EXTENSION DU DICTIONNAIRE ÉLECTRONIQUE GREC DE TERMES BOURSIERS À PARTIR D'UN CORPUS SPÉCIALISÉ

Evangelia FISTA
Eleni TZIAFA
Université Aristote de Thessalonique

Tita KYRIACOPOULOU
Université Paris-Est Marne-la-Vallée

Abstract (En): The problem of unknown words (words not recognized by automated language analysis systems) is one of great importance for Natural Language Processing (NLP). In this paper, we consider as unknown those words which are not recognized in a given corpus, the corpus of Greek Stock Exchange texts, since they are not included in the general dictionaries and terminologies for the Greek language, as used by the NLP systems. In this special domain, it is a critical issue, due to the rapid development of technical and scientific languages. In order to expand our resources, especially as regards new domains, it is necessary to acquire new terms as soon as possible and include them among the existing resources. Many of the unknown words are actually neologisms, and also loan words, written in the Latin or Greek alphabets, words in hybrid form (both Latin and Greek alphabet), proper names, abbreviations, incorrectly spelled words, words without accents etc. The aim of this work is to study the unknown words comprised in the Stock Exchange corpus (CoBourse) and to make them part of the dictionary of Stock Exchange terms. In this paper, we focus especially on simple words, as multiword expressions require a different approach.

Résumé (Fr) : Un des problèmes essentiels en traitement automatique des langues (TAL) est celui des mots non reconnus par les systèmes d'analyse automatique, quelle que soit l'approche adoptée, linguistique, statistique ou hybride. Dans ce travail, nous définissons comme mots inconnus les mots non reconnus dans un corpus donné, précisément dans le corpus boursier grec, car ils ne sont pas répertoriés dans les dictionnaires électroniques généraux et terminologiques du grec auxquels ont recours les systèmes de TAL. Dans un domaine de spécialité, ce problème s'avère l'un des plus délicats du fait de l'évolution rapide des langues techniques ou scientifiques. Pour l'enrichissement de ces ressources et afin d'exploiter de nouveaux domaines, il est nécessaire d'acquérir rapidement la nouvelle terminologie et de mettre à jour les ressources existantes.

Parmi les mots inconnus, figurent des néologismes, mais aussi des mots étrangers, transcrits en grec ou en alphabet latin, des mots en écriture hybride (caractères grecs et latins), des noms propres, des sigles, des mots mal orthographiés et en principe des mots non accentués. Ces mots non reconnus freinent l'analyse automatique des textes boursiers. L'objet du présent travail est l'étude de mots inconnus du corpus boursier (CoBourse), ce qui nous permettra l'ajout de termes néologiques dans le dictionnaire électronique des termes du domaine boursier. Nous nous limitons aux mots simples, les unités polylexicales demandant une approche de traitement différente. À partir de données extraites, nous proposons des heuristiques pour l'annotation semi-automatique des mots inconnus détectés à l'aide du système Unitex, afin de les intégrer dans le dictionnaire de termes boursiers.

Keywords: Unknown words; Expansion of electronic dictionaries; Specialized corpus; Specialized language

Mots-clés : mots inconnus ; extension des dictionnaires électroniques ; corpus spécialisé ; langue de spécialité

Introduction

Dans le domaine des langues de spécialité, les progrès scientifique, technique et culturel ont pour effet la création incessante d'un nombre important de termes nouveaux qui reflètent toutes les composantes essentielles de la spécialité (DINCĂ, 2009). Selon Lerat (1993 : 132), le rôle des néologismes est d'enrichir et de moderniser le vocabulaire pour les besoins de dénomination, d'expression et de communication. Un des problèmes essentiels en traitement automatique des langues de spécialité est celui des mots inconnus, c'est-à-dire des mots qui ne sont pas répertoriés dans les dictionnaires électroniques de termes auxquels ont recours des systèmes d'analyse comme Unitex (PAUMIER, 2003).

Des recherches ont été menées sur le problème de ces mots étiquetés comme inconnus et sur la typologie des néologismes identifiés à partir des corpus statiques ou dynamiques (cf. DISTER & FAIRON, 2004 ; WALTHER & SAGOT, 2011 ; BLANCAFORT *et al.*, 2010). Parmi les mots inconnus, figurent des néologismes, mais aussi des mots étrangers, transcrits en grec ou en alphabet latin, des mots en écriture hybride (caractères grecs et latins), des noms propres, des sigles, des mots mal orthographiés et en principe des mots non accentués.

Notre approche s'inscrit dans les travaux de la linguistique de corpus, proposant une étude de mots inconnus à partir de données (corpus et dictionnaires) particulières (CARTONI, 2006 ; DISTER & FAIRON, 2004 ; MAUREL, 2004). Il est évident que, même si les néologismes *ψευτοάνοδος* (fausse hausse), *φουσκοεταιρεία* (société bulle), *ξενόχαρτο* (action étrangère), *μακροπρόβλεψη* (prévision macroéconomique)[1], sont absents de nos dictionnaires et par conséquent ne sont pas identifiés par Unitex, ils expriment des réalités bien nouvelles et communes. De plus, leur introduction dans un dictionnaire électronique nécessite pour chaque entrée une validation par un spécialiste, et cette validation n'est pas entièrement automatisable[2]. Dans la section 1, nous présentons brièvement le corpus et les ressources lexicales terminologiques existantes pour la langue grecque dans le domaine de la bourse, c'est-à-dire le corpus sur lequel notre travail repose et le dictionnaire électronique des termes boursiers.

En section 2 nous décrivons les mots repérés comme inconnus par Unitex ainsi que les solutions semi-automatiques de filtrage de ces mots inconnus. Nous présentons ensuite, en section 3, les termes néologiques du domaine boursier dans le but de les intégrer dans le dictionnaire des termes de la bourse. Puis, nous concluons avec la section 4 en présentant quelques applications vers lesquelles nous nous engagerons.

[1] La traduction est mot-à-mot et il faut noter qu'en grec le sens de ces mots composés est transparent.

[2] Le rôle de l'informatique est certes d'automatiser certaines tâches, mais en l'occurrence, et pour des dictionnaires électroniques d'une couverture étendue, comme il en existe pour le français, des décisions d'inclusion de mots nouveaux prises de façon entièrement automatique auraient manifestement un taux d'erreur excessif (LAPORTE, 2009).

1. Ressources textuelles et lexicales pour le grec

1.1. Le corpus du domaine boursier (CoBourse)

Certes, le grec est l'une des langues de faible présence sur Internet et par conséquent les textes disponibles numérisés ainsi que les ressources textuelles et lexicales existantes sont d'une couverture relativement limitée, étant donné que l'anglais s'impose de plus en plus comme la *lingua franca* des marchés internationaux. Quand il s'agit d'une langue de spécialité, ces ressources sont beaucoup moins étendues. Notre corpus est constitué[3] de textes spécialisés du domaine boursier tirés de sources et de registres très divers dont la publication s'échelonne sur 11 ans : de 1999 à 2010, une période marquée par deux crises majeures en Grèce, la crise boursière et la crise de la dette. Ces événements ont permis l'émergence d'un grand nombre de néologismes. Comme corpus de référence nous avons utilisé le Corpus de Textes Grecs (*Σώμα Ελληνικών Κειμένων – ΣΕΚ*, environ 30 millions de mots) (GOUTSOS, 2003) et celui formé du journal « Ta Nea » (environ 120 millions de mots)[4]. Le premier est un corpus équilibré, tandis que l'autre est un corpus de source unique. Notre corpus est relativement de grande taille puisqu'il comporte environ 19 millions de mots grecs et il se compose de quatre sous-corpus explicités ci-après. Les textes sont complets et authentiques.

1.2. La structure du corpus boursier

Le sous-corpus A est constitué de messages publiés dans les débats publics dans deux forums sur Internet, tous deux, consacrés à la bourse.

Le sous-corpus B provient de textes journalistiques, numérisés et couvre la période 1999-2000. Il a été complété par des articles sous format électronique de 2000 à 2010, écrits dans le même registre de langue.

Le sous-corpus C provient du site de la Bourse d'Athènes et contient des avis, des rapports annuels et des articles parus en 2000. Le sous-corpus C pourrait constituer une base pour une étude plus approfondie des textes parallèles, puisque les textes inclus sont accompagnés de leurs traductions en anglais.

Le sous-corpus D contient des textes académiques essentiellement axés sur les marchés monétaires et les marchés boursiers dérivés, fournis à partir de modules universitaires. De plus ont été utilisées des thèses de troisième cycle et de doctorats, disponibles en ligne.

1.3. Dictionnaire boursier des termes grecs

À ce jour, le dictionnaire boursier des termes grecs simples et composés comprend 71.717 formes différentes classées sous 9.526 lemmes. Les analyses portent sur un corpus de 18.800.000 occurrences, tous textes confondus. Les termes du domaine boursier étudiés sont soit des mots simples, soit des unités polylexicales (KYRIACOPOULOU & TZIAFA, 2011)[5].

[3] À l'aide de *WordSmith* (SCOTT, 2011).

[4] Ce corpus nous a été fourni par Cédric Fairon, Université Catholique de Louvain.

[5] À noter que dans le dictionnaire des termes boursiers que nous avons constitué, les unités polylexicales représentent 80% des entrées.

Tous ces termes spécialisés sont codés et formalisés dans le dictionnaire électronique du grec et par conséquent des informations comme le genre (m pour masculin, f pour féminin et n pour neutre) le nombre (*s* pour singulier et *p* pour pluriel) ou le cas (*N* pour Nominatif, *G* pour Génitif, *A* pour Accusatif, *V* pour Vocatif dans les exemples ci-après) sont traitées. Voici un extrait du dictionnaire électronique grec des termes simples et composés :

spread,spread.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
spreads,spread.N+[Eco]:Nnp:Gnp:Anp:Vnp
stock option,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
stock option,.N+[Eco]:Nfs:Gfs:Afs:Vfs:Nfp:Gfp:Afp:Vfp
Α ομόλογο,.N+[Eco]:Nns:Ans:Vns
Α ομολόγου,Α ομόλογο.N+[Eco]:Gns[6]

2. Mots inconnus dans le CoBourse et requêtes de filtrage

2. 1. Mots inconnus

Le prétraitement de notre corpus a été effectué par le système Unitex. La première étape a été la segmentation automatique du CoBourse de 19.000.000 mots. Nous avons ainsi obtenu trois fichiers comportant : a. la liste des mots reconnus, b. la liste des mots non reconnus et c. la liste des mots (tokens) par fréquence et par ordre alphabétique[7]. Parmi les 14.450.259 occurrences, Unitex a identifié 151.753 mots simples et 23.746 mots composés à partir de 34.548 occurrences. La liste des mots inconnus qui contenait 156.493 mots différents comportait des coquilles [*κατα**κρ**ύλα (<κατρακύλα)* (forte baisse)], des fautes d'orthographe [*κατανάλ**λ**ωση (<κατανάλωση)* (consommation)], des mots non accentués ou mal accentués [*απομειωση (< απομε**ί**ωση)* (réduction)], des noms propres *(Marfin, Vernikos)*, des abréviations [*cds (<* crédit default swaps)], des mots étrangers souvent en caractères grecs ou en écriture hybride – caractères grecs et latins [*split, haircut* (marge de sécurité), ***sort****άρισμα* (action de vendre)] et des termes néologiques. Plus précisément, nous avons constaté la présence des mots dérivés par suffixation [*χαρ**τάκι*** (petite action)] et également des mots construits à l'aide de préfixes soudés [***υπερ****απορρόφηση* (absorption excessive), ***ψευτο****άνοδος* (fausse hausse)]. La distribution des mots inconnus est représentée à la Figure 1.

[6] Où N indique le nom, A indique l'adjectif et [Eco] désigne le trait sémantique *économie*.
[7] Il faut noter que le grand nombre de mots inconnus est dû au sous-corpus A qui, rappelons-le, est constitué de messages publiés dans des débats publics dans deux forums sur internet, consacrés à la bourse. Ce genre de textes présente des similitudes avec le discours oral : de nombreux mots grecs transcrits en alphabet latin (*greeklish*), des abréviations, des coquilles et des fautes d'orthographe.

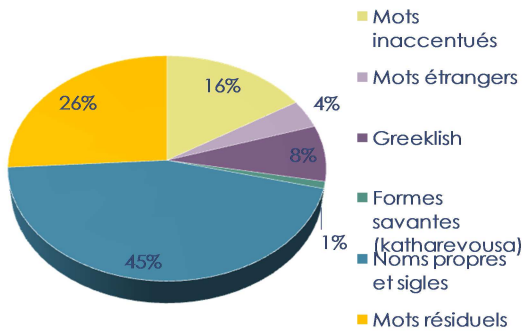

Figure 1 : Distribution des mots inconnus

Parmi ces mots inconnus figurent, comme nous avons déjà mentionné, des néologismes qui pourraient être des termes néologiques candidats. C'est pourquoi nous avons procédé au filtrage du bruit pour diminuer le nombre de mots inconnus et faciliter l'identification de « vrais » termes néologiques.

2.2. Requêtes de filtrage

À partir du fichier des mots inconnus qui contenait 156.493 mots simples, nous avons implémenté des applications appropriées afin de traiter de manière différente les mots inconnus de diverses origines et repérer les termes néologiques attestés dans le corpus.

Il est connu que les noms propres sont fortement représentés dans les textes. Spriet et al. (1996) a démontré que, parmi un pourcentage de 6% d'erreurs relevées dans un corpus donné, 58% des mots inconnus étaient des noms propres. Afin de reconnaitre les noms propres, nous avons appliqué au corpus le graphe (cf. Figure 2) de reconnaissance automatique des noms propres du grec, élaboré par Mavropoulos (2012).

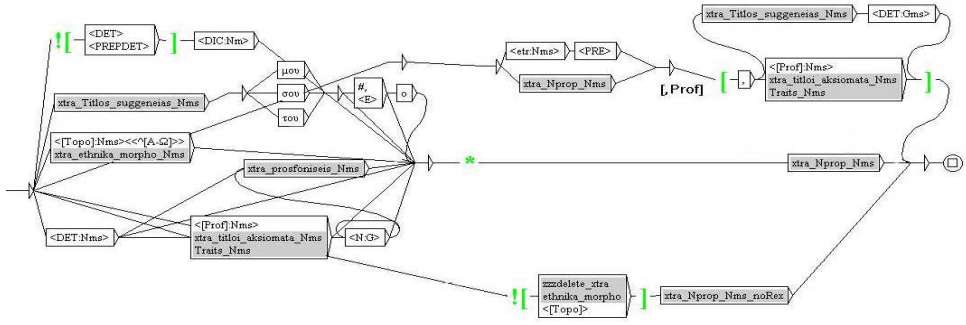

Figure 2 : Exemple d'un graphe de reconnaissance de noms propres (MAVROPOULOS, 2012)

De la même façon, nous avons extrait les abréviations du corpus, créant un graphe qui reconnait automatiquement des sigles et des acronymes des mots simples commençant par au moins une lettre majuscule. Ainsi, nous avons reconnu environ 70.000 noms propres et abréviations, ce qui représente 45% de mots inconnus.

Par la suite, nous avons reconnu les mots non accentués, ce qui représentait 16% de mots inconnus (24.675 formes). Pour le faire, nous avons inclus dans le dictionnaire du grec moderne (DELAF) les formes de tous les mots non accentués, augmentant ainsi sa taille de 2.029.797 à 3.858.936 formes fléchies[8].

Un grand nombre des mots restants étaient des mots écrits en alphabet latin, c'est-à-dire des emprunts (*buy, hedging*), des mots en greeklish (*metoxi/*action, *agora/*marché) ou en graphie hybride (*shortάρισμα*)[9]. 62% des termes anglais se réfèrent essentiellement à des outils économiques se terminant en –*ing* comme par exemple *hedging, rehedging, bookbuilding, churning, decoupling, scalping.* Pour distinguer les mots anglais des autres mots écrits en alphabet latin, nous avons utilisé le dictionnaire anglais DELAF[10]. Ensuite, nous avons supprimé les mots écrits en greeklish grâce à un script perl.

Il faut noter aussi que des noms et des adjectifs anglais ont servi de base pour la construction de verbes grecs en *-άρω* comme en attestent les exemples : *short* > *shortαρω*, *option*> *optionαρω*.

On obtient ainsi une liste de noms, verbes et adjectifs. Pour leur reconnaissance, nous avons créé deux graphes, un pour les verbes et l'autre pour les noms et adjectifs. Le graphe des verbes décrit une liste de 5.867 terminaisons caractéristiques, basée sur le fichier contenant la description des classes morphologiques, utilisé par le programme GFF[11].

À ce stade, dans la liste des mots non reconnus restaient des noms, mais également des adjectifs et des participes passés[12]. Nous avons identifié les participes passés par l'application d'un graphe qui contenait les terminaisons caractéristiques des participes passés du grec moderne, mais aussi celles de la langue savante[13] ce qui nous a amenés à en extraire 2.287 participes passés.

Les noms et les adjectifs ont été reconnus par un graphe (cf. Figure 3) qui décrit les suffixes et les deuxièmes composants les plus fréquents dans la formation des termes nominaux : *-ότητα, -ποίηση, -λογία, -σμα, -ισμός*, etc.

[8] Pourtant il n'y a pas dédoublement des entrées car, il y avait déjà des mots monosyllabiques sans accent (conformément aux règles d'accentuation du grec moderne) grâce à un script perl.

[9] Ces mots sont détectés par le système comme des erreurs, mais comme ils appartiennent au langage particulier utilisé, au langage de messages aux blogs, sms etc, ils méritent une étude plus approfondie.

[10] Nous avons regroupé les emprunts reconnus dans un fichier à part pour une étude ultérieure. Nous constatons aussi la présence d'un grand nombre de mots anglais relevant du vocabulaire de la langue générale.

[11] Programme de Génération de Formes Fléchies.

[12] Les mots mal orthographiés n'ont pas été traités et feront l'objet d'une étude ultérieure, basée sur leur degré de similitude avec des mots existants.

[13] Beaucoup de participes passés de la langue savante sont largement utilisés dans le grec moderne.

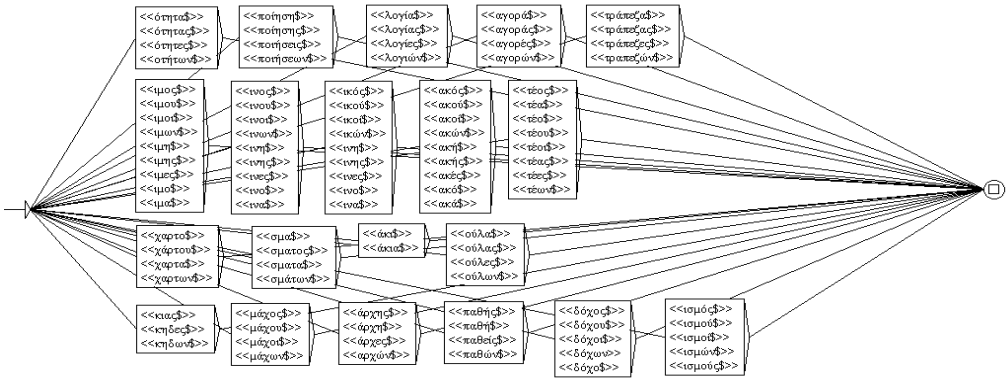

Figure 3 : Graphe de suffixes nominaux et de deuxièmes composants pour l'extraction des termes candidats.

3. Formes des termes néologiques

L'analyse automatique du corpus et malgré l'utilisation du dictionnaire DELAF[14] et du dictionnaire boursier (71.717 formes fléchies) s'est heurtée à plusieurs problèmes relevant pour la plupart des spécificités de la langue boursière, comme des emprunts à l'anglais, des symboles, des abréviations, des noms propres, des néologismes morphologiques de création récente[15].

3.1. Emprunts à l'anglais

Selon Anastasiadi-Symeonidi (1986 : 61), « un vocabulaire de spécialité d'une autre langue peut être « source de néologismes dans un vocabulaire de spécialité ». La création néologique, soit sous forme d'emprunts soit sous forme de calques est étroitement liée à l'internationalisation de la science et de la technologie ».

Dans le marché boursier grec, dominé par le marché boursier anglo-américain, on rencontre des néonymes[16] comme *split*, mais également leurs dérivés suffixaux ou flexionnels du grec comme *split* < *σπλιτάρω ή splitάρω* (faire un split), *short* < *shortarω*, *shortάρω*, *sortάρω* et/ou *σορτάρω* (être vendeur), *σορτάρισμα*, *sortαρισμα*, *σορτάκιας*[17], *hedge* < *hedgarω* et/ou *χετζάρω* (effectuer une opération de couverture), *long* < *longarω*, *λογκάρω* et/ou *λονγκάρω* (être acheteur). Les exemples cités ci-dessus montrent que de nombreux néologismes apparaissent sous plusieurs formes graphiques.

3.2. Termes néologiques de création récente

La création lexicale a pour moteur le besoin de désigner des objets nouveaux, des phénomènes nouveaux, des idées nouvelles, des concepts nouveaux dans un monde en évolution constante. De nouveaux termes, recensés dans le corpus

[14] Dictionnaire électronique des formes fléchies du vocabulaire général du grec moderne, élaboré suivant la méthodologie du LADL (COURTOIS & SILBERZTEIN, 1990).
[15] Nous n'étudions pas les néologismes sémantiques, comme par exemple le nom *φούσκα* (bulle).
[16] Selon Kocourek (1991 : 174) le terme de « néonyme » est utilisé par Cellard et Sommaelt (1979) mais Rondeau (1984) réserve à la néologie terminologique la dénomination de « néonymie».
[17] Le terme *σορτάκιας* est formé sur le terme anglais *short selling* et désigne la personne qui vend des actions qui ne possède pas mais qui espère « qu'elles baissent pour tirer un bénéfice important » (MATHIOPOULOS, 1999 : 164).

boursier (CoBourse), illustrent la dynamique néologique toujours en relation étroite avec les changements socio-économiques, culturels qui ont marqué ce domaine jusqu'en 1999. Citons par exemple : *ελδάρχης, ελδεάρχης, ελδετζής*[18] (propriétaire d'une société de type SICAV), *δεικτοβαρής* (relatif à une obligation indexée). Il s'agit de formes qui ont une existence éphémère puisqu'elles apparaissent en fonction des événements occasionnels liés à des métiers particuliers et par conséquent leur longévité n'est pas assurée.

3.3. Termes dérivés par suffixation

Nous citons comme particulièrement productifs les suffixes *-ότητα, -ποίηση, -λογία, -σμα, -ισμός, -τράπεζα, -χαρτο : διακυμανσιμότητα* (volatilité ou instabilité), *αποπαγιοποίηση* (déconsolidation), *βιοχαλκολογία* (discussion autour de la société Viohalco), *λογκάρισμα* (action d'achat), *μαρφινοτράπεζα* (Marfin Banque).

Le suffixe nominal *-σμα* entre dans la formation de nombreux substantifs neutres qui expriment une action ou le résultat de cette action : *λογκάρισμα* (action d'être acheteur), *σορτάρισμα* (action d'être vendeur), *μανατζάρισμα* (manœuvre de bourse). Les suffixes *-ικός, -ιμος* sont productifs dans la formation des adjectifs à base de noms ou de verbes : *αξιογραφικός* (relatif aux titres), *απορρυθμιστικός* (qui peut provoquer une déréglementation), *αποτιμήσιμος* (quantifiable).

Nous avons remarqué que, par exemple, l'acronyme *ΕΛΔΕ* (SICAV) a servi de base pour la formation du mot *ελδάρχης* (ANASTASIADI-SYMEONIDI, 1986 : 54 et 241). On note plusieurs termes, dérivés ou composés, issus de noms propres tels que *βγενό**χαρτο**, βαγγελό**χαρτο**, πανουσό**χαρτο**, παπαελληνό**χαρτο*** (< Βγενόπουλος/Vgenopoulos, Βαγγέλης/Vangelis, Πανούσης/Panousis, Παπαέλληνας/Papaellinas). Le formant ***-χαρτο*** (*χαρτί* = papier, *l'action* au jargon boursier) est très productif et il a servi pour l'identification des néologismes (cf. Figure 2).

3.4. Termes dérivés par préfixation

Après avoir examiné la liste de mots candidats (3.836 mots), nous avons eu à résoudre un autre problème : un grand nombre de mots n'a pas été reconnu parce que ces mots étaient préfixés. Les préfixes formant les termes nouveaux sont, pour la plupart, d'origine grecque (préfixes savants). On a recensé un grand nombre de préfixes dont les plus productifs[19] sont :

υπερ- (hyper) : *υπερτράπεζα* (banque énorme)
επανα- [επι-+ανα-] (re) : *επαναπώληση* (revente)
υπο- (hypo) : *υποαπόδοση* (baisse de revenu)
ενδο- (endo) : *ενδοσυνεδριακά* (pendant la séance boursière)
εξω- (exo/extra) : *εξωχρηματιστηριακός* (hors bourse)
ιδιο- (idio) : *ιδιοχρηματοδοτούμενος* (autofinancé)
αυτο- (auto) : *αυτοπαλινδρόμηση* (auto-régression)
βραχυ- (brachy) : *βραχυμεσοπρόθεσμος* (court et moyen terme)
ημι- (hemi) *: ημιδιακύμανση* (semi fluctuation)
μικρο- (micro) : *μικροάνοδος* (petite hausse)

[18] Les suffixes *-τζής, -άρχης* forment des noms de métier dans un registre de langue populaire.
[19] À titre indicatif, nous avons relevé 104 formes préfixées en *υπερ-* (hyper), 60 formes en *επανα-*, 32 formes en *υπο-* (hypo), 24 formes en *ψιλο-* (psilo), 22 formes en *μικρο-* (micro), 15 formes en *νεο-* (neo) et 13 formes en *αυτο-* (auto).

μακρο- (macro) : *μακροπρόβλεψη* (prévision à long terme)
μεγαλο- (megalo) : *μεγαλομετοχοαπατεώνας* (gros-actionnaire-escroc)[20]
μεσο- (meso) : *μεσομακροπρόθεσμος* (ayant un impact à moyen et long terme)
νεο- (neo) : *νεοεισαγόμενος* (nouvellement introduit)
πολυ- (poly) : *πολυδιασπορά* (grande dispersion)
πρωτο- *(proto) : πρωτοεισάγω* (introduire pour la première fois)
ψευτο- και ψευδο- (pseudo-) : *ψευτοάνοδος* (fausse hausse)
ψιλο- (psilo) : *ψιλοκλειδώνω* (en train de clôturer)

Comme résultat final nous avons obtenu deux listes de mots, la première contenant 2.158 formes verbales et la deuxième 1.678 formes nominales. Ces listes sont beaucoup moins volumineuses que les listes initiales des mots inconnus. Cela s'explique d'une part par la taille relativement limitée[21] du corpus boursier et d'autre part par le fait que les mots appartenant à la langue générale ont été exclus. Après avoir fait une validation manuelle, nous avons identifié au total 1.087 termes néologiques (413 formes verbales et 674 formes nominales) ce qui représente une augmentation de 40% de la taille du dictionnaire électronique grec de termes boursiers (cf. Figure 4).

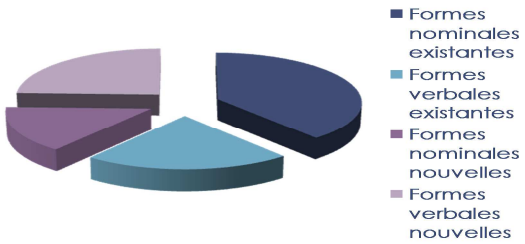

Figure 4 : Enrichissement du dictionnaire boursier de mots simples

4. Conclusions et perspectives

L'objectif de notre travail était d'identifier les termes spécifiques au corpus utilisé et qui ne sont pas présents dans les dictionnaires électroniques généraux et terminologiques du système Unitex. Nous avons développé diverses stratégies afin de repérer les termes néologiques dans le but d'enrichir le dictionnaire de termes boursiers. Nous avons extrait 1.087 termes néologiques parmi les mots inconnus (156.493 mots), en nous appuyant particulièrement sur leurs suffixes. Les procédures néologiques les plus significatives de notre corpus étaient l'emprunt et la dérivation par suffixation et préfixation.

Notre recherche démontre que le vocabulaire spécialisé de la bourse en grec moderne contient non seulement des noms et des adjectifs (674 formes de termes néologiques attestées), mais aussi un nombre important des verbes (413 formes verbales attestées). Ces informations seront intégrées à un outil d'annotation et de lemmatisation[22] qui est en cours de développement. Cet outil permettra

[20] Traduction littérale.
[21] Si on le compare avec de grands corpus annoncés (FERRARESI et al. 2008, 2010, BARONI et al. 2009, POMIKÁLEK 2009).
[22] La lemmatisation consiste à associer un lemme à chaque mot du texte. Si le mot ne peut pas être lemmatisé (nombre, mot étranger, mot inconnu), aucune information ne lui est associée.

la liaison et l'interopérabilité des dictionnaires DELAF grecs avec d'autres systèmes comme Antconc (ANTHONY, 2011) et Wordsmith Tools (SCOTT, 2011).

Bibliographie

ANASTASIADI-SYMEONIDI Anna (1986), *Η Νεολογία στην Κοινή Νεοελληνική. Epistimoniki Epetirida Filosofikis Scholis.* Thessaloniki : Aristotle University of Thessaloniki.

ANTHONY Laurence (2011), *AntConc* (Version 3.2.2) [Computer Software], Tokyo, Japan: Waseda University, http://www.antlab.sci.waseda.ac.jp.

BARONI Marco ; BERNARDINI Silvia ; FERRARESI Adriano ; ZANCHETTA Eros (2009), The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora, in : *Language Resources and Evaluation* 43(3), p. 209-226.

BLANCAFORT Helena ; RECOURCE Gaëlle ; COUTO Javier ; SAGOT Benoît ; STERN Rosa ; TEYSSOU Denis (2010), Traitement des inconnus : une approche systématique de l'incomplétude lexicale, in : *TALN 2010*, Montréal, Canada.

CARTONI Bruno (2006), Constance et variabilité de l'incomplétude lexicale, in : *RECITAL 2006*, Leuven, Belgium, TALN 2006.

CELLARD Jacques ; SOMMAELT Micheline (1979), *500 mots nouveaux définis et expliqués*, Paris-Gembloux, Duculot.

DINCA Daniela (2009), La néologie et ses mécanismes de création lexicale, in *Analele Universităţii din Craiova, Seria Lingvistică*, nr. 1-2, 2009, p. 79-91.

DISTER Anne ; FAIRON Cédrick (2004), Extension des ressources lexicales grâce à un corpus dynamique, *Lexicometrica*.

FAIRON Cédrick ; COURTOIS Blandine (2000), Extension de la couverture lexicale des dictionnaires électroniques du LADL à l'aide de GlossaNet, in : *Actes du Colloque JADT 2000 : 5es Journées Internationales d'Analyse Statistique des Données Textuelles*, Lausanne.

FERRARESI Adriano ; ZANCHETTA Eros ; BARONI Marco ; BERNARDINI Silvia (2008), Introducing and evaluating ukWaC, a very large web-derived corpus of English, in : EVERT Stefan, KILGARRIFF Adam & SHAROFF Serge (éd.) *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?* Marrakech.

FERRARESI Adriano ; BERNARDINI Silvia ; PICCI Giovanni ; BARONI Marco (2010), Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation, in : XIAO Richard (éd.), *Using Corpora in Contrastive and Translation Studies*, Newcastle, Cambridge Scholars Publishing.

GOUTSOS Dionysis (2010), The Corpus of Greek Texts: A reference corpus for Modern Greek, in : *Corpora 5 (1)*, p. 29-44.

KOCOUREK Rostislav (1991), *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. Wiesbaden, Brandsletter.

KYRIACOPOULOU Tita ; TZIAFA Eleni (2011), Dictionnaires électroniques et terminologie : le cas du vocabulaire « boursier », 9èmes *Journées Scientifiques du réseau Lexicologie, Terminologie, Traduction*, 15-16 septembre 2011, Université Paris 13.

LAPORTE Éric (2009), Concordanciers et flexion automatique, in *Cahiers de Lexicologie*, 94 (1), p. 91-106.

LERAT Pierre (1993), *Les langues spécialisées*, Paris, PUF.

MATHIEU Yvette Yannick ; GROSS Gaston ; FOUQUERE Christophe (1998), Vers une extraction automatique des néologismes, in : *Cahiers de Lexicologie*, n° 72, p. 199-208.

MATHIOPOULOS Haris (1999), *Μικρό Εγχειρίδιο του Επενδυτή*, Athens, Estia.

MAUREL Denis (2004), Les mots inconnus sont-ils des noms propres?, in : *Actes des JADT* 2004.

MAVROPOULOS Athanasios (2012), *Ένα σύστημα αυτόματης ανάλυσης κειμένων της Νέας Ελληνικής. Μέθοδοι αναπαράστασης των κύριων ονομάτων προσώπων*, Thessaloniki, Aristotle University of Thessaloniki, thèse de doctorat.

PAUMIER Sébastien (2003), *Unitex. Manuel d'utilisation*, Paris, Université Paris-Est Marne-la-Vallée, http://igm.univ-mlv.fr/~unitex/UnitexManual.pdf.

POMIKÁLEK Jan ; RYCHLÝ Pavel ; KILGARRIFF Adam (2009), Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics, in : *Special Issue of Research in Computing Science Vol 41*, http://pics.cicling.org/2009/RCS-41/003-014.pdf.

RONDEAU Guy (1984), *Introduction à la terminologie*, Québec, Gaetan Morin.

SCOTT Mike (2011), *WordSmith Tools version 6*, Liverpool, Lexical Analysis Software.

SPRIET Thierry ; BECHET Frédéric ; EL-BEZE Marc ; De LOUPY Claude ; KHOURI Liliane (1996), Traitement automatique des mots inconnus, in : *Proceedings of TALN'96*, Marseille, p. 170-179.

WALTHER Géraldine ; SAGOT Benoît (2011), Problèmes d'intégration morphologique d'emprunts d'origine anglaise en français, in : *Proceedings of the 30th Lexis and Grammar Conference*, Nicosia, Cyprus.

ÉCHO DES ÉTUDES ROMANES

Revue semestrielle de linguistique et littératures romanes

Publié par l'Institut d'études romanes
de la Faculté des Lettres
de l'Université de Bohême du Sud,
České Budějovice

ISSN : 1801-0865 (Print)
1804-8358 (Online)

Numéro du volume : Vol. IX / Num. 2
2013