅
DRK+FKZY	∅	∅	∅
ZS+RCT	∅	∅	∅
CP+BB	∅	∅	∅
FSS+JAC	∅	∅	∅
MD+DRK	∅	∅	∅
PRS+SAL	∅	∅	∅
FSS+TT	∅	∅	∅
PMI+FAG	∅	∅	∅
TT+FKZY	∅	∅	∅
USUB+KLO	∅	∅	∅
BB+FAG	∅	∅	∅
BB+SAL	∅	∅	∅
PMI	∅	∅	∅
SAL+FAG	∅	∅	∅
PRS+SIM	∅	∅	∅
TT+SIM	∅	∅	∅
CP+ZS	∅	∅	∅
RCT+JAC	∅	∅	∅
FSS+MD	∅	∅	∅
TT+PMI	∅	∅	∅
CP	∅	∅	∅
PRS+DRK	∅	∅	∅
TT	∅	∅	∅
MD+JAC	∅	∅	∅
CP+JAC	∅	∅	∅
BB+SIM	∅	∅	∅
MD	∅	∅	∅
SAL+FKZY	∅	∅	∅
ZS+DRK	∅	∅	∅
TT+KLO	∅	∅	∅
PMI+SIM	∅	∅	∅
JAC	∅	∅	∅
FSS+ODR	∅	∅	∅
CP+MD	∅	∅	∅
JAC+ttest	∅	∅	∅
RCT+PMI	∅	∅	∅
SAL	∅	∅	∅
FSS+BB	∅	∅	∅
PRS+ZS	∅	∅	∅
SIM+SAL	∅	∅	∅
ODR+SIM	∅	∅	∅
CP+KLO	∅	∅	∅
SIM+KLO	∅	∅	∅
USUB+D freq	∅	∅	∅
RCT+SIM	∅	∅	∅
PMI+FKZY	∅	∅	∅
ODR+PMI	∅	∅	∅
FSS+DRK	∅	∅	∅
TT+D freq	∅	∅	∅
MTD	∅	∅	∅
DRK+SIM	∅	∅	∅
PRS+KLO	∅	∅	∅
SAL+D freq	∅	∅	∅
FSS+SIM	∅	∅	∅
FSS	∅	∅	∅
ZS+TT	∅	∅	∅
USUB+CP	∅	∅	∅
FSS+USUB	∅	∅	∅
DRK+BB	∅	∅	∅
ZS+KLO	∅	∅	∅
PRS+FSS	∅	∅	∅
BB	∅	∅	∅
USUB+JAC	∅	∅	∅
D freq+FAG	∅	∅	∅
PRS+ODR	∅	∅	∅
TT+BB	∅	∅	∅
CP+D freq	∅	∅	∅
PRS+FAG	∅	∅	∅
SIM+FKZY	∅	∅	∅
USUB+DRK	∅	∅	∅
PRS+MD	∅	∅	∅
BB+FKZY	∅	∅	∅
CP+MTD	∅	∅	∅
PRS+FKZY	∅	∅	∅
MTD+D freq	∅	∅	∅
MD+SAL	∅	∅	∅
D freq+FKZY	∅	∅	∅
PRS+MTD	∅	∅	∅
TT+SAL	∅	∅	∅
ODR+MTD	∅	∅	∅
MD+ZS	∅	∅	∅
PMI+D freq	∅	∅	∅
RCT+MTD	∅	∅	∅
TT+DRK	∅	∅	∅
MD+RCT	∅	∅	∅
PRS+USUB	∅	∅	∅
RCT+TT	∅	∅	∅
CP+ODR	∅	∅	∅
MD+KLO	∅	∅	∅
CP+SIM	∅	∅	∅
PRS+RCT	∅	∅	∅
ODR+JAC	∅	∅	∅
FAG	∅	∅	∅
ZS+MTD	∅	∅	∅
PRS+CP	∅	∅	∅
MTD+SAL	∅	∅	∅
RCT+BB	∅	∅	∅
MTD+DRK	∅	∅	∅
ODR+FKZY	∅	∅	∅
MTD+JAC	∅	∅	∅
FSS+FAG	∅	∅	∅
ODR+BB	∅	∅	∅
MD+PMI	∅	∅	∅
MD+BB	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
JAC+FKZY	∅	∅	∅
RCT+SAL	∅	∅	∅
DRK+SAL	∅	∅	∅
MD+FKZY	∅	∅	∅
JAC+DRK	∅	∅	∅
MTD+BB	∅	∅	∅
ODR+SAL	∅	∅	∅
D freq+KLO	∅	∅	∅
MD+SIM	∅	∅	∅
PRS+TT	∅	∅	∅
FKZY	∅	∅	∅
USUB+ZS	∅	∅	∅
TT+MTD	∅	∅	∅

TABLE B.14 – Résultat de la première phase d'évaluation pour l'arabe (UTE franches).

Modèle	Précision	Rappel	F-score
Sfreq+ttest	0.813	0.945	0.874
CP+Sfreq	0.830	0.922	0.874
PRS+Sfreq	0.828	0.918	0.871
FSS+Sfreq	0.830	0.916	0.871
TT+Sfreq	0.818	0.928	0.869
MD+Sfreq	0.780	0.952	0.857
JAC+Sfreq	0.880	0.833	0.856
DRK+Sfreq	0.776	0.953	0.856
ZS+Sfreq	0.773	0.954	0.854
Sfreq+FAG	0.774	0.948	0.852
Sfreq+SAL	0.770	0.954	0.852
Sfreq+FKZY	0.886	0.820	0.852
Sfreq+KLO	0.767	0.953	0.850
ODR+Sfreq	0.768	0.950	0.850
PMI+Sfreq	0.763	0.957	0.849
USUB+Sfreq	0.761	0.952	0.846
RCT+Sfreq	0.751	0.955	0.841
Sfreq	0.741	0.953	0.834
Sfreq+BB	0.735	0.950	0.828
Sfreq+SIM	0.726	0.942	0.820
MTD+Sfreq	0.709	0.967	0.818
Sfreq+D freq	0.700	0.963	0.810
ttest+KLO	0.621	0.851	0.718
DRK+ttest	0.590	0.878	0.706
ttest+FAG	0.600	0.782	0.679
MD+MTD	0.532	0.901	0.669
ZS+SAL	0.537	0.865	0.662
PRS+FKZY	0.539	0.839	0.656
MTD+SAL	0.496	0.957	0.653
TT+FKZY	0.534	0.807	0.643
MTD+PMI	0.538	0.797	0.642
RCT+DRK	0.596	0.677	0.634
RCT+MTD	0.569	0.706	0.630
USUB+BB	0.572	0.694	0.627
PRS+FSS	0.502	0.820	0.623
PMI+SAL	0.554	0.709	0.622
ODR+BB	0.503	0.807	0.620
ZS+RCT	0.502	0.806	0.619
ZS+MTD	0.526	0.743	0.616
PRS+TT	0.496	0.793	0.610
CP+SIM	0.536	0.697	0.606
CP+JAC	0.509	0.734	0.601
JAC+SIM	0.566	0.632	0.597
SIM+FKZY	0.573	0.618	0.595
PRS+SIM	0.528	0.678	0.594
PRS+JAC	0.544	0.626	0.582
MTD+DRK	0.566	0.590	0.578
CP+BB	0.576	0.488	0.528
DRK+SAL	0.564	0.440	0.494
SAL+KLO	0.551	0.430	0.483
ZS+DRK	0.593	0.317	0.413
FSS+BB	0.612	0.268	0.373
JAC+FKZY	0.596	0.218	0.320
USUB+JAC	0.542	0.203	0.296
USUB+FKZY	0.551	0.165	0.254
USUB+SIM	0.508	0.156	0.238
CP+FKZY	0.580	0.136	0.221
PRS+BB	0.557	0.092	0.157
FSS+TT	0.473	0.005	0.010
USUB+TT	0.271	0.003	0.005
TT+SIM	0.497	0.001	0.001
CP+TT	0.331	0.001	0.001
FSS+FKZY	0.000	0.000	0.000
TT+JAC	0.000	0.000	0.000
BB+SIM	0.000	0.000	0.000
FSS+SIM	0.000	0.000	0.000
TT+BB	0.000	0.000	0.000
BB+FKZY	0.000	0.000	0.000
PRS+CP	0.000	0.000	0.000
ODR+D freq	∅	∅	∅
FSS+CP	∅	∅	∅
ODR+FAG	∅	∅	∅
ttest+SIM	∅	∅	∅
ZS+ttest	∅	∅	∅
USUB+FAG	∅	∅	∅
RCT+FAG	∅	∅	∅
DRK+D freq	∅	∅	∅
SIM	∅	∅	∅
USUB+MTD	∅	∅	∅
ZS+FAG	∅	∅	∅
PMI+DRK	∅	∅	∅
PMI+BB	∅	∅	∅
ZS+JAC	∅	∅	∅
RCT+D freq	∅	∅	∅
MTD+ttest	∅	∅	∅
FKZY+KLO	∅	∅	∅
JAC+KLO	∅	∅	∅
ODR	∅	∅	∅
MD+ODR	∅	∅	∅
ODR+TT	∅	∅	∅
FSS+KLO	∅	∅	∅
USUB+MD	∅	∅	∅
BB+D freq	∅	∅	∅
CP+RCT	∅	∅	∅
ZS+BB	∅	∅	∅
D freq	∅	∅	∅
PRS+PMI	∅	∅	∅
MTD+FKZY	∅	∅	∅
USUB+RCT	∅	∅	∅
MTD+KLO	∅	∅	∅
FSS+MTD	∅	∅	∅
DRK+KLO	∅	∅	∅
BB+KLO	∅	∅	∅
ZS+ODR	∅	∅	∅
CP+SAL	∅	∅	∅
SIM+FAG	∅	∅	∅
FSS+RCT	∅	∅	∅
FKZY+FAG	∅	∅	∅
SIM+D freq	∅	∅	∅
MD+D freq	∅	∅	∅
CP+PMI	∅	∅	∅
ODR+RCT	∅	∅	∅
DRK+FAG	∅	∅	∅
DRK	∅	∅	∅
FSS+D freq	∅	∅	∅
USUB+PMI	∅	∅	∅
ZS+SIM	∅	∅	∅
FSS+SAL	∅	∅	∅
JAC+SAL	∅	∅	∅
RCT+FKZY	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
USUB+ODR	∅	∅	∅
ZS+FKZY	∅	∅	∅
TT+FAG	∅	∅	∅
ttest+SAL	∅	∅	∅
PMI+JAC	∅	∅	∅
RCT+KLO	∅	∅	∅
ZS+D freq	∅	∅	∅
JAC+BB	∅	∅	∅
RCT+ttest	∅	∅	∅
ZS+PMI	∅	∅	∅
JAC+FAG	∅	∅	∅
FSS+PMI	∅	∅	∅
ttest+D freq	∅	∅	∅
RCT	∅	∅	∅
FSS+ZS	∅	∅	∅
ODR+KLO	∅	∅	∅
USUB	∅	∅	∅
CP+DRK	∅	∅	∅
USUB+SAL	∅	∅	∅
ODR+DRK	∅	∅	∅
MD+FAG	∅	∅	∅
ZS	∅	∅	∅
MD+TT	∅	∅	∅
MTD+SIM	∅	∅	∅
PMI+KLO	∅	∅	∅
DRK+FKZY	∅	∅	∅
FSS+JAC	∅	∅	∅
MD+DRK	∅	∅	∅
ttest+BB	∅	∅	∅
PRS+SAL	∅	∅	∅
PMI+FAG	∅	∅	∅
USUB+KLO	∅	∅	∅
BB+FAG	∅	∅	∅
BB+SAL	∅	∅	∅
PMI	∅	∅	∅
SAL+FAG	∅	∅	∅
JAC+D freq	∅	∅	∅
CP+ZS	∅	∅	∅
RCT+JAC	∅	∅	∅
FSS+MD	∅	∅	∅
TT+PMI	∅	∅	∅
CP	∅	∅	∅
PRS+DRK	∅	∅	∅
TT+ttest	∅	∅	∅
TT	∅	∅	∅
MD+JAC	∅	∅	∅
MD	∅	∅	∅
SAL+FKZY	∅	∅	∅
PRS+D freq	∅	∅	∅
ttest	∅	∅	∅
PRS+ttest	∅	∅	∅
TT+KLO	∅	∅	∅
PMI+SIM	∅	∅	∅
JAC	∅	∅	∅
FSS+ODR	∅	∅	∅
CP+MD	∅	∅	∅
JAC+ttest	∅	∅	∅
RCT+PMI	∅	∅	∅
SAL	∅	∅	∅
PRS+ZS	∅	∅	∅
SIM+SAL	∅	∅	∅
ODR+SIM	∅	∅	∅
CP+KLO	∅	∅	∅
SIM+KLO	∅	∅	∅
USUB+D freq	∅	∅	∅
RCT+SIM	∅	∅	∅
PMI+FKZY	∅	∅	∅
ODR+PMI	∅	∅	∅
FSS+DRK	∅	∅	∅
TT+D freq	∅	∅	∅
MTD	∅	∅	∅
DRK+SIM	∅	∅	∅
PRS+KLO	∅	∅	∅
SAL+D freq	∅	∅	∅
FSS	∅	∅	∅
PMI+ttest	∅	∅	∅
ZS+TT	∅	∅	∅
USUB+CP	∅	∅	∅
FSS+USUB	∅	∅	∅
DRK+BB	∅	∅	∅
ZS+KLO	∅	∅	∅
BB	∅	∅	∅
D freq+FAG	∅	∅	∅
PRS+ODR	∅	∅	∅
CP+D freq	∅	∅	∅
PRS+FAG	∅	∅	∅
CP+ttest	∅	∅	∅
USUB+DRK	∅	∅	∅
PRS+MD	∅	∅	∅
CP+MTD	∅	∅	∅
FSS+ttest	∅	∅	∅
MTD+D freq	∅	∅	∅
MD+SAL	∅	∅	∅
D freq+FKZY	∅	∅	∅
PRS+MTD	∅	∅	∅
TT+SAL	∅	∅	∅
ODR+MTD	∅	∅	∅
MD+ZS	∅	∅	∅
ODR+ttest	∅	∅	∅
PMI+D freq	∅	∅	∅
TT+DRK	∅	∅	∅
MD+RCT	∅	∅	∅
PRS+USUB	∅	∅	∅
RCT+TT	∅	∅	∅
CP+ODR	∅	∅	∅
MD+KLO	∅	∅	∅
PRS+RCT	∅	∅	∅
ODR+JAC	∅	∅	∅
FAG	∅	∅	∅
USUB+ttest	∅	∅	∅
MD+ttest	∅	∅	∅
RCT+BB	∅	∅	∅
ODR+FKZY	∅	∅	∅
MTD+JAC	∅	∅	∅
FSS+FAG	∅	∅	∅
MD+PMI	∅	∅	∅
MD+BB	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
RCT+SAL	∅	∅	∅
MD+FKZY	∅	∅	∅
JAC+DRK	∅	∅	∅
MTD+BB	∅	∅	∅
ttest+FKZY	∅	∅	∅
ODR+SAL	∅	∅	∅
D freq+KLO	∅	∅	∅
MD+SIM	∅	∅	∅
FKZY	∅	∅	∅
USUB+ZS	∅	∅	∅
TT+MTD	∅	∅	∅

TABLE B.15 – Résultat de la première phase d'évaluation pour l'allemand (UTE franches).

Modèle	Précision	Rappel	F-score
Sfreq+ttest	0.470	0.979	0.635
Sfreq+SIM	0.468	0.985	0.635
Sfreq+D freq	0.468	0.984	0.634
RCT+Sfreq	0.457	0.981	0.624
ODR+Sfreq	0.456	0.983	0.623
TT+Sfreq	0.456	0.981	0.622
PRS+Sfreq	0.454	0.981	0.621
JAC+Sfreq	0.454	0.977	0.620
Sfreq+FAG	0.453	0.982	0.620
DRK+Sfreq	0.453	0.980	0.619
CP+Sfreq	0.451	0.984	0.619
CP+D freq	0.475	0.884	0.618
ZS+D freq	0.462	0.931	0.617
PRS+D freq	0.459	0.939	0.617
MTD+Sfreq	0.449	0.981	0.617
Sfreq	0.447	0.980	0.614
MTD+D freq	0.507	0.779	0.614
Sfreq+KLO	0.447	0.982	0.614
ZS+Sfreq	0.444	0.985	0.612
PMI+Sfreq	0.443	0.984	0.611
Sfreq+FKZY	0.443	0.981	0.610
MD+Sfreq	0.441	0.984	0.609
Sfreq+BB	0.439	0.979	0.606
Sfreq+SAL	0.437	0.981	0.605
FSS+Sfreq	0.437	0.983	0.605
USUB+Sfreq	0.437	0.982	0.605
ttest+D freq	0.437	0.972	0.603
JAC+D freq	0.432	0.966	0.597
D freq	0.429	0.972	0.595
SAL+D freq	0.483	0.756	0.590
TT+BB	0.433	0.807	0.564
PRS+TT	0.431	0.815	0.564
ZS+FAG	0.428	0.817	0.561
ZS+DRK	0.432	0.800	0.561
ZS+MTD	0.432	0.796	0.560
ZS+PMI	0.428	0.811	0.560
TT+JAC	0.429	0.807	0.560
BB+D freq	0.393	0.973	0.560
JAC+DRK	0.422	0.822	0.558
PRS+KLO	0.430	0.792	0.557
ODR+BB	0.425	0.806	0.556
BB+SAL	0.424	0.808	0.556
ODR+JAC	0.428	0.793	0.556
PRS+ODR	0.426	0.794	0.555
PMI+JAC	0.423	0.800	0.553
DRK+BB	0.418	0.817	0.553
JAC+SAL	0.425	0.792	0.553
USUB+FKZY	0.423	0.796	0.553
PMI	0.435	0.755	0.552
BB+KLO	0.416	0.807	0.549
ZS+ODR	0.391	0.921	0.549
MTD+SIM	0.431	0.756	0.549
PMI+SIM	0.396	0.890	0.548
JAC+KLO	0.422	0.778	0.547
TT+KLO	0.387	0.933	0.547
FSS+FKZY	0.428	0.755	0.546
CP+SIM	0.424	0.756	0.543
PRS+SAL	0.386	0.915	0.543
FSS+FAG	0.426	0.739	0.540
FSS+MD	0.414	0.776	0.540
MTD	0.416	0.766	0.539
ttest+SIM	0.369	0.989	0.538
USUB+DRK	0.382	0.902	0.536
DRK+SIM	0.375	0.935	0.535
TT+ttest	0.364	0.990	0.532
USUB+SIM	0.370	0.941	0.532
PMI+ttest	0.363	0.992	0.531
DRK+SAL	0.367	0.948	0.529
ttest+FKZY	0.362	0.981	0.529
PRS+ttest	0.360	0.990	0.529
ODR+ttest	0.360	0.992	0.528
MTD+ttest	0.359	0.991	0.528
DRK+ttest	0.359	0.991	0.528
RCT+ttest	0.359	0.991	0.527
ZS+ttest	0.359	0.990	0.527
CP+ttest	0.358	0.991	0.526
FSS+ttest	0.359	0.988	0.526
FAG	0.413	0.721	0.526
JAC+ttest	0.357	0.991	0.525
ttest	0.357	0.989	0.525
MD+ttest	0.356	0.993	0.524
USUB	0.406	0.737	0.524
ttest+FAG	0.356	0.992	0.524
JAC+FAG	0.360	0.955	0.523

Modèle	Précision	Rappel	F-score
MD+JAC	0.358	0.962	0.522
USUB+ttest	0.354	0.991	0.522
ttest+SAL	0.354	0.991	0.522
ttest+KLO	0.354	0.991	0.521
ttest+BB	0.353	0.989	0.521
PMI+FKZY	0.357	0.960	0.521
FKZY	0.355	0.956	0.518
ZS+TT	0.351	0.979	0.517
PRS+DRK	0.352	0.970	0.517
FSS+RCT	0.350	0.972	0.515
ZS+RCT	0.344	0.986	0.510
FSS+ODR	0.338	0.981	0.502
MD	0.338	0.968	0.501
FSS+USUB	0.336	0.981	0.501
PMI+BB	0.337	0.978	0.501
CP+SAL	0.334	0.983	0.499
ZS+KLO	0.333	0.987	0.498
FSS+SIM	0.332	0.986	0.496
MTD+FKZY	0.330	0.985	0.495
FSS+TT	0.329	0.988	0.494
MTD+SAL	0.329	0.987	0.494
TT+FKZY	0.328	0.991	0.493
USUB+RCT	0.328	0.989	0.492
CP+FKZY	0.328	0.986	0.492
PMI+SAL	0.327	0.988	0.491
TT+SIM	0.325	0.993	0.490
FSS+ZS	0.325	0.993	0.490
DRK+FKZY	0.326	0.987	0.490
PRS	0.326	0.984	0.490
PRS+PMI	0.325	0.989	0.489
USUB+TT	0.325	0.989	0.489
TT+SAL	0.324	0.991	0.489
ODR+FKZY	0.323	0.990	0.487
SAL+FAG	0.323	0.989	0.486
USUB+MD	0.322	0.988	0.486
KLO	0.322	0.988	0.485
MD+BB	0.321	0.990	0.485
SIM+FAG	0.321	0.992	0.485
RCT+SIM	0.321	0.994	0.485
USUB+FAG	0.321	0.986	0.484
ZS+SIM	0.320	0.994	0.484
SAL+KLO	0.320	0.991	0.484
PRS+ZS	0.320	0.994	0.484
FSS+KLO	0.320	0.991	0.484
MD+SIM	0.320	0.993	0.484
SIM+KLO	0.319	0.994	0.483
PRS+MTD	0.320	0.990	0.483
ODR+SAL	0.319	0.992	0.483
ZS+JAC	0.318	0.994	0.482
PRS+RCT	0.318	0.996	0.482
RCT+JAC	0.317	0.994	0.481
MD+SAL	0.318	0.992	0.481
ZS+SAL	0.317	0.994	0.481
ODR+SIM	0.317	0.994	0.481
USUB+KLO	0.317	0.990	0.481
SIM	0.317	0.993	0.480
PRS+FAG	0.317	0.993	0.480
PRS+CP	0.316	0.994	0.480
USUB+ZS	0.316	0.995	0.480
SIM+SAL	0.316	0.995	0.480
FKZY+FAG	0.316	0.991	0.480
ZS+FKZY	0.316	0.995	0.479
ZS	0.315	0.995	0.479
RCT+FKZY	0.315	0.995	0.479
USUB+ODR	0.316	0.990	0.479
RCT+SAL	0.315	0.994	0.479
ZS+BB	0.315	0.995	0.478
CP+JAC	0.315	0.994	0.478
PRS+MD	0.314	0.993	0.478
BB+FAG	0.315	0.992	0.478
FSS+SAL	0.314	0.994	0.477
JAC	0.314	0.990	0.477
RCT	0.314	0.994	0.477
MTD+JAC	0.314	0.993	0.477
RCT+BB	0.313	0.995	0.476
MD+FKZY	0.313	0.993	0.476
MTD+BB	0.313	0.993	0.476
FKZY+KLO	0.312	0.994	0.475
CP	0.312	0.994	0.475
SAL	0.312	0.994	0.475
SAL+FKZY	0.312	0.994	0.475
USUB+SAL	0.311	0.994	0.474
CP+BB	0.310	0.995	0.473
BB	0.310	0.992	0.473

Modèle	Précision	Rappel	F-score
TT	0.539	0.045	0.083
FSS	0.362	0.010	0.020
ODR	0.487	0.006	0.012
DRK	0.379	0.002	0.005
SIM+D freq	0.000	0.000	0.000
ODR+D freq	∅	∅	∅
FSS+CP	∅	∅	∅
ODR+FAG	∅	∅	∅
RCT+DRK	∅	∅	∅
JAC+SIM	∅	∅	∅
RCT+FAG	∅	∅	∅
DRK+D freq	∅	∅	∅
USUB+MTD	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
CP+TT	∅	∅	∅
RCT+D freq	∅	∅	∅
MD+ODR	∅	∅	∅
ODR+TT	∅	∅	∅
CP+RCT	∅	∅	∅
MTD+KLO	∅	∅	∅
FSS+MTD	∅	∅	∅
DRK+KLO	∅	∅	∅
MD+D freq	∅	∅	∅
CP+PMI	∅	∅	∅
ODR+RCT	∅	∅	∅
DRK+FAG	∅	∅	∅
FSS+D freq	∅	∅	∅
USUB+PMI	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
TT+FAG	∅	∅	∅
RCT+KLO	∅	∅	∅
JAC+BB	∅	∅	∅
MD+MTD	∅	∅	∅
FSS+PMI	∅	∅	∅
ODR+KLO	∅	∅	∅
CP+DRK	∅	∅	∅
ODR+DRK	∅	∅	∅
MD+FAG	∅	∅	∅
MD+TT	∅	∅	∅
MTD+PMI	∅	∅	∅
PMI+KLO	∅	∅	∅
FSS+JAC	∅	∅	∅
MD+DRK	∅	∅	∅
PMI+FAG	∅	∅	∅
PRS+JAC	∅	∅	∅
PRS+SIM	∅	∅	∅
CP+ZS	∅	∅	∅
TT+PMI	∅	∅	∅
BB+SIM	∅	∅	∅
CP+MD	∅	∅	∅
RCT+PMI	∅	∅	∅
FSS+BB	∅	∅	∅
CP+KLO	∅	∅	∅
USUB+D freq	∅	∅	∅
ODR+PMI	∅	∅	∅
FSS+DRK	∅	∅	∅
TT+D freq	∅	∅	∅
USUB+CP	∅	∅	∅
PRS+FSS	∅	∅	∅
USUB+JAC	∅	∅	∅
D freq+FAG	∅	∅	∅
SIM+FKZY	∅	∅	∅
BB+FKZY	∅	∅	∅
CP+MTD	∅	∅	∅
PRS+FKZY	∅	∅	∅
D freq+FKZY	∅	∅	∅
ODR+MTD	∅	∅	∅
MD+ZS	∅	∅	∅
PMI+D freq	∅	∅	∅
RCT+MTD	∅	∅	∅
TT+DRK	∅	∅	∅
MD+RCT	∅	∅	∅
PRS+USUB	∅	∅	∅
RCT+TT	∅	∅	∅
CP+ODR	∅	∅	∅
MD+KLO	∅	∅	∅
MTD+DRK	∅	∅	∅
MD+PMI	∅	∅	∅
JAC+FKZY	∅	∅	∅
D freq+KLO	∅	∅	∅
TT+MTD	∅	∅	∅

Table B.16 – Résultat de la première phase d'évaluation pour l'anglais (UTE franches).

Modèle	Précision	Rappel	F-score
MTD+Sfreq	0.632	0.947	0.758
Sfreq+D freq	0.633	0.932	0.754
ODR+Sfreq	0.620	0.942	0.748
Sfreq+FKZY	0.621	0.933	0.745
Sfreq+ttest	0.619	0.936	0.745
PRS+Sfreq	0.618	0.934	0.744
RCT+Sfreq	0.611	0.948	0.743
ZS+Sfreq	0.612	0.943	0.743
Sfreq+SIM	0.617	0.931	0.743
TT+Sfreq	0.612	0.944	0.742
DRK+Sfreq	0.613	0.937	0.741
MD+Sfreq	0.613	0.938	0.741
Sfreq+FAG	0.611	0.942	0.741
PMI+Sfreq	0.607	0.947	0.740
USUB+Sfreq	0.609	0.936	0.738
FSS+Sfreq	0.604	0.943	0.736
Sfreq+BB	0.606	0.931	0.735
Sfreq+SAL	0.603	0.937	0.734
Sfreq	0.607	0.924	0.733
JAC+Sfreq	0.603	0.919	0.728
Sfreq+KLO	0.593	0.943	0.728
CP+Sfreq	0.591	0.944	0.727
SIM+D freq	0.600	0.880	0.714
ZS+D freq	0.569	0.940	0.709
MD+D freq	0.568	0.932	0.706
RCT+D freq	0.557	0.962	0.705
ttest+D freq	0.561	0.935	0.701
MTD+D freq	0.560	0.926	0.698
SAL+D freq	0.554	0.937	0.696
PMI+D freq	0.546	0.956	0.695
TT+D freq	0.550	0.940	0.694
ODR+D freq	0.547	0.942	0.692
D freq+FKZY	0.552	0.923	0.691
PRS+D freq	0.547	0.936	0.690
D freq+KLO	0.533	0.939	0.680
CP+D freq	0.530	0.944	0.679
D freq+FAG	0.527	0.941	0.675
MD+ttest	0.525	0.944	0.675
USUB+D freq	0.618	0.738	0.673
MTD+ttest	0.520	0.947	0.671
BB+D freq	0.525	0.929	0.671
ZS+ttest	0.515	0.960	0.670
PMI+ttest	0.545	0.864	0.668
RCT+ttest	0.508	0.970	0.667
DRK+ttest	0.543	0.863	0.667
ZS	0.525	0.906	0.664
ODR+ttest	0.516	0.918	0.661
ZS+DRK	0.522	0.896	0.659
FSS+D freq	0.506	0.946	0.659
FSS+ttest	0.498	0.974	0.659
ttest	0.493	0.968	0.653
PRS+ttest	0.491	0.969	0.652
ttest+SIM	0.491	0.967	0.651
TT+ttest	0.495	0.952	0.651
ttest+KLO	0.490	0.963	0.650
ttest+FAG	0.489	0.966	0.649
ttest+SAL	0.487	0.967	0.647
ttest+FKZY	0.487	0.965	0.647
MD+ZS	0.508	0.884	0.645
USUB+SIM	0.510	0.874	0.644
ODR+DRK	0.494	0.916	0.642
RCT+MTD	0.486	0.944	0.642
USUB+ttest	0.479	0.969	0.641
ZS+MTD	0.492	0.919	0.641
MTD+FAG	0.496	0.900	0.640
JAC+ttest	0.478	0.962	0.638
DRK+D freq	0.481	0.944	0.637
ODR+JAC	0.540	0.776	0.636
ttest+BB	0.475	0.961	0.636
TT+PMI	0.478	0.950	0.636
CP+ttest	0.473	0.968	0.635
RCT+DRK	0.480	0.931	0.634
DRK+FAG	0.468	0.955	0.628
PRS+USUB	0.471	0.903	0.619
ZS+TT	0.448	0.979	0.615
ODR+MTD	0.455	0.940	0.614
PRS+FAG	0.447	0.974	0.612
PRS+FSS	0.452	0.939	0.610
TT+KLO	0.444	0.965	0.609
MTD+KLO	0.450	0.939	0.608
DRK+SAL	0.442	0.972	0.608
ZS+RCT	0.441	0.980	0.608
ZS+ODR	0.444	0.961	0.608
CP+ZS	0.439	0.985	0.607

Modèle	Précision	Rappel	F-score
TT+JAC	0.444	0.948	0.604
MD+TT	0.439	0.966	0.604
TT+SAL	0.438	0.969	0.604
RCT+TT	0.438	0.971	0.603
PMI+SAL	0.438	0.965	0.603
ODR+FAG	0.436	0.974	0.602
SIM+FAG	0.436	0.972	0.602
CP+TT	0.437	0.967	0.602
MD+ODR	0.438	0.961	0.602
TT+DRK	0.440	0.949	0.601
MD+KLO	0.437	0.964	0.601
FSS+MTD	0.434	0.975	0.600
DRK+KLO	0.437	0.959	0.600
FSS+ZS	0.433	0.979	0.600
TT+FAG	0.435	0.963	0.600
TT+SIM	0.436	0.962	0.600
MD+SIM	0.433	0.971	0.599
ODR+TT	0.434	0.966	0.599
RCT+SAL	0.432	0.974	0.599
TT+MTD	0.433	0.969	0.598
MD+RCT	0.433	0.969	0.598
USUB+FAG	0.431	0.976	0.598
ODR+SAL	0.432	0.971	0.598
MD+FAG	0.435	0.954	0.598
PRS+SAL	0.432	0.967	0.597
PRS+RCT	0.429	0.980	0.597
FSS+RCT	0.430	0.977	0.597
SAL+FAG	0.431	0.970	0.596
RCT+SIM	0.429	0.975	0.596
PRS+MTD	0.430	0.968	0.596
PMI+SIM	0.428	0.978	0.596
SAL+KLO	0.431	0.965	0.595
RCT+FAG	0.429	0.972	0.595
CP+RCT	0.428	0.976	0.595
FSS+PMI	0.429	0.971	0.595
FSS+FAG	0.429	0.969	0.595
FSS+KLO	0.429	0.971	0.595
ZS+SIM	0.428	0.973	0.595
FSS+CP	0.428	0.975	0.595
CP+SAL	0.428	0.974	0.595
RCT+KLO	0.428	0.975	0.594
FSS+TT	0.429	0.963	0.594
PRS+SIM	0.429	0.962	0.594
USUB+SAL	0.428	0.970	0.594
ODR+KLO	0.427	0.970	0.593
FSS+SAL	0.428	0.968	0.593
CP+SIM	0.426	0.976	0.593
MD+FKZY	0.427	0.971	0.593
FSS+SIM	0.426	0.973	0.593
ZS+JAC	0.425	0.978	0.593
RCT+BB	0.425	0.975	0.592
FSS+DRK	0.425	0.974	0.592
DRK+SIM	0.424	0.976	0.592
USUB+RCT	0.424	0.976	0.591
PMI+FKZY	0.424	0.976	0.591
CP+ODR	0.425	0.972	0.591
MTD+SAL	0.426	0.964	0.591
MD+JAC	0.425	0.969	0.591
ZS+SAL	0.425	0.970	0.591
CP+MD	0.424	0.970	0.590
MD+SAL	0.425	0.968	0.590
CP+KLO	0.424	0.968	0.590
PRS+TT	0.428	0.951	0.590
ZS+KLO	0.423	0.972	0.590
SIM+SAL	0.424	0.967	0.589
ZS+FKZY	0.422	0.973	0.589
TT+FKZY	0.425	0.958	0.589
MTD+SIM	0.424	0.964	0.589
USUB+JAC	0.429	0.935	0.588
PRS+MD	0.422	0.971	0.588
USUB+MD	0.422	0.969	0.588
ODR+SIM	0.422	0.969	0.588
SIM+KLO	0.423	0.967	0.588
USUB+MTD	0.422	0.972	0.588
MD+BB	0.423	0.966	0.588
FKZY+FAG	0.422	0.969	0.588
FSS+FKZY	0.422	0.970	0.588
PRS+ZS	0.421	0.969	0.587
SAL+FKZY	0.422	0.964	0.587
FSS+MD	0.422	0.962	0.587
RCT+FKZY	0.420	0.970	0.586
CP+FAG	0.421	0.963	0.586
USUB+PMI	0.420	0.969	0.586
PRS+ODR	0.419	0.972	0.586

Modèle	Précision	Rappel	F-score
ODR+RCT	0.420	0.967	0.586
TT+BB	0.422	0.956	0.585
SAL	0.421	0.963	0.585
DRK+FKZY	0.419	0.968	0.585
BB+SAL	0.420	0.963	0.585
ZS+FAG	0.418	0.970	0.585
JAC+SAL	0.419	0.965	0.584
JAC+FKZY	0.419	0.964	0.584
USUB+TT	0.420	0.956	0.584
JAC+SIM	0.418	0.970	0.584
RCT+JAC	0.418	0.970	0.584
CP+MTD	0.418	0.964	0.583
CP+FKZY	0.418	0.966	0.583
USUB+ZS	0.416	0.974	0.583
MTD+FKZY	0.418	0.962	0.583
FSS+ODR	0.416	0.967	0.582
ODR+FKZY	0.416	0.969	0.582
FKZY+KLO	0.416	0.965	0.581
PRS+FKZY	0.416	0.960	0.581
USUB+ODR	0.415	0.968	0.581
PRS+PMI	0.414	0.969	0.581
ZS+BB	0.414	0.971	0.581
PRS+KLO	0.416	0.957	0.580
BB+FAG	0.414	0.966	0.580
MTD+JAC	0.415	0.959	0.580
JAC+FAG	0.414	0.963	0.579
PMI+JAC	0.413	0.971	0.579
ODR+BB	0.412	0.964	0.578
USUB+KLO	0.412	0.966	0.578
CP+JAC	0.412	0.966	0.577
PRS+CP	0.412	0.963	0.577
CP+DRK	0.412	0.964	0.577
JAC+KLO	0.413	0.959	0.577
USUB+CP	0.410	0.968	0.576
FSS+USUB	0.410	0.967	0.576
CP+BB	0.410	0.961	0.575
USUB+DRK	0.407	0.968	0.573
PMI+BB	0.407	0.966	0.573
MTD+BB	0.409	0.958	0.573
BB+KLO	0.409	0.954	0.572
RCT	0.390	0.020	0.038
SIM	0.000	0.000	0.000
D freq	0.000	0.000	0.000
TT	0.000	0.000	0.000
USUB+FKZY	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
ODR	∅	∅	∅
CP+PMI	∅	∅	∅
DRK	∅	∅	∅
JAC+BB	∅	∅	∅
ZS+PMI	∅	∅	∅
MD+MTD	∅	∅	∅
USUB	∅	∅	∅
MTD+PMI	∅	∅	∅
PMI+KLO	∅	∅	∅
FSS+JAC	∅	∅	∅
MD+DRK	∅	∅	∅
PMI+FAG	∅	∅	∅
PMI	∅	∅	∅
PRS+JAC	∅	∅	∅
JAC+D freq	∅	∅	∅
CP	∅	∅	∅
PRS+DRK	∅	∅	∅
BB+SIM	∅	∅	∅
MD	∅	∅	∅
JAC	∅	∅	∅
RCT+PMI	∅	∅	∅
FSS+BB	∅	∅	∅
ODR+PMI	∅	∅	∅
MTD	∅	∅	∅
FSS	∅	∅	∅
DRK+BB	∅	∅	∅
BB	∅	∅	∅
SIM+FKZY	∅	∅	∅
BB+FKZY	∅	∅	∅
FAG	∅	∅	∅
MTD+DRK	∅	∅	∅
MD+PMI	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
JAC+DRK	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.17 – Résultat de la première phase d'évaluation pour le français (UTE franches).

Modèle	Précision	Rappel	F-score
MTD+Sfreq	0.685	0.972	0.803
Sfreq+D freq	0.679	0.966	0.797
PRS+Sfreq	0.678	0.967	0.797
Sfreq+ttest	0.676	0.970	0.797
ZS+Sfreq	0.671	0.968	0.793
TT+Sfreq	0.668	0.968	0.791
FSS+Sfreq	0.662	0.972	0.787
Sfreq+FKZY	0.661	0.971	0.787
Sfreq+SIM	0.662	0.965	0.786
MD+Sfreq	0.660	0.967	0.784
Sfreq+SAL	0.659	0.968	0.784
RCT+Sfreq	0.658	0.969	0.784
Sfreq+FAG	0.658	0.966	0.783
CP+Sfreq	0.653	0.970	0.781
DRK+Sfreq	0.653	0.968	0.780
Sfreq	0.657	0.957	0.779
JAC+Sfreq	0.653	0.964	0.779
PMI+Sfreq	0.648	0.972	0.778
ODR+Sfreq	0.649	0.970	0.778
Sfreq+KLO	0.648	0.966	0.776
USUB+Sfreq	0.642	0.968	0.772
Sfreq+BB	0.643	0.967	0.772
ttest+D freq	0.614	0.966	0.751
MTD+D freq	0.605	0.963	0.743
MD+D freq	0.596	0.970	0.738
FSS+ttest	0.603	0.942	0.736
ttest+FKZY	0.595	0.955	0.734
PRS+D freq	0.589	0.965	0.731
TT+D freq	0.588	0.966	0.731
RCT+D freq	0.583	0.974	0.729
FSS+D freq	0.582	0.974	0.729
ZS+D freq	0.583	0.973	0.729
SIM+D freq	0.582	0.974	0.729
BB+D freq	0.585	0.965	0.729
SAL+D freq	0.583	0.968	0.727
ODR+D freq	0.581	0.971	0.727
DRK+D freq	0.581	0.970	0.727
ttest+SIM	0.588	0.944	0.725
D freq+KLO	0.578	0.969	0.724
JAC+D freq	0.583	0.955	0.724
FSS+USUB	0.606	0.898	0.723
CP+D freq	0.575	0.972	0.723
MD+ZS	0.644	0.821	0.722
ZS+BB	0.604	0.897	0.722
D freq+FAG	0.573	0.971	0.720
ZS+MTD	0.598	0.901	0.719
D freq+FKZY	0.570	0.973	0.719
FSS+RCT	0.652	0.796	0.717
RCT+MTD	0.587	0.919	0.717
ttest+KLO	0.587	0.918	0.716
MD+ttest	0.569	0.963	0.716
RCT+SIM	0.593	0.900	0.715
PMI+D freq	0.567	0.967	0.715
ttest+BB	0.572	0.953	0.715
SIM+KLO	0.597	0.886	0.713
USUB+D freq	0.561	0.977	0.713
USUB+ZS	0.639	0.806	0.713
ODR+DRK	0.637	0.809	0.712
BB+KLO	0.598	0.880	0.712
ODR+FKZY	0.576	0.929	0.712
SIM+SAL	0.568	0.953	0.711
BB+FAG	0.594	0.886	0.711
MTD	0.586	0.902	0.710
MTD+FKZY	0.558	0.968	0.708
RCT+TT	0.583	0.901	0.708
MD+FKZY	0.582	0.904	0.708
FSS+KLO	0.593	0.878	0.708
MTD+BB	0.623	0.819	0.708
ttest+FAG	0.566	0.944	0.708
MD+TT	0.557	0.970	0.708
PRS+ttest	0.553	0.980	0.707
DRK+ttest	0.555	0.969	0.706
ODR+PMI	0.640	0.788	0.706
TT+FAG	0.563	0.945	0.706
TT+MTD	0.650	0.772	0.705
TT+KLO	0.561	0.949	0.705
MTD+ttest	0.550	0.981	0.705
FKZY+FAG	0.559	0.953	0.704
PRS+DRK	0.645	0.774	0.704
USUB+TT	0.557	0.956	0.704
TT+FKZY	0.549	0.980	0.704
TT+ttest	0.548	0.981	0.703
SIM+FAG	0.625	0.801	0.702
FSS+ZS	0.557	0.950	0.702

Modèle	Précision	Rappel	F-score
PMI+FKZY	0.573	0.903	0.701
TT+BB	0.555	0.952	0.701
TT	0.549	0.971	0.701
CP+TT	0.638	0.778	0.701
ZS+PMI	0.552	0.959	0.701
CP+ttest	0.568	0.915	0.701
PRS+TT	0.552	0.959	0.701
ttest	0.545	0.981	0.700
USUB+FKZY	0.545	0.974	0.699
ZS+RCT	0.554	0.947	0.699
CP+FKZY	0.610	0.815	0.698
USUB+SIM	0.548	0.959	0.698
FSS+FAG	0.617	0.801	0.697
RCT+ttest	0.540	0.979	0.696
PRS+PMI	0.615	0.801	0.696
FSS+SAL	0.545	0.961	0.695
BB+SIM	0.557	0.924	0.695
FSS+ODR	0.546	0.953	0.694
ZS+DRK	0.544	0.958	0.694
ODR+SIM	0.543	0.962	0.694
PRS+ODR	0.548	0.946	0.694
JAC+BB	0.609	0.805	0.693
FSS+SIM	0.540	0.964	0.692
USUB+BB	0.550	0.928	0.691
SIM+FKZY	0.539	0.962	0.690
ODR+SAL	0.535	0.975	0.690
USUB+SAL	0.535	0.970	0.690
RCT+FKZY	0.530	0.981	0.688
BB+SAL	0.535	0.964	0.688
PRS+FKZY	0.532	0.971	0.688
PRS+JAC	0.539	0.949	0.687
FKZY+KLO	0.532	0.970	0.687
PRS+USUB	0.532	0.967	0.687
ZS+ttest	0.528	0.981	0.686
TT+PMI	0.530	0.973	0.686
PMI+JAC	0.534	0.960	0.686
ODR+ttest	0.527	0.981	0.686
PMI+ttest	0.526	0.984	0.686
PRS+MD	0.528	0.977	0.685
ttest+SAL	0.527	0.980	0.685
PRS+ZS	0.529	0.971	0.685
USUB+KLO	0.641	0.735	0.685
ODR	0.535	0.952	0.685
PRS+SAL	0.523	0.986	0.684
FSS+FKZY	0.524	0.981	0.683
TT+DRK	0.527	0.971	0.683
USUB+JAC	0.542	0.924	0.683
RCT+BB	0.530	0.959	0.683
ZS+FKZY	0.525	0.976	0.683
PRS+BB	0.535	0.943	0.682
TT+SAL	0.522	0.984	0.682
FSS+TT	0.523	0.979	0.682
PMI	0.599	0.790	0.681
SAL+FKZY	0.519	0.979	0.678
SAL	0.517	0.979	0.677
PRS+MTD	0.517	0.979	0.676
FSS+BB	0.522	0.960	0.676
JAC+SIM	0.521	0.963	0.676
PRS+SIM	0.517	0.975	0.676
RCT	0.516	0.971	0.674
JAC+DRK	0.518	0.963	0.674
JAC+ttest	0.514	0.977	0.674
PRS+RCT	0.512	0.981	0.673
TT+SIM	0.512	0.978	0.672
PRS+KLO	0.511	0.979	0.671
ODR+TT	0.511	0.979	0.671
MTD+JAC	0.509	0.982	0.670
ZS+SIM	0.510	0.975	0.670
PRS+FSS	0.509	0.977	0.669
ZS+TT	0.508	0.980	0.669
PRS+FAG	0.507	0.979	0.668
PRS+CP	0.507	0.980	0.668
ZS	0.507	0.975	0.667
FSS+JAC	0.509	0.967	0.667
TT+JAC	0.505	0.978	0.666
JAC+FKZY	0.506	0.973	0.666
ZS+SAL	0.502	0.980	0.664
JAC+SAL	0.500	0.981	0.662
ZS+ODR	0.500	0.981	0.662
JAC+KLO	0.499	0.979	0.661
ZS+JAC	0.497	0.983	0.660
ODR+JAC	0.498	0.979	0.660
RCT+JAC	0.497	0.978	0.659
MD+JAC	0.490	0.980	0.654

Modèle	Précision	Rappel	F-score
JAC+FAG	0.491	0.978	0.653
CP+JAC	0.488	0.972	0.649
MD	0.706	0.061	0.113
DRK	0.769	0.055	0.103
RCT+SAL	0.606	0.047	0.087
D freq	0.785	0.019	0.037
MTD+SAL	0.606	0.019	0.037
PMI+SAL	0.703	0.018	0.036
MD+SAL	0.624	0.012	0.024
ZS+FAG	0.632	0.009	0.018
ZS+KLO	0.664	0.009	0.018
FAG	0.598	0.008	0.016
KLO	0.554	0.007	0.014
USUB+ttest	0.514	0.006	0.012
DRK+SAL	0.575	0.005	0.010
CP+ZS	0.587	0.005	0.010
SAL+KLO	0.475	0.004	0.008
SAL+FAG	0.558	0.003	0.007
ODR+KLO	0.792	0.003	0.007
RCT+FAG	0.492	0.002	0.005
SIM	0.495	0.002	0.004
RCT+KLO	0.563	0.002	0.004
RCT+DRK	0.698	0.002	0.004
CP+SAL	0.446	0.001	0.003
RCT+PMI	0.700	0.001	0.002
MD+RCT	0.700	0.001	0.002
DRK+FAG	0.745	0.001	0.002
MD+FAG	0.714	0.001	0.001
DRK+KLO	0.708	0.001	0.001
PMI+KLO	0.659	0.001	0.001
MD+KLO	0.426	0.001	0.001
ODR+RCT	0.750	0.000	0.001
ODR+MTD	0.500	0.000	0.001
CP	0.481	0.000	0.001
PMI+FAG	0.500	0.000	0.001
MTD+PMI	0.666	0.000	0.001
PMI+DRK	1.000	0.000	0.000
CP+PMI	1.000	0.000	0.000
FSS	0.300	0.000	0.000
CP+RCT	0.000	0.000	0.000
MTD+KLO	0.000	0.000	0.000
MTD+DRK	0.000	0.000	0.000
FSS+CP	∅	∅	∅
ODR+FAG	∅	∅	∅
USUB+FAG	∅	∅	∅
USUB+MTD	∅	∅	∅
PMI+BB	∅	∅	∅
MD+ODR	∅	∅	∅
USUB+MD	∅	∅	∅
USUB+RCT	∅	∅	∅
FSS+MTD	∅	∅	∅
USUB+PMI	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
USUB+ODR	∅	∅	∅
MD+MTD	∅	∅	∅
FSS+PMI	∅	∅	∅
USUB	∅	∅	∅
CP+DRK	∅	∅	∅
MTD+SIM	∅	∅	∅
DRK+FKZY	∅	∅	∅
CP+BB	∅	∅	∅
MD+DRK	∅	∅	∅
FSS+MD	∅	∅	∅
PMI+SIM	∅	∅	∅
JAC	∅	∅	∅
CP+MD	∅	∅	∅
CP+KLO	∅	∅	∅
FSS+DRK	∅	∅	∅
DRK+SIM	∅	∅	∅
USUB+CP	∅	∅	∅
DRK+BB	∅	∅	∅
BB	∅	∅	∅
USUB+DRK	∅	∅	∅
BB+FKZY	∅	∅	∅
CP+MTD	∅	∅	∅
CP+ODR	∅	∅	∅
CP+SIM	∅	∅	∅
ODR+BB	∅	∅	∅
MD+PMI	∅	∅	∅
MD+BB	∅	∅	∅
PRS	∅	∅	∅
MD+SIM	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.18 – Résultat de la première phase d'évaluation pour le polonais (UTE franches).

Modèle	Précision	Rappel	F-score
RCT+Sfreq	0.676	0.955	0.791
PRS+Sfreq	0.671	0.960	0.790
Sfreq+KLO	0.671	0.957	0.789
Sfreq+SIM	0.663	0.972	0.789
ODR+Sfreq	0.668	0.959	0.788
JAC+Sfreq	0.667	0.957	0.787
PMI+Sfreq	0.666	0.959	0.786
Sfreq+FAG	0.664	0.962	0.786
Sfreq+ttest	0.668	0.951	0.785
Sfreq+SAL	0.663	0.961	0.785
USUB+Sfreq	0.664	0.959	0.784
Sfreq+FKZY	0.662	0.962	0.784
Sfreq	0.662	0.958	0.783
MTD+Sfreq	0.659	0.961	0.782
TT+Sfreq	0.658	0.962	0.782
ZS+Sfreq	0.655	0.965	0.781
Sfreq+BB	0.656	0.963	0.780
FSS+Sfreq	0.656	0.961	0.779
DRK+Sfreq	0.652	0.966	0.779
Sfreq+D freq	0.651	0.969	0.778
MD+Sfreq	0.648	0.964	0.775
CP+Sfreq	0.641	0.969	0.771
ttest+D freq	0.544	0.963	0.695
RCT+D freq	0.567	0.891	0.693
SIM+D freq	0.584	0.851	0.692
D freq+FAG	0.559	0.897	0.689
ZS+D freq	0.555	0.905	0.688
MTD+D freq	0.553	0.904	0.686
MD+D freq	0.555	0.896	0.686
SAL+D freq	0.547	0.916	0.685
BB+D freq	0.575	0.842	0.684
CP+D freq	0.541	0.923	0.682
DRK+D freq	0.566	0.848	0.679
PMI+D freq	0.567	0.840	0.677
TT+D freq	0.558	0.852	0.674
D freq+KLO	0.555	0.849	0.671
PRS+D freq	0.571	0.811	0.670
TT+ttest	0.499	0.972	0.659
FSS+D freq	0.577	0.740	0.648
ttest+SIM	0.484	0.979	0.648
DRK+ttest	0.482	0.978	0.645
D freq	0.598	0.698	0.644
ODR+D freq	0.582	0.720	0.644
ttest+KLO	0.479	0.979	0.643
RCT+ttest	0.480	0.973	0.643
PMI+ttest	0.479	0.971	0.642
ODR+ttest	0.478	0.975	0.641
TT+SIM	0.488	0.932	0.641
MTD+ttest	0.478	0.973	0.641
ttest	0.476	0.977	0.640
PRS+ttest	0.479	0.964	0.640
USUB+TT	0.497	0.895	0.639
ZS+ttest	0.476	0.973	0.639
JAC+ttest	0.476	0.971	0.638
FSS+ttest	0.474	0.976	0.638
USUB+ttest	0.473	0.975	0.637
ttest+FKZY	0.473	0.970	0.636
CP+ttest	0.470	0.976	0.635
ttest+BB	0.470	0.974	0.635
ttest+SAL	0.471	0.971	0.634
PRS+FAG	0.506	0.845	0.633
ttest+FAG	0.463	0.976	0.629
PRS+TT	0.521	0.788	0.627
TT	0.509	0.815	0.627
ODR+TT	0.463	0.964	0.626
MD+ttest	0.459	0.980	0.626
PRS+ZS	0.486	0.874	0.625
FSS+TT	0.467	0.942	0.624
D freq+FKZY	0.585	0.668	0.624
DRK+SIM	0.472	0.916	0.623
JAC+D freq	0.588	0.661	0.622
TT+BB	0.496	0.830	0.621
ZS+JAC	0.477	0.888	0.620
TT+FKZY	0.497	0.816	0.618
ZS	0.474	0.885	0.618
ZS+BB	0.476	0.879	0.617
TT+JAC	0.500	0.806	0.617
RCT+JAC	0.480	0.863	0.617
PRS+RCT	0.481	0.857	0.616
ZS+FKZY	0.477	0.868	0.616
PRS+MD	0.475	0.870	0.614
FKZY+FAG	0.474	0.870	0.614
SIM+KLO	0.451	0.958	0.614
USUB+DRK	0.476	0.863	0.613
SIM+FAG	0.449	0.960	0.612
RCT+BB	0.470	0.875	0.612
PRS+MTD	0.470	0.876	0.611
TT+PMI	0.446	0.972	0.611
PRS+SAL	0.487	0.820	0.611
PMI+SIM	0.448	0.960	0.610
USUB+SIM	0.502	0.778	0.610
FKZY+KLO	0.472	0.863	0.610
MTD+JAC	0.467	0.879	0.610
BB+FAG	0.471	0.866	0.610
SAL+FKZY	0.461	0.902	0.610
SAL	0.455	0.924	0.610
PRS+CP	0.465	0.883	0.610
BB+SAL	0.460	0.901	0.609
USUB+PMI	0.469	0.869	0.609
FSS+DRK	0.450	0.943	0.609
ODR+DRK	0.451	0.939	0.609
MTD	0.453	0.930	0.609
MTD+FKZY	0.463	0.889	0.609
PMI+DRK	0.449	0.944	0.608
PRS+DRK	0.487	0.807	0.608
MD+BB	0.464	0.880	0.608
PRS+PMI	0.512	0.742	0.606
FSS+PMI	0.445	0.947	0.606
USUB+KLO	0.451	0.921	0.606
MTD+BB	0.457	0.898	0.606
MD+FKZY	0.462	0.873	0.605
RCT+SIM	0.439	0.971	0.604
MD+SIM	0.437	0.971	0.603
CP+JAC	0.463	0.861	0.602
FSS+RCT	0.437	0.963	0.602
CP	0.455	0.888	0.601
ZS+SIM	0.435	0.975	0.601
JAC+SAL	0.481	0.800	0.601
TT+KLO	0.434	0.977	0.601
ODR+SIM	0.497	0.760	0.601
ODR+FAG	0.436	0.962	0.600
ODR+RCT	0.435	0.968	0.600
TT+FAG	0.433	0.975	0.600
DRK+BB	0.490	0.771	0.599
FSS+SIM	0.496	0.757	0.599
JAC+DRK	0.488	0.777	0.599
ODR+KLO	0.434	0.964	0.599
PRS+KLO	0.484	0.780	0.598
DRK+FKZY	0.485	0.779	0.597
TT+DRK	0.430	0.975	0.597
FSS+KLO	0.433	0.962	0.597
ODR+PMI	0.434	0.950	0.596
RCT+TT	0.429	0.979	0.596
USUB+SAL	0.431	0.964	0.595
RCT+FKZY	0.482	0.775	0.595
JAC+FAG	0.489	0.758	0.594
FSS+USUB	0.488	0.760	0.594
RCT+DRK	0.428	0.972	0.594
PMI+KLO	0.428	0.969	0.594
CP+BB	0.471	0.804	0.594
SIM+SAL	0.426	0.980	0.594
ODR+FKZY	0.479	0.778	0.593
ZS+TT	0.424	0.980	0.592
PMI+BB	0.483	0.761	0.591
FSS+FKZY	0.475	0.783	0.591
MD+ODR	0.425	0.972	0.591
USUB+ODR	0.487	0.752	0.591
DRK+KLO	0.426	0.967	0.591
FSS+ODR	0.483	0.763	0.591
RCT+PMI	0.425	0.972	0.591
FSS+MD	0.426	0.964	0.591
ODR+SAL	0.422	0.981	0.590
ZS+RCT	0.422	0.979	0.590
PMI+JAC	0.489	0.743	0.590
FSS+FAG	0.424	0.967	0.590
USUB+RCT	0.423	0.971	0.589
JAC+KLO	0.473	0.781	0.589
USUB+FAG	0.422	0.974	0.589
TT+SAL	0.421	0.979	0.589
MD+DRK	0.423	0.969	0.589
MD+TT	0.421	0.978	0.588
CP+FKZY	0.434	0.911	0.588
BB+KLO	0.479	0.760	0.588
PMI+FKZY	0.488	0.740	0.588
CP+SIM	0.421	0.977	0.588
DRK+FAG	0.420	0.974	0.587
RCT+KLO	0.420	0.977	0.587
CP+TT	0.419	0.983	0.587
FSS+SAL	0.419	0.978	0.587
RCT+FAG	0.419	0.977	0.587
TT+MTD	0.417	0.982	0.585
USUB+CP	0.419	0.965	0.584
FSS+ZS	0.416	0.978	0.584
SAL+FAG	0.416	0.977	0.584
ZS+DRK	0.416	0.974	0.583
PMI+FAG	0.417	0.971	0.583
USUB+ZS	0.415	0.980	0.583
DRK+SAL	0.416	0.976	0.583
ZS+ODR	0.415	0.979	0.583
MTD+SIM	0.415	0.979	0.582
RCT+SAL	0.414	0.982	0.582
MTD+PMI	0.414	0.980	0.582
ZS+KLO	0.413	0.983	0.581
PMI+SAL	0.413	0.978	0.581
ZS+SAL	0.412	0.982	0.581
USUB+MD	0.415	0.966	0.581
MTD+FAG	0.414	0.974	0.581
ZS+FAG	0.412	0.982	0.580
CP+DRK	0.412	0.981	0.580
RCT+MTD	0.411	0.981	0.580
SAL+KLO	0.412	0.978	0.579
ZS+PMI	0.411	0.978	0.579
MTD+KLO	0.410	0.979	0.578
ZS+MTD	0.408	0.986	0.578
MD+RCT	0.409	0.981	0.577
ODR+MTD	0.409	0.980	0.577
USUB+MTD	0.409	0.982	0.577
MTD+DRK	0.409	0.981	0.577
FSS+CP	0.409	0.975	0.577
MD+KLO	0.409	0.976	0.576
MD+PMI	0.409	0.974	0.576
MD+FAG	0.409	0.974	0.576
MD+ZS	0.407	0.983	0.576
FSS+MTD	0.406	0.980	0.574
CP+KLO	0.406	0.980	0.574
MTD+SAL	0.406	0.979	0.574
CP+RCT	0.405	0.981	0.573
CP+ODR	0.404	0.982	0.573
PRS+ODR	0.521	0.634	0.572
CP+PMI	0.403	0.980	0.572
MD+SAL	0.403	0.979	0.571
CP+ZS	0.401	0.982	0.569
MD+JAC	0.507	0.644	0.568
CP+FAG	0.400	0.977	0.567
MD+MTD	0.399	0.978	0.566
CP+SAL	0.398	0.983	0.566
CP+MTD	0.397	0.982	0.565
CP+MD	0.395	0.983	0.563
SIM+FKZY	0.528	0.597	0.560
PRS+FSS	0.520	0.596	0.555
ODR+BB	0.518	0.575	0.545
FSS+BB	0.515	0.570	0.541
PRS+BB	0.532	0.515	0.523
JAC+BB	0.551	0.458	0.500
SIM	0.444	0.003	0.007
MD	0.916	0.003	0.006
RCT	0.588	0.003	0.005
KLO	0.889	0.002	0.004
ODR	1.000	0.001	0.002
FAG	0.600	0.001	0.002
DRK	1.000	0.001	0.001
FSS	1.000	0.000	0.001
PRS+SIM	0.000	0.000	0.000
USUB+FKZY	∅	∅	∅
JAC+SIM	∅	∅	∅
USUB+BB	∅	∅	∅
USUB	∅	∅	∅
FSS+JAC	∅	∅	∅
PMI	∅	∅	∅
PRS+JAC	∅	∅	∅
BB+SIM	∅	∅	∅
JAC	∅	∅	∅
USUB+D freq	∅	∅	∅
BB	∅	∅	∅
USUB+JAC	∅	∅	∅
BB+FKZY	∅	∅	∅
PRS+FKZY	∅	∅	∅
PRS+USUB	∅	∅	∅
ODR+JAC	∅	∅	∅
PRS	∅	∅	∅
JAC+FKZY	∅	∅	∅
FKZY	∅	∅	∅

TABLE B.19 – Résultat de la première phase d'évaluation pour le turc (UTE franches).

Modèle	Précision	Rappel	F-score
DRK+Sfreq	0.887	0.974	0.928
MD+Sfreq	0.882	0.974	0.926
Sfreq+FAG	0.879	0.975	0.925
Sfreq+KLO	0.878	0.975	0.924
Sfreq+SIM	0.875	0.977	0.923
ZS+Sfreq	0.876	0.976	0.923
Sfreq	0.880	0.970	0.923
PMI+Sfreq	0.875	0.977	0.923
ODR+Sfreq	0.874	0.976	0.922
FSS+Sfreq	0.874	0.975	0.922
USUB+Sfreq	0.872	0.976	0.921
RCT+Sfreq	0.871	0.977	0.921
Sfreq+BB	0.869	0.976	0.920
Sfreq+FKZY	0.868	0.978	0.920
JAC+Sfreq	0.866	0.977	0.918
PRS+Sfreq	0.863	0.980	0.918
Sfreq+SAL	0.864	0.977	0.917
Sfreq+D freq	0.863	0.978	0.917
TT+Sfreq	0.861	0.978	0.916
MTD+Sfreq	0.861	0.978	0.916
Sfreq+ttest	0.860	0.974	0.913
CP+Sfreq	0.847	0.981	0.909
ttest+D freq	0.691	0.892	0.779
ttest+BB	0.650	0.935	0.767
ttest+FAG	0.638	0.934	0.758
RCT+ttest	0.625	0.959	0.757
ttest+SIM	0.623	0.955	0.754
USUB+ttest	0.623	0.955	0.754
ODR+ttest	0.614	0.977	0.754
TT+ttest	0.613	0.977	0.753
ttest+FKZY	0.621	0.955	0.753
PMI+ttest	0.612	0.971	0.751
FSS+ttest	0.606	0.982	0.749
ttest+KLO	0.609	0.972	0.749
DRK+ttest	0.601	0.985	0.746
PRS+ttest	0.647	0.866	0.741
ttest+SAL	0.590	0.988	0.739
ZS+ttest	0.590	0.988	0.739
CP+ttest	0.624	0.897	0.736
MD+ttest	0.595	0.963	0.736
MTD+ttest	0.584	0.989	0.735
JAC+D freq	0.653	0.699	0.675
PRS+D freq	0.663	0.663	0.663
PRS+JAC	0.605	0.732	0.663
ttest	0.777	0.043	0.081
JAC+BB	0.000	0.000	0.000
USUB+FKZY	∅	∅	∅
ODR+D freq	∅	∅	∅
FSS+CP	∅	∅	∅
ODR+FAG	∅	∅	∅
RCT+DRK	∅	∅	∅
JAC+SIM	∅	∅	∅
PMI+SAL	∅	∅	∅
USUB+FAG	∅	∅	∅
RCT+FAG	∅	∅	∅
DRK+D freq	∅	∅	∅
SIM	∅	∅	∅
USUB+MTD	∅	∅	∅
ZS+FAG	∅	∅	∅
PMI+DRK	∅	∅	∅
PRS+BB	∅	∅	∅
USUB+BB	∅	∅	∅
CP+TT	∅	∅	∅
SAL+KLO	∅	∅	∅
PMI+BB	∅	∅	∅
ZS+JAC	∅	∅	∅
RCT+D freq	∅	∅	∅
FKZY+KLO	∅	∅	∅
JAC+KLO	∅	∅	∅
ODR	∅	∅	∅
MD+ODR	∅	∅	∅
ODR+TT	∅	∅	∅
FSS+KLO	∅	∅	∅
USUB+MD	∅	∅	∅
BB+D freq	∅	∅	∅
CP+RCT	∅	∅	∅
ZS+BB	∅	∅	∅
USUB+TT	∅	∅	∅
D freq	∅	∅	∅
PRS+PMI	∅	∅	∅
MTD+FKZY	∅	∅	∅
USUB+RCT	∅	∅	∅
MTD+KLO	∅	∅	∅
FSS+MTD	∅	∅	∅
DRK+KLO	∅	∅	∅
BB+KLO	∅	∅	∅
FSS+FKZY	∅	∅	∅
ZS+ODR	∅	∅	∅
CP+FKZY	∅	∅	∅
CP+SAL	∅	∅	∅
SIM+FAG	∅	∅	∅
FSS+RCT	∅	∅	∅
FKZY+FAG	∅	∅	∅
SIM+D freq	∅	∅	∅
MD+D freq	∅	∅	∅
CP+PMI	∅	∅	∅
ODR+RCT	∅	∅	∅
DRK+FAG	∅	∅	∅
DRK	∅	∅	∅
FSS+D freq	∅	∅	∅
USUB+PMI	∅	∅	∅
ZS+SIM	∅	∅	∅
FSS+SAL	∅	∅	∅
ZS+SAL	∅	∅	∅
JAC+SAL	∅	∅	∅
RCT+FKZY	∅	∅	∅
MTD+FAG	∅	∅	∅
CP+FAG	∅	∅	∅
USUB+ODR	∅	∅	∅
ZS+FKZY	∅	∅	∅
TT+FAG	∅	∅	∅
PMI+JAC	∅	∅	∅
RCT+KLO	∅	∅	∅
ZS+D freq	∅	∅	∅
ZS+PMI	∅	∅	∅
USUB+SIM	∅	∅	∅
MD+MTD	∅	∅	∅
TT+JAC	∅	∅	∅
JAC+FAG	∅	∅	∅
FSS+PMI	∅	∅	∅
RCT	∅	∅	∅
FSS+ZS	∅	∅	∅
ODR+KLO	∅	∅	∅
USUB	∅	∅	∅
CP+DRK	∅	∅	∅
USUB+SAL	∅	∅	∅
ODR+DRK	∅	∅	∅
MD+FAG	∅	∅	∅
ZS	∅	∅	∅
MD+TT	∅	∅	∅
MTD+SIM	∅	∅	∅
MTD+PMI	∅	∅	∅
PMI+KLO	∅	∅	∅
DRK+FKZY	∅	∅	∅
ZS+RCT	∅	∅	∅
CP+BB	∅	∅	∅
FSS+JAC	∅	∅	∅
MD+DRK	∅	∅	∅
PRS+SAL	∅	∅	∅
FSS+TT	∅	∅	∅
PMI+FAG	∅	∅	∅
TT+FKZY	∅	∅	∅
USUB+KLO	∅	∅	∅
BB+FAG	∅	∅	∅
BB+SAL	∅	∅	∅
PMI	∅	∅	∅
SAL+FAG	∅	∅	∅
PRS+SIM	∅	∅	∅
TT+SIM	∅	∅	∅
CP+ZS	∅	∅	∅
RCT+JAC	∅	∅	∅
FSS+MD	∅	∅	∅
TT+PMI	∅	∅	∅
CP	∅	∅	∅
PRS+DRK	∅	∅	∅
TT	∅	∅	∅
MD+JAC	∅	∅	∅
CP+JAC	∅	∅	∅
BB+SIM	∅	∅	∅
MD	∅	∅	∅
SAL+FKZY	∅	∅	∅
ZS+DRK	∅	∅	∅
TT+KLO	∅	∅	∅
PMI+SIM	∅	∅	∅
JAC	∅	∅	∅
FSS+ODR	∅	∅	∅
CP+MD	∅	∅	∅
JAC+ttest	∅	∅	∅
RCT+PMI	∅	∅	∅
SAL	∅	∅	∅
FSS+BB	∅	∅	∅
PRS+ZS	∅	∅	∅
SIM+SAL	∅	∅	∅
ODR+SIM	∅	∅	∅
CP+KLO	∅	∅	∅
SIM+KLO	∅	∅	∅
USUB+D freq	∅	∅	∅
RCT+SIM	∅	∅	∅
PMI+FKZY	∅	∅	∅
ODR+PMI	∅	∅	∅
FSS+DRK	∅	∅	∅
TT+D freq	∅	∅	∅
MTD	∅	∅	∅
DRK+SIM	∅	∅	∅
PRS+KLO	∅	∅	∅
SAL+D freq	∅	∅	∅
FSS+SIM	∅	∅	∅
FSS	∅	∅	∅
ZS+TT	∅	∅	∅
USUB+CP	∅	∅	∅
FSS+USUB	∅	∅	∅
DRK+BB	∅	∅	∅
ZS+KLO	∅	∅	∅
PRS+FSS	∅	∅	∅
BB	∅	∅	∅
USUB+JAC	∅	∅	∅
D freq+FAG	∅	∅	∅
PRS+ODR	∅	∅	∅
TT+BB	∅	∅	∅
CP+D freq	∅	∅	∅
PRS+FAG	∅	∅	∅
SIM+FKZY	∅	∅	∅
USUB+DRK	∅	∅	∅
PRS+MD	∅	∅	∅
BB+FKZY	∅	∅	∅
CP+MTD	∅	∅	∅
PRS+FKZY	∅	∅	∅
MTD+D freq	∅	∅	∅
MD+SAL	∅	∅	∅
D freq+FKZY	∅	∅	∅
PRS+MTD	∅	∅	∅
TT+SAL	∅	∅	∅
ODR+MTD	∅	∅	∅
MD+ZS	∅	∅	∅
PMI+D freq	∅	∅	∅
RCT+MTD	∅	∅	∅
TT+DRK	∅	∅	∅
MD+RCT	∅	∅	∅
PRS+USUB	∅	∅	∅
RCT+TT	∅	∅	∅
CP+ODR	∅	∅	∅
MD+KLO	∅	∅	∅
CP+SIM	∅	∅	∅
PRS+RCT	∅	∅	∅
ODR+JAC	∅	∅	∅
FAG	∅	∅	∅
ZS+MTD	∅	∅	∅
PRS+CP	∅	∅	∅
MTD+SAL	∅	∅	∅
RCT+BB	∅	∅	∅
MTD+DRK	∅	∅	∅
ODR+FKZY	∅	∅	∅
MTD+JAC	∅	∅	∅
FSS+FAG	∅	∅	∅
ODR+BB	∅	∅	∅
MD+PMI	∅	∅	∅
MD+BB	∅	∅	∅
PRS	∅	∅	∅
KLO	∅	∅	∅
JAC+FKZY	∅	∅	∅
RCT+SAL	∅	∅	∅
DRK+SAL	∅	∅	∅
MD+FKZY	∅	∅	∅
JAC+DRK	∅	∅	∅
MTD+BB	∅	∅	∅
ODR+SAL	∅	∅	∅
D freq+KLO	∅	∅	∅
MD+SIM	∅	∅	∅
PRS+TT	∅	∅	∅
FKZY	∅	∅	∅
USUB+ZS	∅	∅	∅
TT+MTD	∅	∅	∅

Annexe C

F-scores des expériences sur la portabilité des modèles entre les langues

Langue support : l'arabe						
	Modèles calques			*Modèles contre-épreuves*		
langue image	Tokens	EPU mod.	EPU fra.	Tokens	EPU mod.	EPU fra.
deu	0.902	0.741	0.765	0.937	0.732	0.743
eng	0.945	0.837	0.791	0.870	0.866	0.793
fra	0.871	0.802	0.759	0.934	0.785	0.774
pol	0.903	0.798	0.798	0.931	0.806	0.821
tur	0.951	0.843	0.841	0.961	0.839	0.873
zho	0.937	0.898	0.884	0.000	0.898	0.877

Langue support : l'allemand

	Modèles calques			*Modèles contre-épreuves*		
langue image	Tokens	EPU mod.	EPU fra.	Tokens	EPU mod.	EPU fra.
ara	0.964	0.746	0.729	0.963	0.798	0.000
eng	0.961	0.721	0.672	0.867	0.720	0.694
fra	0.827	0.721	0.667	0.898	0.703	0.661
pol	0.952	0.788	0.701	0.944	0.760	0.725
tur	0.959	0.812	0.735	0.966	0.820	0.741
zho	0.919	0.818	0.718	0.000	0.836	0.000

Langue support : l'anglais

	Modèles calques			*Modèles contre-épreuves*		
langue image	Tokens	EPU mod.	EPU fra.	Tokens	EPU mod.	EPU fra.
ara	0.879	0.835	0.792	0.877	0.856	0.000
deu	0.892	0.808	0.749	0.885	0.770	0.732
fra	0.864	0.795	0.763	0.893	0.796	0.783
pol	0.902	0.831	0.796	0.902	0.819	0.786
tur	0.896	0.833	0.805	0.879	0.000	0.000
zho	0.884	0.848	0.754	0.000	0.842	0.000

Langue support : le français

	Modèles calques			*Modèles contre-épreuves*		
langue image	Tokens	EPU mod.	EPU fra.	Tokens	EPU mod.	EPU fra.
ara	0.915	0.856	0.835	0.913	0.849	0.000
deu	0.927	0.855	0.794	0.898	0.839	0.777
eng	0.917	0.845	0.811	0.913	0.859	0.823
pol	0.934	0.861	0.832	0.907	0.865	0.820
tur	0.930	0.872	0.847	0.911	0.874	0.000
zho	0.925	0.844	0.780	0.000	0.000	0.000

Langue support : le polonais

	Modèles calques			*Modèles contre-épreuves*		
langue image	Tokens	EPU mod.	EPU fra.	Tokens	EPU mod.	EPU fra.
ara	0.951	0.771	0.794	0.949	0.000	0.000
deu	0.945	0.769	0.727	0.945	0.750	0.679
eng	0.949	0.737	0.744	0.949	0.748	0.742
fra	0.870	0.772	0.737	0.968	0.783	0.723
tur	0.965	0.819	0.808	0.951	0.000	0.000
zho	0.934	0.849	0.820	0.000	0.000	0.000

Langue support : le turc

	Modèles calques			*Modèles contre-épreuves*		
langue image	Tokens	EPU mod.	EPU fra.	Tokens	EPU mod.	EPU fra.
ara	0.957	0.883	0.894	0.956	0.872	0.864
deu	0.941	0.789	0.779	0.936	0.787	0.742
eng	0.951	0.816	0.820	0.933	0.776	0.776
fra	0.873	0.832	0.815	0.957	0.804	0.750
pol	0.958	0.859	0.856	0.953	0.867	0.845
zho	0.936	0.939	0.940	0.962	0.938	0.938

Langue support : le chinois

	Modèles calques			*Modèles contre-épreuves*		
langue image	Tokens	EPU mod.	EPU fra.	Tokens	EPU mod.	EPU fra.
ara	0.567	0.650	0.597	0.823	0.677	0.655
deu	0.766	0.443	0.404	0.782	0.414	0.372
eng	0.573	0.459	0.532	0.600	0.570	0.595
fra	0.672	0.561	0.561	0.692	0.706	0.570
pol	0.673	0.455	0.526	0.557	0.475	0.438
tur	0.695	0.518	0.587	0.562	0.740	0.626

Annexe D

Extraction automatique de Traductions à partir de Wiktionnaires

Il est possible de paramétrer un ensemble de règles simples spécifiques à une langue pour l'identification et l'extraction de traduction et de synonymes à partir de pages de Wiktionnaires. Notre procédé d'extraction repose en premier lieu sur des « déclencheurs » à même d'identifier des zones d'intérêt dans les pages contenant nos liens de traduction et de synonymie. La figure D.1 présente l'ensemble des déclencheurs nécessaires à l'extraction, ainsi que les actions qu'ils provoquent.

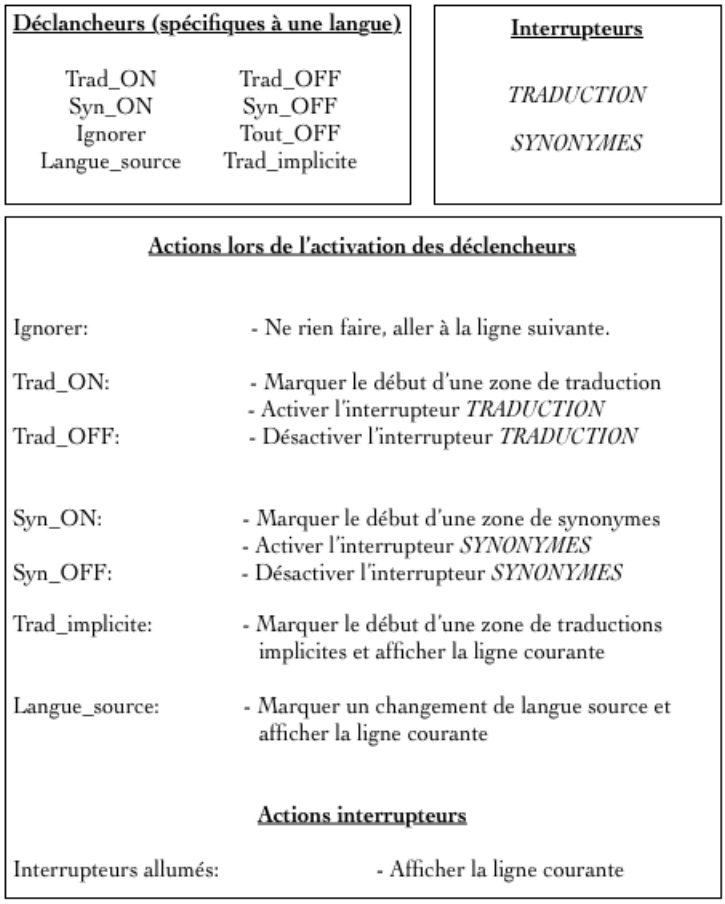
Ces déclencheurs sont instanciés par des expressions régulières. Par exemple, les déclencheurs de l'édition anglaise sont :

- Trad_ON : {{trans-top.*
- Trad_OFF : {{trans-bottom}}
- Syn_ON : ==*Synonyms==*
- Syn_OFF : {{checksyns}}
- Tout_OFF : ^\\s*\$
- Ignorer : {{trans-[^tb].+}}
- Langue_source : ^[^>]*>?==([^=]+)==\$
- Trad_implicite : ^#\s*\[\[.+\]\].?$

Pour l'édition hindi, nous avons instancié ces déclencheurs comme suit :

- Trad_ON :
  ^\({{-trans-}}\)\|\(*==* *अन्य भाषाओं मे*==*\)\|\(*==* *अनुवाद *==*\)

FIGURE D.1 – Ensemble de déclencheurs (*triggers*) dépendants de la langue, interrupteurs (*switches*) et effets de bord nécessaires à l'extraction de traductions et de synonymes depuis des éditions de wiktionnaires.



— Trad_OFF : ^\(*{{-.*\)\|\(*==\)\|\(^\[\[\)
— Syn_ON : ^ *==* * समानार्थी ! *==*
— Syn_OFF : ^\[\[.*
— Tout_OFF : ^=
— Ignorer :
— Langue_source : <text xml:space=\"preserve\">{?{?-?(.[^-]*)-?}?}?
— Trad_implicite : ^ *#\s*\(([^)]*)\)\s*\[\[.*

Utiliser ces indicateurs personnalisés pour chaque article de wiktionnaires en tenant compte de l'édition de langue permet d'extraire des données brutes concernant les traductions et synonymes. La figure D.2 présente un exemple de données brutes extraites de l'édition hindi à l'aide de l'ensemble de déclencheurs spécifiques présentés ci-dessus.

FIGURE D.2 – Fragment d'une extraction brute de traductions à partir de l'édition hindi de wiktionnaire.

```
#ITEM# आसमान          #LANG# == हिन्दी==
 #IT# # [[ अम्बर]]
 #IT# # [[ आकाश]]
 #IT# # [[ गगन]]
 #IT# # [[ नभ]]
 #TRANS#
 * {{en}} : [[sky]]  [[:en:sky]]
 * {{gu}} : [[ આકાશ]] [[:gu: આકાશ]]
 * {{bn}} : [[ আকাশ]] [[:bn: আকাশ]]
 * {{es}} : [[firmamento]] [[:es:firmamento]]
 * {{fr}} : [[ciel]] [[:fr:ciel]]
 * {{fa}} : [[  آسمان ]] [[:fa:  آسمان ]]
```

Cette extraction brute est par la suite nettoyée : tout d'abord, les caractères indésirables (*[*, *{*, *:* ...) sont supprimés, puis les noms de langues sont normalisés en utilisant les codes ISO 639-2 (alpha-3) [1]. Par exemple, « हिन्दी » [2] devient « hin ». Enfin, le tout est mis dans un format n'acceptant qu'un(e) traduction ou synonyme par ligne avec l'origine de la traduction (cf. tableau D.1).

TABLE D.1 – Fragment normalisé d'information extrait de l'édition hindi du wiktionnaire.

Langue source		Langue cible		Wiktionnaire
Lang.	Terme	Lang.	Terme	édition (lang.)
hin	आसमान	hin	अम्बर	hi
hin	आसमान	hin	आकाश	hi
hin	आसमान	hin	गगन	hi
hin	आसमान	hin	नभ	hi
hin	आसमान	eng	sky	hi
hin	आसमान	guj	આકાશ	hi
hin	आसमान	ben	আকাশ	hi
hin	आसमान	spa	firmamento	hi
hin	आसमान	fra	ciel	hi
hin	आसमान	fas	آسمان	hi

1. Liste disponible à l'adresse http://www.loc.gov/standards/iso639-2/php/code_list.php.
2. हिन्दीest le terme hindi désignant la langue hindi.

D.0.0.1 Évaluation du processus d'extraction

Le tableau D.2 met en correspondance le nombre d'articles de wiktionnaires (total et contenant une information potentiellement exploitable), le nombre d'articles pour lesquels nous avons extrait une information (au moins une traduction ou un synonyme), et la proportion correspondante d'articles concernés pour les 21 éditions utilisées pour la construction de YAMTG 1.0.

TABLE D.2 – Proportion des articles de wiktionnaire dont au moins une traduction ou un synonyme a été extrait par notre système.

Éd. Wikt. (lang)	#articles	#articles articles	#non-vides p. lesquels au moins 1 trad./syn. a été extrait	%
BUL	821873	85096	14812	17.,4
CES	53448	53442	16549	31
DAN	16708	14974	10672	71,3
DEU	341474	341006	60951	17,9
ELL	428746	421762	49335	11,7
ENG	3609456	3599203	661166	18,4
FRE	2399766	2393544	426952	17,8
HIN	24266	23901	4853	20,3
HUN	205646	203401	154893	76,2
ITA	119223	119126	48180	40,4
JAP	108800	105942	57635	54,4
NLD	373532	373531	79433	21,3
POL	393627	391626	44743	11,4
POR	188905	188728	170053	90,1
RON	115753	111305	89999	80,9
RUS	594425	488631	146296	29,9
SLK	4518	4517	1906	42,2
SPA	482886	482690	63807	13,2
SWE	368885	367964	62075	16,9
TUR	307225	305437	106535	34,9
VIE	218032	216313	117085	54,1

Pour les éditions les plus volumineuses (anglais et français), près d'un quart de leurs articles non vides bénéficient de l'extraction (respectivement 18,4% and 17,8%). Les meilleurs taux d'extraction (90,1% and 80,9%) ont été obtenus pour les éditions portugaise et le roumaine. Les éditions grecque et polonaise ont quant à elles les plus mauvais taux d'exaction (11,7% and 11,4%).

Pour le grec, ce faible taux semble être principalement le fait de l'influence d'un format de traductions implicites qui n'est pas pris en charge par notre système. Plus précisément, nous ne retenons pour cette langue que le déclencheur de traductions implicites `^#\s*\[\[.+\]`

\].?$. Toutefois, de nombreuses traductions implicites étaient marquées par le motif ^* \[\[, que nous avons choisi de ne pas inclure car il dénote également les termes connexes (également dénotés par le terme grec « *συγγενικά* » dans certains cas de figure). L'inclure aurait généré plus de bruit que d'information correcte dans nos traductions.

Le processus d'extraction a été évalué sur ces 21 éditions de langue en sélectionnant aléatoirement 3150 liens de traductions et/ou de synonymie (150 par édition de langue) pour manuellement vérifier la pertinence de l'extraction par rapport aux wiktionnaires concernés. Le tableau D.3 présente les résultats de cette évaluation ainsi qu'une analyse d'erreurs. La proportion de liens correctement extraits va de 78% (espagnol) à 97.3% (néerlandais). Ces chiffres concernent l'évaluation avant la phase de filtre, décrite à la section 9.2.2 (p.9.2.2).

Comme cela a déjà été évoqué à la section 9.2.2.1 (p. 181), nous avons identifié 4 grandes classes d'erreurs :

1. les traductions dont la langue source ou cible n'a pas été correctement identifiée (colonne *lang.* dans le tableau D.3) ;
2. des définitions prises à tort pour des traductions (colonne *déf.* dans le tableau D.3) ;
3. des traductions contenant du bruit, principalement des caractères indésirables (colonne *bruit* dans le tableau D.3) ;
4. des erreurs variées, non répertoriées ci-dessus (colonne *misc.* dans le tableau D.3).

Le tableau D.3 indique un score moyen de 87,6% pour l'extraction non filtrée. Toutefois, ce score est biaisé parce que les éditions de langues les plus modestes sont sur-représentées. En pondérant les scores par la taille de l'extraction (tableau D.2), on obtiens une estimation de score plus fiable, de l'ordre de 89,4% d'extractions correctes.
Comme cela est développé à la section 9.2.2 (p.9.2.2), il est possible de filtrer de nombreuses erreurs émanant des classes 1 à 3. Si l'on estime que la phase d'extraction de traductions et de synonymes issues de wiktionnaires est nécessairement suivie d'une étape de filtrage, il est envisageable de gagner près de 8 points de précision (cf. section 9.2.2.2 et tableau D.3).

TABLE D.3 – Résultats de l'évaluation de traductions/synonymes non-filtrés et filtrés issus de 21 éditions de wiktionnaires (taille de l'échantillon pour chaque langue : 150).

	Avant filrage						Après filtrage						
	Erronnés				Corrects		Erronnés				Corrects		Gardés
édition	misc.	lang.	déf.	bruit	total	%	misc.	lang.	déf.	bruit	total	%	%
BUL	8	1	1	4	136	90,7	3	0	1	0	93	95,9	64,7
CES	4	0	19	5	122	81,3	0	0	2	1	16	84,2	12,7
DAN	2	0	3	2	143	95,3	1	0	0	0	65	98,5	44
DEU	4	1	0	2	143	95,3	1	0	0	0	46	97,9	31,3
ELL	4	0	5	2	139	92,7	1	0	2	1	65	94,2	46
ENG	1	5	0	4	140	93,3	1	1	0	0	111	97,4	76
FRE	4	0	8	0	138	92	2	0	3	0	95	95	66,7
HIN	9	3	3	4	131	87,3	1	0	0	0	42	97,7	28,7
HUN	3	9	1	1	136	90,7	1	2	0	0	56	91,8	40,7
ITA	0	4	14	2	130	86,7	0	1	0	0	58	96,7	40
JAP	2	10	26	15	97	64,7	1	3	8	0	38	71,7	35,3
NLD	0	0	0	4	146	97,3	0	0	0	0	68	100	45,3
POL	3	3	0	2	142	94,7	0	0	0	0	68	100	45,3
POR	5	3	18	4	120	80	2	0	0	0	89	97,8	60,7
RON	0	0	13	3	134	89,3	0	0	0	0	116	100	77,3
RUS	10	1	4	2	133	88,7	1	0	0	0	95	99	64
SLK	5	7	12	0	126	84	2	0	0	0	39	95,1	27,3
SPA	5	0	15	13	117	78	0	0	0	1	46	97,9	31,3
SWE	4	4	9	0	133	88,7	2	3	0	0	69	89,6	51,3
TUR	2	11	4	0	133	88,7	0	1	1	0	83	96,5	57,3
VIE	9	0	20	0	121	80,7	0	0	0	0	31	100	20,7
Total	84	62	175	69	2760	—	19	11	17	3	1389	—	—
%	2,7	1,9	5,6	2,2	87,6	—	1,3	0,8	1,2	0,2	96,5	—	—

Bibliographie

(1995). *MUC6 '95 : Proceedings of the 6th Conference on Message Understanding*. Association for Computational Linguistics.

Aberdeen J., Bayer S., Yeniterzi R., Wellner B., Clark C., Hanauer D., Malin B. & Hirschman L. (2010). The mitre identification scrubber toolkit : design, training, and assessment. *International journal of medical informatics*, 79(12), 849–859.

Ahmad A., Halawani S. M. & Albidewi I. A. (2012). Novel ensemble methods for regression via classification problems. *Expert Systems with Applications*, 39(7), 6396 – 6401.

Ahmad K., Gillam L. & Tostevin L. (1999). University of Surrey Participation in TREC8 : Weirdness indexing for logical document extrapolation and retrieval (wilder). In *TREC*.

Al-Sughaiyer I. A. & Al-Kharashi I. A. (2004). Arabic morphological analysis techniques : A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3), 189–213.

Al-Sulaiti L. & Atwell E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, 11(2), 135–171.

Alajmi A., Saad E. & Darwish R. (2012). Toward an arabic stop-words list generation. *International Journal of Computer Applications*, 46.

Alamgir M. & Von Luxburg U. (2010). Multi-agent random walks for local clustering on graphs. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, p. 18–27 : IEEE.

Alexeeva L. (2006). Proceedings of the theoretical foundations of terminology comparison between eastern europe and western countries in conjunction with the 14th European Symposium on Language for Special Purposes (LSP). In Budin *et al.* (2006), chapter Interaction Between Terminology and Philosophy.

Allwood J., Hendrikse A. & Ahlsén E. (2010). Words and alternative basic units for linguistic analysis. *Linguistic Theory and Raw Sound*, 40, 9–25.

Ananiadou S. (1994). A methodology for automatic term recognition. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, p. 1034–1038 : Association for Computational Linguistics.

Anderson S. (1992). *A-Morphous Morphology*. Cambridge Studies in Linguistics. Cambridge University Press.

Anderson S. R. (2013). The morpheme : Its nature and use. *The Oxford Handbook of Inflection, Oxford : Oxford UniversityPress*.

Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B. & Vossen P. (2004). The meaning multilingual central repository. In *In Proceedings of the Second International WordNet Conference*, p. 80–210.

Aubin S. & Hamon T. (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, p. 380–387. Springer.

Baccouche T. & Mejri S. (2007). Norme grammaticale et description linguistique : le cas de l'arabe. *Langages*, (3), 27–37.

Badawi E., Carter M., Carter M. & Gully A. (2013). *Modern Written Arabic : A Comprehensive Grammar*. Routledge Comprehensive Grammars. Taylor & Francis.

Bakker D., Müller A., Velupillai V., Wichmann S., Brown C. H., Brown P., Egorov D., Mailhammer R., Grant A. & Holman E. W. (2009). Adding typology to lexicostatistics : a combined approach to language classification. *Linguistic Typology*, 13, 169–181.

Baldwin T., Pool J. & Colowick S. M. (2010). Panlex and lextract : Translating all words of all languages of the world. In *COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, 23-27 August 2010, Beijing, China*, p. 37–40.

Baroni A. (2011). Alphabetic vs. non-alphabetic writing : Linguistic fit and natural tendencies. *Rivista di Linguistica*, 23(2), 127–159.

Baroni M., Bernardini S., Ferraresi A. & Zanchetta E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3), 209–226.

Baroni M. & Kilgarriff A. (2006). Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics : Posters & Demonstrations*, p. 87–90 : Association for Computational Linguistics.

Basili R., Moschitti A. & Pazienza M. T. (1999). A text classifier based on linguistic processing.

Bastian M., Heymann S., Jacomy M. *et al.* (2009). Gephi : an open source software for exploring and manipulating networks. *ICWSM*, 8, 361–362.

Baudouin de Courtenay J. (1895). A Baudouin de Courtenay anthology. Bloomington, IN : Indiana University Press.

Bay S. D. (2000). Multivariate discretization of continuous variables for set mining. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, p. 315–319, New York, NY, USA : ACM.

Beesley K. R. (2001). Finite-state morphological analysis and generation of Arabic at Xerox Research : Status and plans in 2001. In *The Arabic Language Processing : Status and Prospect—39th Annual Meeting of the Association for Computational Linguistics*, p. 1–8.

Benczes R. (2006). *Creative Compounding in English : The Semantics of Metaphorical and Metonymical Noun-noun Combinations*. Human cognitive processing. John Benjamins Publishing Company.

Bender E. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(0).

Bentz C., Kiela D., Hill F. & Buttery P. (2014). Zipf's law and the grammar of languages (abstract). *A quantitative study of Old and Modern*.

Bernhard D. (2010). Apprentissage non supervisé de familles morphologiques : Comparaison de méthodes et aspects multilingues. *Traitement Automatique des Langues*, 51(2), 11–39.

Bhat B., Poddar L. & Bhattacharyya P. (2013). IndoNet : A Multilingual Lexical Knowledge Network for Indian Languages. In *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Bisetto A. & Scalise S. (2009). Classification of compounds. university of bologna.

Bond F., Chang Z. & Uchimoto K. (2008). Extracting bilingual terms from mainly monolingual data. In *14th Annual Meeting of the Association for Natural Language Processing*, Tokyo, Japan.

Boulaknadel S., Daille B. & Aboutajdine D. (2008). A multi-word term extraction program for Arabic language. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias, Eds., *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA).

Bourigault D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 3*, COLING '92, p. 977–981 : Association for Computational Linguistics.

Bourigault D., Aussenac-Gilles N. & Charlet J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18(1), 87–110.

Bourigault D. & Fabre C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de grammaire*, 25, 131–151.

Bourigault D. & Jacquemin C. (1999). Term extraction + term clustering : An integrated platform for computer-aided terminology. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, p. 15–22 : Association for Computational Linguistics.

Buckwalter T. (2002). Buckwalter Arabic morphological analyzer version 1.0. *Linguistic Data Consortium (LDC) catalog number LDC2002L49*.

G. Budin, C. Laurén, H. Picht, N. Pilke, M. Rogers & B. Toft, Eds. (2006). *Proceedings of The Theoretical Foundations of Terminology Comparison Between Eastern Europe and Western Countries in Conjunction with the 14th European Symposium on Language for Special Purposes (LSP)*. Content and Communication. Ergon-Verlag.

Bynon T. (2004). Approaches to morphological typology. In *Morphologie / Morphology. Ein internationales Handbuch zur Flexion und Wortbildung*. Berlin, Germany : De Gruyter.

Cabré M. T. (1998). *La terminologie : théorie, méthode et applications*. Les Presses de l'Université d'Ottawa, armand colin edition.

Cabré M. T. (1995). On diversity and terminology. *Terminology*, 2(1), 1–16.

Campenhoudt M. V. (2006). Que nous reste-t-il d'Eugen Wüster ? Colloque international Eugen Wüster et la terminologie de l'École de Vienne.

Candel D. (2004). Wüster par lui-même. In C. Cortès, Ed., *Terminologie : problèmes théoriques*, p. 15–32. Cahiers du CIEL.

Cao G., Gao J. & Nie J.-Y. (2007). A system to mine large-scale bilingual dictionaries from monolingual web pages. *Proceedings of thé MT Summit XI - The Eleventh Machine Translation Summit*, p. 57–64.

Cao L., Zhao X., Zheng H. & Zhao B. Y. (2011). *Atlas : Approximating shortest paths in social graphs*. Rapport interne, Department of Computer Science, University of California, Santa Barbara.

Caraballo S. A. & Charniak E. (1999). Determining the specificity of nouns from text. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, p. 63–70.

Carletta J. (1996). Assessing agreement on classification tasks : the kappa statistic. *Computational linguistics*, 22(2), 249–254.

Catach N. (1997). Les Histoires de l'Écriture – Panorama critique –. *Histoire Épistémologie Langage*, 19(2), 177–185.

Cen Y., Han Z. & Ji P. (2008). Chinese term recognition and extraction based on hidden markov model. In *Pacific-Asia Workshop on Computational Intelligence and Industrial Application (PACIIA'08)*, volume 2, p. 219–224 : IEEE.

Chawla N., K.W. B., L.O. H. & Kegelmeyer W. (2002). SMOTE : synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16.

Chen M., Chang B. & Pei W. (2014). A Joint Model for Unsupervised Chinese Word Segmentation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 854–863 : Association for Computational Linguistics.

Chung F. (2007). Random walks and local cuts in graphs. *Linear Algebra and its applications*, 423(1), 22–32.

Clark A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, p. 59–66 : Association for Computational Linguistics.

Claveau V. & L'Homme M.-C. (2005). Structuring terminology using analogy-based machine learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE*, p. 17–18.

Clément L. & Villemonte de La Clergerie E. (2005). MAF : a morphosyntactic annotation framework. In *Proc. of the 2nd Language & Technology Conference (LT'05)*, p. 90–94, Poznan, Poland : Wydawnictwo Poznańskie.

Comrie B. (1989). *Language universals and linguistic typology*. Chicago : University of Chicago press.

Conrado M. D. S., Pardo T. A. S. & Rezende S. O. (2013). A machine learning approach to automatic term extraction using a rich feature set. In *Proceedings of the NAACL HLT 2013 Student Research Workshop*, p. 16–23, Atlanta, Georgia.

Constant M., Tellier I., Duchier D., Dupont Y., Sigogne A., Billot S. *et al.* (2011). Intégrer des connaissances linguistiques dans un crf : application à l'apprentissage d'un segmenteur-étiqueteur du français. *TALN2011*, 1.

Corbett G. G. (1983). The number of genders in Polish. *Papers and Studies in Contrastive Linguistics*, VXI, 83–89. © 1983 Greville G. Corbett and the School of English, Adam Mickiewicz University, Poznań, Poland. Note : journal is now called Poznań Studies in Contemporary Linguistics.

Creutz M. & Lagus K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*.

da Silva J. F. & Lopes G. P. (1999). A local maxima method and a fair dispersion normalization for extracting multi-word units from corpora. In *Sixth Meeting on Mathematics of Language*.

Dagan I. & Church K. (1994). Termight : Identifying and translating technical terminology. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, ANLC '94, p. 34–40 : Association for Computational Linguistics.

Dagan I., Itai A. & Schwall U. (1991). Two languages are more informative than one. In *in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 130–137.

Daille B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7.

Daille B. (2003). Conceptual structuring through term variations. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment - Volume 18*, MWE '03, p. 9–16 : Association for Computational Linguistics.

Daille B. (2005). Variations and application-oriented terminology engineering. *Terminology*, 11(1), 181–197.

Daille B. (2012). Building bilingual terminologies from comparable corpora : The ttc termsuite. In *The 5th Workshop on Building and Using Comparable Corpora*, p.29.

Daille B. & Blancafort H. (2013). Knowledge-poor and knowledge-rich approaches for multilingual terminology extraction. In *Proceedings, 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2013)*.

Daniels P. T. (2003). *The Handbook of Linguistics*, chapter Chapter 3. Blackwell.

David S. & Plante P. (1990). De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence artificielle et sciences cognitives au Québec*, 3(3), 140–154.

Davidov D. & Rappoport A. (2009). Enhancement of lexical concepts using cross-lingual web mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2-Volume 2*, p. 852–861 : Association for Computational Linguistics.

Dayan P. (1999). Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*.

de Borda J. C. (1781). Mémoire sur les élections au scrutin.

de Melo G. (2012). *Graph-based Methods for Large-Scale Multilingual Knowledge Integration*. Saarbrücken, Germany : universaar - Saarland University Press.

de Melo G. & Weikum G. (2009a). Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09 : ACM.

de Melo G. & Weikum G. (2009b). Towards a Universal Wordnet by Learning from Combined Evidence. In *CIKM'09*.

de Paiva V., Rademaker A. & de Melo G. (2012). Openwordnet-pt : An open brazilian wordnet for reasoning. In *Proceedings of the 24th International Conference on Computational Linguistics*.

Déjean H., Gaussier E. & Sadat F. (2002). Bilingual Terminology Extraction : An Approach based on a Multilingual thesaurus Applicable to Comparable Corpora. In *In Proceedings of the 19th International Conference on Computational Linguistics COLING 2002*, p. 218–224.

Diab M. T. (2004). The feasibility of bootstrapping an Arabic wordnet leveraging parallel corpora and and an English wordnet. In *Proceedings of the Arabic Language Technologies and Resources*, Cairo.

Dong Z., Dong Q. & Hao C. (2010). Word segmentation needs change-from a linguist's view. In *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing*, p. 1–7.

Dougherty J., Kohavi R. & Sahami M. (1995). Supervised and unsupervised discretization of continuous features. In e. Armand Prieditis & Stuart Russell, Ed., *Machine Learning : Proceedings of the Twelfth International Conference*, p. 194–202, San Francisco, USA : Morgan Kaufmann Publishers.

Dryer M. S. (1995). Frequency and pragmatically unmarked word order. *Word order in discourse*, 30, 105.

Dryer M. S. (2007). Word order. *Language typology and syntactic description*, 1, 61–131.

Dryer M. S. (2013). *Prefixing vs. Suffixing in Inflectional Morphology*, In M. S. Dryer & M. Haspelmath, Eds., *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology : Leipzig.

Drzazga G. (2013). *The Puzzle of Grammatical Gender : Insights from the Cognitive Theory of Translation and the Nature of Polish Hybrid Nouns*. PhD thesis, McMaster University, Hamilton, Ontario.

Ducrot O. & Schaeffer J. (1995). *Nouveau dictionnaire encyclopédique des sciences du langage*. Points (Paris). Editions du Seuil.

Dyvik H. (1998). Translations as semantic mirrors : from parallel corpus to wordnet. In *Proceedings of the Workshop Multilinguality in the lexicon II at the 13th biennial European Conference on Artificial Intelligence (ECAI'98)*, p. 24–44, Brighton, UK.

Edachery J., Sen A. & Brandenburg F. J. (1999). Graph clustering using distance-k cliques. In *Graph drawing*, p. 98–106 : Springer.

Ehrmann M. (2008). *Les Entités Nommées, de la Linguistique au TAL : Statut théorique et méthodes de désambiguïsation*. PhD thesis, Université Paris 7.

Eifring H. B. & Theil R. (2004). *Linguistics for students of Asian and African languages*. Institutt for østeuropeiske og orientalske studier.

El Ayari S. (2009). *Évaluation transparente du traitement des éléments de réponse à une question factuelle*. PhD thesis, Université Paris Sud-Paris XI.

El Hadi W. M., Timimi I., Dabbadie M., Choukri K., Hamon O. & Chiao Y.-C. (2006). Terminological resources acquisition tools : Toward a user-oriented evaluation model. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, p. 945–948, Genova, Italy : European Language Resources Association (ELRA).

Estivill-Castro V. (2002). Why so many clustering algorithms : a position paper. *ACM SIGKDD Explorations Newsletter*, 4(1), 65–75.

Etzioni O., Reiter K., Soderland S. & Sammer M. (2007). Lexical translation with application to image search on the web.

Evert S. (2005). *The Statistics of Word Cooccurrences : Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Evert S. (2007). Corpora and collocations (extended manuscript). Institute of Cognitive Science, University of Osnabrück.

Fan X., Shimizu N. & Nakagawa H. (2009). Automatic extraction of bilingual terms from a Chinese-Japanese parallel corpus. In *Proceedings of the 3rd International Universal Communication Symposium*, p. 41–45 : ACM.

Fang J., Sui L.-N. & Jian H.-Y. (2013). Comparative analysis of continuous entropy estimation with different unsupervised discretization methods. In *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering* : Atlantis Press.

Farreres X., German Rigau & Rodríguez H. (1998). Using wordnet for building wordnets. *Computing Research Repository*, cmp-lg/980.

C. Fellbaum, Ed. (1998). *WordNet : An Electronic Lexical Database*. Cambridge, Massachusetts : MIT Press.

Feuillet J. (2006). *Introduction à la typologie linguistique*. Bibliothèque de grammaire et de linguistique. Honoré Champion.

Firth J. R. (1957). A synopsis of linguistic theory 1930-55. *The Philological Society*, 1952-59, 1–32.

Foo J. (2012). Computational terminology : Exploring bilingual and monolingual term extraction. Master's thesis, Linköping University.

Foo J. & Merkel M. (2010). Using machine learning to perform automatic term recognition. In *Proceedings of the LREC 2010 Workshop on Methods for automatic acquisition of Language Resources and their evaluation methods*, p. 49–54.

Fortunato S. (2010). Community detection in graphs. *Physics Reports*, 486(3), 75–174.

Frantzi K., Ananiadou S. & Tsuji J. (1998). The c-value/nc-value method of automatic recognition for multi-word terms. In *Proceedings of the ECDL*.

Fung P. & Cheung P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04 : Association for Computational Linguistics.

Fung P. & McKeown K. (1997). Finding terminology translations from non-parallel corpora. In *Proceedings of the 5th Annual Workshop on Very Large Corpora*, p. 192–202.

Gaillard B., Gaume B. & Navarro E. (2011). Invariants and variability of synonymy networks : Self mediated agreement by confluence. In *Proceedings of TextGraphs-6 : Graph-based Methods for Natural Language Processing*, p. 15–23 : Association for Computational Linguistics.

Gaume B. (2004). Balades Aléatoires dans les petits mondes lexicaux. *I3 Information Interaction Intelligence*.

Gaume B. (2008). Mapping the forms of meaning in small worlds. *International Journal of Intelligent Systems*, 23(7), 848–862.

Gelb I. J. (1952). *A study of writing : The foundations of grammatology*. "The" University of Chicago Press.

Gil D. (2001). *Linguistic Fieldwork*, chapter Escaping Eurocentrism : fieldwork as a process of unlearning. Cambridge University Press.

Gilbert A. L., Regier T., Kay P. & Ivry R. B. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2), 489–494.

Göksel A. & Kerslake C. (2005). *Turkish : A Comprehensive Grammar*. Comprehensive grammars. Routledge.

Goldsmith J. (2000). Linguistica : An automatic morphological analyzer. In *Proceedings of 36th meeting of the Chicago Linguistic Society*.

Greenberg J. H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 26(3), pp. 178–194.

Greenberg J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg, Ed., *Universals of Human Language*, p. 73–113. Cambridge, Mass : MIT Press.

Greenberg J. H. (1966). *Language universals (With special reference to feature hierarchies)*. Mouton, The Hague.

Grigonytė G., Rimkutė E., Utka A. & Boizou L. (2011). Experiments on Lithuanian term extraction. In *Proceedings of NODALIDA 2011 Conference*, p. 82–89 : Northern European Association for Language Technology (NEALT).

Grouin C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. PhD thesis, Université Pierre et Marie Curie-Paris VI.

Gruber T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2), 199–220.

Guarino N., Oberle D. & Staab S. (2009). What is an ontology ? In *Handbook on ontologies*, p. 1–17. Springer.

J. Gumperz & S. Levinson, Eds. (1996). *Rethinking Linguistic Relativity*. Studies in the Social and Cultural Foundations of Language. Cambridge University Press.

Hammarström H. & Borin L. (2011). Unsupervised learning of morphology. *Comput. Linguist.*, 37(2), 309–350.

Hamp B. & Feldweg H. (1997). Germanet - a lexical-semantic net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, p. 9–15.

Han J. & Kamber M. (2006). *Data Mining : Concepts and Techniques, Second Edition*. Morgan kaufmann.

Hanoka V. & Sagot B. (2012). Wordnet extension made simple : A multilingual lexicon-based approach using wiki resources. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey : European Language Resources Association (ELRA).

Hanoka V. & Sagot B. (2014). YaMTG : An Open-Source Heavily Multilingual Translation Graph Extracted from Wiktionaries and Parallel Corpora. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, Eds., *Proceedings of the Ninht International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).

Hartigan J. A. & Wong M. A. (1979). Algorithm as 136 : A k-means clustering algorithm. *Journal of the Royal Statistical Society - Serie C : Applied Statistics*, 28(1), 100–108.

Haspelmath M. (2009). An empirical test of the agglutination hypothesis. In S. Scalise, E. Magni & A. Bisetto, Eds., *Universals of Language Today*, volume 76 of *Studies in Natural Language and Linguistic Theory*, p. 13–29. Springer Netherlands.

Healy D. (2012). *Complete Vietnamese : Teach Yourself*. Complete Languages. Hodder & Stoughton.

Heid U. (1999). A linguistic bootstrapping approach to the extraction of term candidates from German text.

Hjelm H. (2007). Identifying cross language term equivalents using statistical machine translation and distributional association measures. In *Proceedings of NODALIDA*, p. 97–104.

Hu Y., Li M., Zhang P., Fan Y. & Di Z. (2008). Community detection by signaling on complex networks. *Physical Review E*, 78(1), 016115.

Humphries M. D. & Gurney K. (2008). Network 'small-world-ness' : a quantitative method for determining canonical network equivalence. *PLoS One*, 3(4), e0002051.

Ide N., Erjavec T. & Tufiş D. (2002). Sense discrimination with parallel corpora. In *Proceedings of the ACL-02 workshop on Word sense disambiguation : recent successes and future directions - Volume 8*, WSD '02, p. 61–66 : Association for Computational Linguistics.

Ideue M., Yamamoto K., Utiyama M. & Sumita E. (2011). A comparison of unsupervised bilingual term extraction methods using phrase tables. *Proc. MT Summit XIII, Xiamen*.

Isahara H., Bond F., Uchimoto K., Utiyama M. & Kanzaki K. (2008). Development of the Japanese WordNet. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias, Eds., *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA).

Ismail M. (2003). An empirical investigation of the impact of discretization on common data distributions. Master's thesis, RMIT University, Melboune, Australia.

Ittoo A. & Bouma G. (2013). Term extraction from sparse, ungrammatical domain-specific documents. *Expert Syst. Appl.*, 40(7), 2530–2540.

Jacquemin C. (1999). Syntagmatic and paradigmatic representations of term variation. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, p. 341–348 : Association for Computational Linguistics.

Jain A., Nandakumar K. & Ross A. (2005). Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12), 2270–2285.

Jones K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1), 11–21.

Justeson J. S. & Katz S. M. (1995). Technical terminology : some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1, 9–27.

Kageura K. (1995). Toward the theoretical study of terms : A sketch from the linguistic viewpoint. *Terminology*, 2(2), 239–257.

Kageura K. & Umino B. (1996). Methods of automatic term recognition : A review. *Terminology*, 3(2), 259–289.

Kamholz D., Pool J. & Colowick S. M. (2014). Panlex : Building a resource for panlingual lexical translation. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).

Kando N., Kuriyama K., Nozue T., Eguchi K., Kato H., Hidaka S. & Adachi J. (1999). The NTCIR workshop : the first evaluation workshop on Japanese text retrieval and cross-lingual information retrieval. In *Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages (IRAL'99)*.

Khaliq B. & Carroll J. (2013). Induction of root and pattern lexicon for unsupervised morphological analysis of Arabic. *International Joint Conference on Natural Language Processing*, p. 1012–1016.

Khoja S. & Garside R. (1999). Stemming arabic text. *Lancaster, UK, Computing Department, Lancaster University*.

Kim S. N., Baldwin T. & Min-Yen K. (2009). An unsupervised approach to domain-specific term extraction. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, p.94.

Kirkpatrick S., Vecchi M. *et al.* (1983). Optimization by simmulated annealing. *science*, 220(4598), 671–680.

Klir G. & Wierman M. (1999). *Uncertainty-Based Information : Elements of Generalized Information Theory*. Studies in Fuzziness and Soft Computing. Physica-Verlag HD.

Knowles M. & Moon R. (2006). *Introducing Metaphor*. Routledge.

Koehn P. & Knight K. (2001). Knowledge sources for word-level translation models. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, p. 27–35.

Kohavi R. *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, p. 1137–1145.

Kohonen O., Virpioja S. & Lagus K. (2010). Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on*

Computational Morphology and Phonology, p. 78–86 : Association for Computational Linguistics.

Kotsiantis S. & Kanellopoulos D. (2006). Discretization techniques : A recent survey. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 47–58.

Kozareva Z. & Hovy E. (2010). A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, p. 1110–1118 : Association for Computational Linguistics.

Kudo T., Yamamoto K. & Matsumoto Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. In *EMNLP*, volume 4, p. 230–237.

Kullback S. & Leibler R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86.

M. Kurimo, S. Virpioja, V. T. Turunen & others, Eds. (2010a). *Proceedings of the Morpho Challenge 2010 Workshop*. Aalto University School of Science and Technology.

Kurimo M., Virpioja S., Turunen V. T., Blackwood G. W. & Byrne W. (2010b). Overview and results of morpho challenge 2009. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, p. 578–597. Springer.

Lafferty J., McCallum A. & Pereira F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. p. 282–289.

Laughlin C. (1997). Intuition : The inside story : Interdisciplinary perspectives. chapter The Nature of Intuition : A neuropsychological Approach. Routledge.

Lavergne T., Cappé O. & Yvon F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

Lebart L. & Salem A. (1994). *Statistique textuelle*. Dunod.

Lee C.-M., Huang C.-K., (Fayuan) K.-M. T. & Chen K.-H. (2012). Iterative Machine-Learning Chinese Term Extraction. In H.-H. Chen & G. Chowdhury, Eds., *The Outreach of Digital Libraries : A Globalized Resource Network - 14th International Conference on Asia-Pacific Digital Libraries (ICADL'12)*, volume 7634 of *Lecture Notes in Computer Science*, p. 309–312 : Springer Berlin Heidelberg.

Leitchik V. & Shelov S. (2006). Proceedings of the theoretical foundations of terminology comparison between eastern europe and western countries in conjunction with the 14th European Symposium on Language for Special Purposes (LSP). In Budin *et al.* (2006), chapter Some Basics Concepts of Terminology : Tradition and Innovations.

Li L. S., Dang Y. Z., Zhang J. & Li D. (2012). Domain term extraction based on conditional random fields combined with active learning strategy. *Journal of Information & Computational Science*, 9(7), 1931–1940.

R. Lieber & P. Štekauer, Eds. (2005). *Handbook of Word-Formation*. Springer.

R. Lieber, P. Štekauer & M. C. Baker, Eds. (2009). *The Oxford handbook of compounding*. Oxford University Press Oxford/New York.

Lin D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, p. 317–324 : Association for Computational Linguistics.

Liu F., Pennell D., Liu F. & Liu Y. (2009). Unsupervised approaches for automatic keyword extraction using meeting transcripts. In *Proceedings of human language technologies : The 2009 annual conference of the North American chapter of the association for computational linguistics*, p. 620–628 : Association for Computational Linguistics.

Liu P.-P., Li W.-J., Lin N. & Li X.-S. (2013). Do Chinese Readers Follow the National Standard Rules for Word Segmentation during Reading ? *PLoS ONE*, 8(2).

Liu W., Weichselbraun A., Scharl A. & Chang E. (2005). Semi-automatic ontology extension using spreading activation. *Journal of Universal Knowledge Management*, 1, 50–58.

Lo R. T.-W., He B. & Ounis I. (2005). Automatically building a stopword list for an information retrieval system. In *Journal on Digital Information Management : Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, volume 5, p. 17–24.

Longadge R. & Dongre S. (2013). Class imbalance problem in data mining review. *arXiv preprint arXiv :1305.1707*.

López L. M., Ruiz I. F., Bueno R. M. & Ruiz F. T. (2000). Dynamic discretization of continuous values from time series. In R. L. de Mántaras & E. Plaza, Eds., *ECML*, volume 1810 of *Lecture Notes in Computer Science*, p. 280–291 : Springer.

Loukachevitch N. V. (2012). Automatic term recognition needs multiple evidence. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, Eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, p. 2401–2407, Istanbul, Turkey : European Language Resources Association (ELRA).

Lud M.-C. & Widmer G. (2000). Relative unsupervised discretization for association rule mining. In *Principles of data mining and knowledge discovery*, p. 148–158. Springer.

Lurie D. B. (2006). Language, writing, and disciplinarity in the critique of the "ideographic myth" : Some proleptical remarks. *Language & Communication*, 26(3), 250–269.

L'Homme M.-C. (2005). Sur la notion de «terme». *Meta : Journal des traducteursMeta :/ Translators' Journal*, 50(4), 1112–1132.

Magistry P. (2013). *Unsupervised Word Segmentation and Wordhood Assessment. The case for Mandarin Chinese*. PhD thesis, Université Paris Diderot (Paris 7).

Magistry P. & Sagot B. (2012). Unsupervized Word Segmentation : the case for Mandarin Chinese. In *ACL - Annual Meeting of the Association for Computational Linguistics - 2012*, Jeju, Corée, République De : ACL.

Maimon O. & Rokach L. (2005). *Data Mining and Knowledge Discovery Handbook*. The Kluwer International Series in Engineering and Computer Science. Springer.

Makkai A. (1972). *Idiom Structure in English*. Number 48 in Janua Linguarum. Series Maior. De Gruyter.

Mandelbrot B. B. (1953). An informational theory of the statistical structure of languages. In W. Jackson, Ed., *Communication theory : papers read at a Symposium on "Applications of Communication Theory" held at the Institution of Electrical Engineers, London, September 22nd–26th 1952*, p. 486–502, London, UK : Butterworths.

Manning C. & Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, 1 edition.

Manning C. D., Raghavan P. & Schütze H. (2008). *Scoring, term weighting, and the vector space model*, In *Introduction to Information Retrieval*, chapter Ch. 6. Cambridge University Press.

Martin S., Brown W. M., Klavans R. & Boyack K. W. (2011). Openord : an open-source toolbox for large graph layout. In *IS&T/SPIE Electronic Imaging*, p. 786806–786806 : International Society for Optics and Photonics.

Mausam, Soderland S., Etzioni O., Weld D. S., Skinner M. & Bilmes J. (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th International Conference on NLP of the AFNLP*, ACL '09, p. 262–270, Suntec, Singapore.

McCallum A. & Li W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, p. 188–191 : Association for Computational Linguistics.

McCallum A. K. (2002). Mallet : A machine learning for language toolkit. http ://mallet.cs.umass.edu.

McCarthy J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic inquiry*, 12, 373–418.

McEnery A. & Xiao Z. (2004). The Lancaster Corpus of Mandarin Chinese : A corpus for monolingual and contrastive language study. *Religion*, 17, 3–4.

McInnes B. T. (2004). Extending the log likelihood measure to improve collocation identification. Master's thesis, University of Minnesota.

Mel'čuk I. A. (1997). *Leçon inaugurale faite le vendredi 10 janvier 1997, Collège de France, Chaire internationale : vers une linguistique sens-texte*. Collège de France.

Mel'čuk I. A. (1998). Collocations and lexical functions. *2001 [1998]*, p. 23–54.

Meyer C. M. & Gurevych I. (2012). Wiktionary : a new rival for expert-built lexicons ? exploring the possibilities of collaborative lexicography. In S. Granger & M. Paquot, Eds., *Electronic Lexicography*, chapter 13, p. 259–291. Oxford : Oxford University Press.

Meyer I. (1993). In R. Strehlow & S. Wright, Eds., *Standardizing Terminology for Better Communication : Practice, Applied Theory, and Results*, chapter Concept Management for Terminology : A Knowledge Engineering Approach. American Society for Testing and Materials (ASTM).

Milgram S. (1967). The small world problem. *Psychology Today*, 67(1), 61–67.

Moens M. (2006). *Information Extraction : Algorithms and Prospects in a Retrieval Context*. The Information Retrieval Series. Springer.

Molinero, Miguel A., Sagot B. & Nicolas L. (2009). A morphological and syntactic wide-coverage lexicon for Spanish : The Leffe. In *RANLP 2009 - Recent Advances in Natural Language Processing*, Borovets, Bulgaria.

Mollineda V. G. J. S. R. & Sotoca R. A. J. (2007). The class imbalance problem in pattern classification and learning. *Simposio de Inteligencia Computacional, SICO'2007 (IEEE Computational Intelligence Society, SC). Congreso Español de Informática (CEDI 2007).*

Mondary T., Nazarenko A., Zargayouna H. & Barreaux S. (2012). The Quaero Evaluation Campaign on Term Extraction. In *The eighth international conference on Language Resources and Evaluation (LREC)*, p. 663–669, Istanbul, Turquie.

Moravcsik E. A. (2007). What is universal about typology ? *Linguistic Typology*, 11, 27–41.

Muller P. & Langlais P. (2011). Comparaison d'une approche miroir et d'une approche distributionnelle pour l'extraction de mots sémantiquement reliés. *Traitement Automatique des Langues Naturelles (TALN), Montpellier*, 1, 235–246.

Nagata M., Saito T. & Suzuki K. (2001). Using the web as a bilingual dictionary. In *Proceedings of the Workshop on Data-driven Methods in Machine Translation - Volume 14*, DMMT '01, p. 1–8 : Association for Computational Linguistics.

Naït-ali A. & Fournier R. (2012). *Traitement du signal et de l'image pour la biométrie*. Lavoisier.

Nakagawa H. & Mori T. (2002). A simple but powerful automatic term extraction method. In *COLING-02 on COMPUTERM 2002 : Second International Workshop on Computational Terminology - Volume 14*, COMPUTERM '02, p. 1–7 : Association for Computational Linguistics.

Navigli R. & Ponzetto S. P. (2010). BabelNet : Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 216–225, Uppsala, Sweden.

Navigli R. & Ponzetto S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217–250.

Navigli R., Velardi P. & Faralli S. (2011). A graph-based algorithm for inducing lexical taxonomies from scratch. In *IJCAI*, p. 1872–1877.

Nazar R., Wanner L. & Vivaldi J. (2008). Two step flow in bilingual lexicon extraction from unrelated corpora. In *Proceedings of the EAMT (European Association for Machine Translation) Conference*, Hamburg, Germany.

Nazarenko A. & Zargayouna H. (2009). Evaluating term extraction. In *Proceegings of International Conference Recent Advances in Natural Language Processing (RANLP'09)*, p. 299–304, Borovets, Bulgarie.

Nazarenko A., Zargayouna H., Hamon O. & van Puymbrouck J. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions . *Traitement Automatique des Langues (TAL)*, 50(1), 257–281.

Nichols J. (2007). What, if anything, is typology ? *Linguistic Typology*, 11, 231–238.

Nicholson J., Cohn T. & Baldwin T. (2012). Evaluating a morphological analyser of inuktitut. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, NAACL HLT '12, p. 372–376 : Association for Computational Linguistics.

Niemann H. (1990). *Pattern analysis and understanding*, volume 4 of *Springer series in information sciences*. Springer-Verlag.

Noor N. H. B. M., Sapuan S. & Bond F. (2011). Creating the Open Wordnet Bahasa. In H. H. Gao & M. Dong, Eds., *PACLIC*, p. 255–264 : Digital Enhancement of Cognitive Development, Waseda University.

Oh J.-H., Lee K. & Choi K.-S. (2000). Term recognition using technical dictionary hierarchy. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, p. 496–503 : Association for Computational Linguistics.

Ordan N., Ilan B., Ordan N. & Wintner S. (2007). S. : Hebrew wordnet : a test case of aligning lexical databases across languages. *International Journal of Translation*, 19, 39–58.

Packard J. L. (2000). *The Morphology of Chinese A Linguistic and Cognitive Approach*. Cambridge University Press.

Pakendorf B. *et al.* (2007). *Contact in the prehistory of the Sakha (Yakuts) : Linguistic and genetic perspectives*. PhD thesis, Netherlands Graduate School of Linguistics,(LOT) Utrecht.

Park Y., Byrd R. J. & Boguraev B. K. (2002). Automatic glossary extraction : Beyond terminology identification. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, p. 1–7 : Association for Computational Linguistics.

Pecina P. & Schlesinger P. (2006). Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, p. 651–658 : Association for Computational Linguistics.

Petrović S., Šnajder J. & Bašić B. D. (2009). Extending lexical association measures for collocation extraction. *Computer Speech & Language*, 24(2), 383–394.

Pianta E., Bentivogli L. & Girardi C. (2002). Multiwordnet : developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Pinnis M., Ljubešić N., Ştefănescu D., Skadiņa I., Tadić M. & Gornostay T. (2012). Term extraction, tagging, and mapping tools for under-resourced languages. In *Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June*, p. 20–21.

Polguère A. (2003). Collocations et fonctions lexicales : pour un modèle d'apprentissage. *Les Collocations. Analyse et traitement*, p. 117–133.

Polinsky M. & Kluender R. (2007). Linguistic typology and theory construction : Common challenges ahead. *Linguistic Typology*, 11, 273–283.

Popescu I.-I. & Altmann G. (2008). Zipf's mean and language typology. *Glottometrics*, (16), 31–37.

Provost F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, p. 1–3.

Przepiórkowski A., Górski R. L., Lewandowska-Tomaszyk B. & Lazinski M. (2008). Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco*.

Rao R. B., Fung G. & Rosales R. (2008). On the dangers of cross-validation. an experimental evaluation. In *SDM*, p. 588–596 : SIAM.

Rapp R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, p. 320–322 : Association for Computational Linguistics.

Ratinov L. & Roth D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, p. 147–155 : Association for Computational Linguistics.

Resnik P. & Yarowsky D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the SIGLEX Workshop*.

Rey A. (1979). *La Terminologie : noms et notions*. Que Sais-Je ? Presses Univ. de France.

Rondeau G. (1984). *Introduction à la terminologie*. Gaëtan Morin, 2 edition.

Rosch E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology : General*, 104, 192–233.

Rosch E. (1978). Principles of categorization. In E. Rosch & B. Lloyd, Eds., *Cognition and categorization*, p. 27–48. Hillsdale, New Jersey : Erlbaum.

Rubio G. (2005). Chasing the semitic root : The skeleton in the closet. *Aula Orientalis*, 23, 45–63.

Sadeghi M. & Vegas J. (2014). Automatic identification of light stop words for persian information retrieval systems. *Journal of Information Science*.

Sager J. (1990). *Practical Course in Terminology Processing*. John Benjamins Publishing Company.

Sagot B. (2009). Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish. In *Human Language Technology. Challenges of the Information Society*, p. 85–95. Springer.

Sagot B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malte.

Sagot B. (2013). Construction de ressources lexicales pour le traitement automatique des langues. In N. Gala & M. Zock, Eds., *Ressources Lexicales – Contenu, construction, utilisation, évaluation*, volume 30 of *Lingvisticæ Investigationes Supplementa*, p. 217–254. John Benjamins.

Sagot B. (2014). Delex, a freely-available, large-scale and linguistically grounded morphological lexicon for German. In N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis, Eds., *Proceedings of*

the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland : European Language Resources Association (ELRA).

Sagot B. & Boullier P. (2008). SxPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, 49(2), 155–188.

Sagot B. & Fišer D. (2008). Building a free French wordnet from multilingual resources. In *Ontolex 2008*, Marrakech, Morocco.

Sagot B. & Fišer D. (2012). Automatic extension of WOLF. In *Proceedings of the 6th Global WordNet Conference*, Matsue, Japan.

Sagot B. & Stern R. (2012). Aleda, a free large-scale entity database for French. In *LREC 2012 : eighth international conference on Language Resources and Evaluation*, p. 4–pages.

Sang E. F. & Veenstra J. (1999). Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, p. 173–179 : Association for Computational Linguistics.

Sapir E. (1921). *Language : An Introduction to the Study of Speech*. Harvest book. Harcourt, Brace.

Saranya C. & Manikandan G. (2013). A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering & Technology*, 5(3).

Savary A. (2000). *Recensement et description des mots composés - méthodes et applications*. Theses, Université de Marne la Vallée.

Say B., Zeyrek D., Oflazer K. & Özge U. (2002). Development of a corpus and a treebank for present-day written Turkish. In *Proceedings of the eleventh international conference of Turkish linguistics*, p. 183–192.

Schachter P. (1996). *The subject in Tagalog : Still none of the above*, volume 15. UCLA.

Schaeffer S. E. (2005). Stochastic local clustering for massive graphs. In *Advances in knowledge discovery and data mining*, p. 354–360. Springer.

Schaeffer S. E. (2007). Graph clustering. *Computer Science Review*, 1(1), 27–64.

Sérasset G. (2012). Dbnary : Wiktionary as a lemon-based multilingual lexical resource in RDF. *Semantic Web Journal-Special issue on Multilingual Linked Open Data*.

Serrano M. Á., Flammini A. & Menczer F. (2009). Beyond Zipf's law : Modeling the structure of human language. *CoRR*, abs/0902.0606.

Sha F. & Pereira F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, p. 134–141 : Association for Computational Linguistics.

Shanon C. (1948). A mathematical theory of communication. the bell systems technical journal, 27.

Šíma J. & Schaeffer S. E. (2006). On the np-completeness of some graph cluster measures. In *SOFSEM 2006 : Theory and Practice of Computer Science*, p. 530–537. Springer.

Skorik P. (1977). *Grammatika čukotskogo jazyka : Fonetika i morfologija imennyx častej reči*. Nauka.

Smadja F. (1993). Retrieving collocations from text : Xtract. *Computational Linguistics*, 19, 143–177.

Spencer A. (1991). *Morphological theory : An introduction to word structure in generative grammar*, volume 2. Blackwell Oxford.

Spielman D. A. & Teng S.-H. (2013). A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1), 1–26.

Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D. & Grigoriadou M. (2002). Balkanet a multilingual semantic network for the Balkan languages. p. 21–25.

Štekauer P., Valera S. & Kőrtvélyessy L. (2012). *Word-Formation in the World's Languages : A Typological Survey*. Cambridge University Press.

Sutton C. & McCallum A. (2010). An introduction to conditional random fields. *arXiv preprint arXiv :1011.4088*.

Syal P. & Jindal D. (2007). *An Introduction to Linguistics : Language, Grammar and Semantics*. Eastern Economy Edition. PHI Learning.

Szmrecsanyi B. & Kortmann B. (2009). The morphosyntax of varieties of English worldwide : A quantitative perspective. *Lingua*, 119(11), 1643 – 1663. The Forests behind the Trees.

Tadić M. & Šojat K. (2003). Finding multiword term candidates in Croatian. In *IESL2003*.

Taghva K., Elkhoury R. & Coombs J. (2005). Arabic stemming without a root dictionary. *Information Technology : Coding and Computing, International Conference on*, 1, 152–157.

Tellier I., Eshkol I., Taalab S., Prost J.-P. *et al.* (2010). POS-tagging for oral texts with CRF and category decomposition. *Research in Computing Science*, 46, 79–90.

Temmerman R. (2000). *Towards New Ways of Terminology Description : The Sociocognitive-Approach*. Terminology and Lexicography Research and Practice Series. Benjamins Publishing Company.

The Unicode Consortium (2011). *The Unicode Standard*. Rapport interne Version 6.0.0, Unicode Consortium, Mountain View, CA.

Tiedemann J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova & R. Mitkov, Eds., *Recent Advances in Natural Language Processing (vol V)*, p. 237–248. Amsterdam/Philadelphia : John Benjamins.

Torgo L. & Gama J. (1997). Search-based class discretization. In M. van Someren & G. Widmer, Eds., *Machine Learning : ECML-97*, volume 1224 of *Lecture Notes in Computer Science*, p. 266–273. Springer Berlin Heidelberg.

Torii M., Wagholikar K. & Liu H. (2011). Using machine learning for concept extraction on clinical documents from multiple data sources. *Journal of the American Medical Informatics Association*, 18(5), 580–587.

Tsarfaty R., Seddah D., Goldberg Y., Kübler S., Candito M., Foster J., Versley Y., Rehbein I. & Tounsi L. (2010). Statistical parsing of morphologically rich languages (SPMRL) : what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, p. 1–12 : Association for Computational Linguistics.

Tsuruoka Y., Tsujii J. & Ananiadou S. (2009). Fast full parsing by linear-chain conditional random fields. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, p. 790–798 : Association for Computational Linguistics.

Tufiş D., Cristea D. & Stamou S. (2004). BalkaNet : Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal on Information Science and Technology. Special Issue on BalkaNet*, volume 7, p. 9–34.

Uschold M. & Gruninger M. (2004). Ontologies and semantics for seamless connectivity. *ACM SIGMod Record*, 33(4), 58–64.

Uszkoreit H. (2002). New chances for deep linguistic processing. In *In Proceedings of COLING 2002* : Association for Computational Linguistics (ACL).

Valderrábanos A. S., Belskis A. & Moreno L. I. (2002). Multilingual terminology extraction and validation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Spain.

van Huyssteen G. B. & Verhoeven B. (2014). A taxonomy for Afrikaans and Dutch compounds. *Proceedings of the First Workshop on Computational Approaches to Compound Analysis*, p.31.

Vapnik V. N. (1995). *The Nature of Statistical Learning Theory*. New York, NY, USA : Springer-Verlag New York, Inc.

Velardi P. & Sclano F. (2007). Termextractor : a web application to learn the common terminology of interest groups and research communities. In *7ème Conférence" Terminologie et intelligence artificielle"*, p. 85–94.

Virpioja O. K. S. & Lagus L. L. K. (2010). Semi-supervised extensions to Morfessor Baseline. In Kurimo *et al.* (2010a), p. 30–34.

Virpioja S., Smit P., Grönroos S.-A., Kurimo M. *et al.* (2013). *Morfessor 2.0 : Python implementation and extensions for Morfessor Baseline*. Rapport interne.

Visa S. & Ralescu A. (2005). Issues in mining imbalanced data sets-a review paper. In *Proceedings of the sixteen midwest artificial intelligence and cognitive science conference*, p. 67–73 : sn.

Vivaldi J., Màrquez L. & Rodríguez H. (2001). Improving term extraction by system combination using boosting. In *Proceedings of European Conference on Machine Learning (ECML'01).*

Vivaldi J. & Rodríguez H. (2007). Evaluation of terms and term extraction systems : A practical approach. *Terminology*, 13(2), 225–248.

von Humboldt W. (1822). *Ueber das Entstehen der grammatischen Formen, und ihren Einfluss auf die Ideenentwicklung, reprinted in : über die Sprache : Ausgewählte Schriften (1985).* Jürgen Trabant. München : Deutscher Taschenbuch Verlag.

von Humboldt W. (1836). *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwickelung des Menschengeschlechtes*. Reprinted.

P. Vossen, Ed. (1998). *EuroWordNet : a multilingual database with lexical semantic networks*. Norwell, MA, USA : Kluwer Academic Publishers.

Wan X., Yang J. & Xiao J. (2007). Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Annual Meeting-Association for Computational Linguistics*, volume 45, p. 552.

Wang A., Kan M.-Y., Andrade D., Onishi T. & Ishikawa K. (2013). Chinese informal word normalization : an experimental study. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing, IJCNLP*, volume 13, p. 127–135.

Wang W., Yaman S., Precoda K., Richey C. & Raymond G. (2011). Detection of agreement and disagreement in broadcast conversations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, p. 374–378 : Association for Computational Linguistics.

Watts D. & Strogatz S. (1998). Collective dynamics of « small-world » networks. *Nature*, (393), 440–442.

Weaver W. (1948). Science and complexity. *American scientist*, 36(4), 536–544.

Weissenhofer P. (1995). *Conceptology in Terminology Theory, Semantics and Wordformation : A Morpho-conceptually Based Approach to Classification as Exemplified by the English Baseball Terminology*. IITF-series / International Institute for Terminology Research : IITF-series. TermNet, International Network for Terminology.

Welch B. L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika*, p. 28–35.

Wermter J. (2009). *Collocation and term extraction using linguistically enhanced statistical methods.* PhD thesis, Friedrich-Schiller-Universität, Jena, Germany.

Wermter J. & Hahn U. (2004). Collocation extraction based on modifiability statistics. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04 : Association for Computational Linguistics.

Wermter J. & Hahn U. (2005). Paradigmatic modifiability statistics for the extraction of complex multi-word terms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 843–850 : Association for Computational Linguistics.

WERMTER J. & HAHN U. (2006). You can't beat frequency (unless you use linguistic knowledge) : a qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 785–792 : Association for Computational Linguistics.

WHORF B. L. (1940). Science and linguistics.

WILBUR W. J. & SIROTKIN K. (1992). The automatic identification of stop words. *Journal of information science*, 18(1), 45–55.

WILLIAMS G. (2001). Sur les caractéristiques de la collocation. *Traitement Automatique du Langage Naturel (TALN) 2001*, 2, 9–16.

WITTEN I. H. & FRANK E. (2005). *Data Mining : Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

WRIGHT S. (1997). In *Handbook of Terminology Management*, number vol. 1 in Basic Aspects of Terminology Management, chapter Term Selection : The Initial Phase of Terminology Management. J. Benjamins.

WÜSTER E. (1931). Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik [International language standardization, especially within electrotechnics]. *Dusseldorf, Germany : VDI-Verlag*.

WÜSTER E. & BAUER L. (1979). *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*. Springer.

XANTHOS A. (2008). *Apprentissage automatique de la morphologie : le cas des structures racine-schème*. Sciences pour la communication. Peter Lang.

XU J., FRASER A. & WEISCHEDEL R. (2002). Empirical studies in strategies for Arabic retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 269–274 : ACM.

XU Y., RINGLSTETTER C. & GOEBEL R. (2009). A continuum-based approach for tightness analysis of chinese semantic units. In O. KWONG, Ed., *PACLIC*, p. 569–578 : City University of Hong Kong Press.

YANG H. & CALLAN J. (2009). A metric-based framework for automatic taxonomy induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and*

the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 1-Volume 1, p. 271–279 : Association for Computational Linguistics.

Yang Y. (2003). *Discretization for naive-bayes learning*. PhD thesis, Monash University.

Zawada B. E. & Swanepoel P. (1994). On the empirical adequacy of terminological concept theories : The case for prototype theory. *Terminology*, 1(2), 253–275.

Zhang X., Song Y. & Fang A. (2010). Term recognition using conditional random fields. In *Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on*, p. 1–6 : IEEE.

Zhang Z., Iria J., Brewster C. & Ciravegna F. (2008). A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco*.

Zhikov V., Takamura H. & Okumura M. (2010). An efficient algorithm for unsupervised word segmentation with branching entropy and MDL. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, p. 832–842 : Association for Computational Linguistics.

Zou F., Wang F. L., Deng X., Han S. & Wang L. S. (2006). Automatic construction of Chinese stop word list. In *Proceedings of the 5th WSEAS international conference on Applied computer science*, p. 1010–1015.