Dictionnaires électroniques et terminologie: le cas du vocabulaire "boursier"

Tita Kyriacopoulou, Eleni Tziafa

Tita Kyriacopoulou,
Département,
Université Paris-Est Marne-la-Vallée,
Paris,
France,
tita@univ-mlv.fr

Eleni Tziafa,
Département de Langue et Littérature françaises,
Université Aristote de Thessalonique,
Thessalonique
Grèce,
etziafa@lit.auth.gr

Dictionnaires électroniques et terminologie : le cas du vocabulaire « boursier »

Résumé

Notre objectif est de présenter le dictionnaire électronique de la bourse, élaboré pour le grec, et d'exposer les problèmes rencontrés ainsi que les solutions adoptées. Ce travail s'avère réalisable grâce à l'évolution de la technologie et surtout grâce à l'alignement des textes et la création des mémoires de traduction. Cette recherche est effectuée au sein de l'Institut Gaspard Monge et de l'Université Aristote de Thessalonique. Ce travail fait partie d'une thèse de doctorat co-financée par l'Union Européenne (Fonds Social Européen) et par l'État Grec (Cadre de Référence Stratégique National 2007-2013), dans le cadre du programme « Heracleitus II. Investissement dans la société de la connaissance / Éducation et Formation Tout au Long de la Vie ».

Mots clés : vocabulaire boursier; dictionnaires électroniques; terminologie.

Introduction

Il est bien connu que les dictionnaires électroniques sont codifiés et formalisés pour qu'ils soient utilisés dans des applications informatiques. Malheureusement les données spécialisées sont souvent traitées, dans le meilleur des cas, dans des bases de données qui sont destinées à une utilisation humaine (cf. IATE par exemple, InterActive Terminology for Europe: www.iate.europa.eu.). Le paradoxe est que le TAL (Traitement Automatique des Langues) vise l'exploitation des documents techniques, alors que les dictionnaires électroniques sont pauvres en terminologie. En effet, si l'enrichissement des dictionnaires généraux est une opération « bien maîtrisée » par les linguistes aujourd'hui, il n'en va pas de même pour les dictionnaires spécifiques dits « terminologiques ». En

TAL, nous avons besoin, en fonction des applications, de dictionnaires (généraux et terminologiques, monolingues et bilingues) formalisés, renseignés par des informations linguistiques et des grammaires (monolingues et de transfert).

Notre objectif est de présenter le dictionnaire électronique de la bourse, élaboré pour le grec, et d'exposer les problèmes rencontrés ainsi que les solutions adoptées. Les termes ont été sélectionnés à partir d'un corpus constitué pour cette étude qui comporte 18.805.328 mots et couvre la période de 1999 à 2010. Ce travail qui est effectué au sein de l'Institut Gaspard Monge et de l'Université Aristote de Thessalonique s'avère réalisable grâce à l'évolution de la technologie et surtout grâce aux outils d'extraction d'information.

1 La langue boursière

La langue des affaires et des finances a toujours suscité l'intérêt, puisque le mouvement des indices boursiers peut être un indicateur, un « baromètre » de la tendance générale de l'économie. Selon Behr et al. (2007) « L'activité boursière fascine le grand public qui peut, lui aussi, accéder aux jeux de la Bourse. L'autre aspect et objet de fascination est le vocabulaire boursier perçu comme un code secret plus que comme une langue de spécialité ». De plus la terminologie touche le problème de la « néologie » avec les nouveaux termes qui naissent tous les jours en fonction de l'évolution des sciences et des technologies. Les tentatives de récupérer automatiquement les bases de données terminologiques demandent un effort considérable de correction manuelle.

Les termes techniques sont souvent des néologismes (*επιλεκτική χρεωκοπία* / défaut sélectif*),* abréviations (*A&O* / Avril & Octobre), ou des symboles (AA+) et touchent « la problématique » des entités nommées dans la mesure où il s'agit de segments « semi-figés ». On observe aussi des cas de nominalisation (*πτώση των τιμών* / baisse des prix*),* de métaphore ou de métonymie [*Η Σοφοκλέους καταγράφει αλλεπάλληλα ρεκόρ* / Sofokleous (la Bourse) enregistre des hausses successives, ou *Η βουτιά του Δείκτη* / le plongeon de l'indice]. Les linguistes ont du mal à connaître tous les emplois des termes spécialisés pour pouvoir les décrire et les formaliser ; en outre ces mots présentent des problèmes particuliers et sont en évolution constante.

2 Le corpus boursier

Selon Kocourek (1991), il y a « deux facteurs irréductibles de la langue de spécialité que sont termes et textes ». Les textes sont la source des données pour l'analyse de tous les niveaux de langue. Ainsi, la méthodologie de linguistique de corpus est nécessaire. Pour la présente étude, nous avons procédé à la construction d'un corpus de textes boursiers, appartenant à des genres différents.

Il s'agit de textes authentiques, c'est-à-dire de textes non traduits. Nous n'avons pas inclus dans notre corpus les textes de la Commission Européenne qui, en général, sont des textes traduits

de l'anglais ou du français. La période que couvre notre corpus est de 1999 à 2010. Cette période a été marquée en Grèce par deux crises majeures, la crise de la Bourse et la crise de la dette. Ces années turbulentes nous donnent aussi l'occasion d'étudier la langue du marché boursier dans ses hauts et ses bas, ou en période « baissière » et « haussière ». Notre corpus final est constitué de quatre sous-corpus (cf. Tableau 1) de textes grecs boursiers et comporte 18 millions de mots. A noter aussi que notre corpus de référence provient d'une source unique : le journal « Ta Nea » (environ 120 millions de mots).

Par rapport aux grands corpus existants (Ferraresi et al. 2008, 2010, Baroni et al. 2009, Pomikalek 2009), notre corpus peut apparaître petit, mais compte tenu du domaine choisi et de la langue, il s'agit d'un corpus relativement important. En effet le Corpus de Textes Grecs (Goutsos 2003) est composé de 30 millions de mots et le Hellenic National Corpus (Hatzigeorgiu et al. 2000) de 47.000.000 mots. La construction d'un grand corpus spécialisé reste une tâche difficile et la disponibilité affecte gravement la taille du corpus.

Nous présentons ci-après les quatre sous-corpus qui constituent notre corpus final, en précisant chaque fois la période et le nombre de mots :

Sous-corpus	*Source*	*Période*	*Nombre de textes*	*Mots*
A. Messages sur Internet (forums de discussion sur la Bourse)	Fora Internet (http://www.neoforum.gr, http://www. capital.gr)	2009-2010 2009-2010	46.912 messages	4.316.172
B. Articles journalistiques, imprimés ou sous forme électronique	Journaux financiers sur la Bourse Sites web sur la Bourse	1999–2000 2000-2010	12.407 articles	5.031.919
C. Communiqués de presse, rapports annuels de la Bourse d' Athènes, documents officiels	Site web de la Bourse d' Athènes (http://www.ase.gr)	2000-2010	18.199 communiqués de presse, 21 rapports annuels, 47 documents officiels	5.612.945
D. Discours académique	Polycopiés, thèses et articles	2002-2010	350 textes	3.547.347

Tableau 1: Structure du corpus

Le sous-corpus A est constitué de messages provenant de discussions publiques de deux forums d'Internet, tous deux dédiés à la bourse. Ce genre de forum est apparu en Grèce ces trois dernières années et relève du discours de communication d'entreprise professionnelle qui, selon Daniushina (2010) effectue « une fonction instrumentale-persuasion ».

Le sous-corpus B provient de textes journalistiques, numérisés à partir des journaux pour la période 1999-2000, complété par des articles disponibles sous format électronique de 2000 à 2010,

du même niveau langagier. Il s'agit du discours des médias d'affaires qui détermine « une fonction informative-polémique ».

Le sous-corpus C provient du site officiel de la Bourse d'Athènes. Il contient des avis, rapports annuels et articles datant de l'année 2000 ce qui correspond au discours écrit (« fonction régulatrice ») des documents officiels (correspondance d'affaires internes et externes, documents corporatifs, règlements et chartes des entreprises, statuts, etc.). Concernant le C et les communiqués de presse en particulier, il faut noter que nous disposions au départ de textes représentant 29 millions de mots mais nous en avons extrait un échantillon aléatoire de 5 millions de mots afin de maintenir un corpus équilibré puisque notre but est aussi de procéder à des comparaisons de style etc.

Enfin, le sous-corpus D contient des textes académiques fournis par les supports de cours universitaires se référant aux marchés monétaires et aux marchés des actions et de produits dérivés. Par ailleurs, des mémoires de troisième cycle et des thèses de doctorat, disponibles en ligne, ont été utilisés. Ce registre correspond au discours de formation et au discours académique et il constitue une « fonction éducative ».

2.1 Classification des textes selon leur genre

Dans cette section, nous étudierions deux points en particulier : d'une part la représentation équilibrée de différents registres de la langue spécialisée (ex. jargon, langue savante etc.), et d'autre part les particularités dues aux origines de la langue boursière. Gavioli (2002) pose le problème de la représentativité de petits corpus et des critères utilisés notamment dans la conception de corpus spécialisés de petite taille. Alors que la discussion sur la représentativité du corpus a été un problème majeur et aussi controversé dans la linguistique de corpus au cours des 20 dernières années, par la suite elle s'est concentrée principalement sur les problèmes concernant la construction de grands corpus, comme le Bank of English Corpus (Sinclair 1991) et le BNC (Burnard 1998). Les corpus spécialisés ont tendance à appartenir à un domaine ou à un genre spécialisé, mais ils doivent être diversifiés incluant un large éventail de types pour être vraiment représentatifs d'une variété particulière de la langue.

De ce fait, l'objectif de cette étude est la représentation adéquate d'un langage spécialisé, comportant différents registres, dans le domaine de la bourse, l'analyse des modèles linguistiques à travers les différents registres étant d'une importance primordiale pour la description linguistique d'une langue de spécialité. Un tel corpus nous donne l'occasion d'étudier et de comparer les variétés linguistiques des discours formels et informels allant de « l'argot de transaction » à la « prose académique ». La collecte et la création d'outils structurés représentent un grand défi technologique, notamment pour les langues avec un petit nombre de locuteurs natifs comme la langue grecque.

L'existence de différents registres est encore plus justifiée en ce qui concerne les bourses de valeurs. Il s'agit d'un langage professionnel et, en même temps, il peut être un langage scientifique.

Dans de nombreux pays, les transactions ont réellement été menées dans les rues, comme la rue Quincampoix à Paris ou dans Exchange Alley à Londres ou encore dans des cafés. Un café a été également la première bourse d'Athènes (nommée Bella Grecia). Ainsi, au cœur de ce discours, il y a « la langue du marché », une langue qui, en grec, est synonyme du mot argot. Au fil du temps, cependant, comme les produits boursiers sont devenus de plus en plus complexes la langue utilisée a suivi cette évolution.

De façon générale nous pouvons dire que pour « toutes les langues » il y a un « jargon spécialisé » utilisé exclusivement par des experts. Mais dans le domaine boursier, experts et non experts parlent la même langue et animent des débats sur les mêmes questions. De plus, en raison de la crise actuelle, tous les mots, considérés comme spécialisés auparavant, ont désormais « envahi » la langue générale. Actuellement les grecs prennent conscience de la terminologie financière, car ils sont quotidiennement préoccupés par la crise et les termes comme *défaut sélectif, temporaire, contrôlé, ordonné, désordonné, restreint* ou *organisé* deviennent familiers.

3 Extraction des unités lexicales et fréquence

Notre corpus a été analysé à l'aide de l'outil Wordsmith Tools (Scott 2008) et le corpus de référence. Les quatre sous-corpus comportent les mêmes unités lexicales (*αγορά /* marché*, ευρώ /* euro, *μετοχές /* actions) ce qui montre qu'il s'agit bien du même domaine technique. En revanche, la variation de fréquence entre les sous-corpus (cf. tableau 3) peut s'expliquer par la différence de registre de langue. Le tableau 2 présente les mots clefs extraits automatiquement (à l'aide de la fonction *Keywords* de l'outil Wordsmith[1]) de notre corpus.

A		B		C		D	
ΦΙΛΕ	ami	ΜΕΤΟΧΏΝ	actions	ΕΥΡΏ	euro	ΤΙΜΉ	prix
ΈΓΡΑΨΕ	il/elle a écrit	ΑΝΑΚΟΊΝΩΣΗ	annonce	ΕΤΑΙΡΊΑΣ	société	ΑΓΟΡΆΣ	marché
ΕΥΡΏ	euro	ΕΥΡΏ	euro	ΜΕΤΟΧΏΝ	actions	ΜΕΤΟΧΏΝ	actions
ΕΤΕ	NBG	ΑΓΟΡΆ	marché	ΤΡΑΠΕΖΑ	banque	ΔΕΊΚΤΗ	indice
SPREAD	écart	ΙΔΊΩΝ	capital en actions	ΑΕ	SA	ΕΤΑΙΡΊΑΣ	société
SHORT	short	ΜΕΤΟΧΉ	action	ΔΙΟΙΚΗΤΙΚΟΎ	administrative	ΕΛΈΓΧΟΥ	contrôle
ΔΝΤ (IMF)	FMI	ΔΙΑΒΑΣΤΕ	lisez	ΣΥΜΜΕΤΟΧΩΝ	holding	ΠΑΡΑΓΏΓΩΝ	dérivés
ΤΡΆΠΕΖΕΣ	Banques	ΔΗΜΟΦΙΛΈΣΤΕΡΕΣ	les plus populaires	ΣΥΝΈΛΕΥΣΗ	réunion	ΚΙΝΔΎΝΟΥ	risque
ΟΜΌΛΟΓΑ	bonds	ΕΤΑΙΡΊΑΣ	société	EURO	euro	ΕΠΕΝΔΥΤΗΣ	investisseur
ΕΥΧΑΡΙΣΤΩ	merci	ΡΥΘΜΙΖΌΜΕΝΗΣ	régulée	ΚΕΦΑΛΑΊΟΥ	capital	ΧΑΡΤΟΦΥΛΑΚΊΟΥ	portefeuille

[1] Cet outil calcule la fréquence de chaque mot relativement à sa fréquence dans un corpus de langue générale. Cette comparaison se fait grâce au test de signification du X^2 (Chi2) avec la correction Yates pour une table 2x2 d'une part et au test Log-Likelihood d'autre part (Dunning 1993).

ΑΜΚ	augmentation du capital	ΕΠΙΦΆΝΕΙΑ	surface	ΑΝΑΚΟΙΝΏΝΕΙ	il/elle annonce	ΕΠΙΧΕΊΡΗΣΗΣ	entreprise
ΓΔ	indice boursier	ΚΈΡΔΗ	profit	ΟΜΊΛΟΥ	groupe	ΣΜΕ	contrats à terme
MARKET	marché	ΓΝΩΣΤΟΠΟΊΗΣΗ	notification	ΜΕΤΟΧΙΚΟΎ	actionnaires	ΑΞΊΑ	valeur
ΒΆΣΗΣ	base	ΒΑΝΚ	banque	ΣΥΜΒΟΥΛΊΟΥ	l'assemblée	ΣΥΝΑΛΛΑΓΏΝ	transactions
ΤΙΜΗ	prix	ΝΑΥΤΕΜΠΟΡΙΚΗ	Naftemporiki	ΆΡΘΡΟ	article	ΕΣΩΤΕΡΙΚΟΎ	intérieur
ΚΑΛΗΜΈΡΑ	bonjour	ΠΛΗΡΟΦΟΡΊΑΣ	information	ΨΉΦΟΥ	vote	ΚΕΦΆΛΑΙΟ	capital
ΕΤΑΙΡΊΑΣ	société	ΕΚΑΤ	million	ΜΕΡΊΣΜΑΤΟΣ	dividende	GRANGER	Granger
CAPITAL	capital	ΟΙΚΟΝΟΜΙΑ	économie	ΕΛΛΑΔΟΣ	Grèce	ΑΝΆΛΥΣΗ	analyse
LONG	longue	ΣΥΝΈΛΕΥΣΗΣ	réunion	ΧΡΉΣΗΣ	exercice	ΑΠΟΔΌΣΕΙΣ	rendement
ΦΙΛΟΙ	amis	WALL	Wall Street	ΣΥΝΈΛΕΥΣΗΣ	réunion	ΔΙΚΑΙΏΜΑΤΟΣ	droit

Tableau 2: Mots clefs extraits du corpus boursier à l'aide de Wordsmith Tools

Dans la suite nous présentons le tableau avec les fréquences des unités lexicales extraites par sous-corpus. Nous pouvons constater par exemple que les mots du vocabulaire courant comme *είμαι* (être), *έχω* (avoir) etc. sont plus fréquents dans le sous-corpus A (rappelons qu'il s'agit des messages pris sur Internet et en particulier des forums de discussion sur la Bourse).

A		B		C		D	
είναι 40397	Il/elle est	είναι 21700	il/elle est	ευρώ 29272	euro	είναι 30182	il/elle est
έχει 10702	Il/elle a	μετοχών 20223	actions	μετοχών 19364	actions	έχει 7784	il/elle a
ευρώ 10405	euro	ευρώ 16983	euro	εταιρείας 15310	société	τιμή 7281	prix
έγραψε 6689	il/elle a écrit	ανακοίνωση 9798	annonce	είναι 14072	il/elle est	αγοράς 6351	marché
Ελλάδα 6377	Grèce	αγορά 9716	marché	μετοχές 11264	actions	πρέπει 6024	il doit
έχουν 5651	ils/elles ont	εκατ 8187	millions	εκατ 10604	millions	μπορεί 5958	il/elle peut
πρέπει 4486	il doit	Αγορά 7739	marché	αύξηση 9541	augmentation	μετοχών 5782	actions
αγορά 4107	marché	εταιρεία 6954	société	Συμβουλίου 8797	conseil	αγορά 5573	marché
μονάδες 4082	unités	εταιρείας 6567	société	Διοικητικού 8758	d'administration	περίπτωση 4031	cas
μπορεί 4012	il/elle peut	κέρδη 6261	profit	ΤΡΑΠΕΖΑ 8407	banque	δείκτη 3930	index
τράπεζες 4011	banques	μετοχή 6205	action	Αθηνών 8305	Athènes	αξία 3774	valeur
σήμερα 3874	aujourd'hui	έχουν 6184	ils/elles ont	Γενική 8158	Général	συναλλαγών 3408	transactions
τώρα 3363	maintenant	μετοχές 5928	actions	Συνέλευση 8148	Réunion	σχέση 3157	relation
εκατ 3267	millions	ιδίων 5812	capital en actions	έχει 8071	il /elle a	επενδυτές 2995	investisseurs
όχι 3169	no	capital 5689	capital	Εταιρίας 7790	Société	στοιχεία 2815	éléments
ΕΤΕ 3066	NBG	Ελλάδα 5425	Grèce	ΑΕ 7647	SA	βάση 2812	base
φιλε 3038	ami	δισ 5362	milliard	κεφαλαίου 7453	capital	ελέγχου 2810	controle
μετοχές 2968	actions	εργασίας 5194	travail	αξίας 7252	valeur	κινδύνου 2794	risque
spread 2728	écart	Capital 5160	Capital	Ομίλου 7187	Groupe	επιχειρήσεις 2650	sociétés
Κάνει 2706	il (elle) fait	αύξηση 5145	augmentation	άρθρο 7164	article	ευρώ 2502	euro

Tableau 3: Fréquences des unités lexicales

4 Vers la constitution d'un dictionnaire électronique

Un dictionnaire électronique doit contenir toutes les informations nécessaires à l'automatisation de la production des formes fléchies. Les adverbes et les conjonctions sont en principe invariables et ne présentent donc pas de grandes difficultés dans un dictionnaire électronique. En revanche, les

verbes, les noms et les adjectifs doivent être décrits morphologiquement du point de vue de leur flexion (genre, nombre et cas pour les noms et les adjectifs). Pour cela, il faut créer à partir d'une liste de noms, d'adjectifs, de verbes et de noms composés toutes les formes fléchies contenant les informations nécessaires à la reconnaissance automatique des noms (simples et composés), des adjectifs et des verbes dans les textes.

Ainsi, par exemple, pour la flexion des noms et adjectifs, nous avons pris en compte toutes les caractéristiques de la langue grecque, à savoir :

- les trois genres (masculin, féminin, neutre)

- les deux nombres (singulier, pluriel)

- les quatre cas -quatre au singulier et quatre au pluriel- (nominatif, génitif, accusatif, vocatif)

- le déplacement de l'accent

Les principales caractéristiques exposées ici sont extraites des grammaires traditionnelles (Holton et al. 2000, Triandaphyllidis 2005). Rappelons qu'en grec moderne, il ne reste plus qu'un seul accent, qui est tonique. Il peut se positionner sur une des trois syllabes de la fin d'un mot (sur la finale : *αγορά* / marché, sur la pénultième : *απάτη* / malhonnêteté, et sur l'antépénultième : *κυβέρνηση* / gouvernement). Le déplacement de l'accent dans un mot, lors de la flexion, revient à la suppression de l'accent existant, puis à l'ajout d'un autre accent à un endroit différent (*απόκομμα-αποκόμματος* / coupon).

- les variantes

Elles sont nombreuses en grec moderne et elles sont soit graphiques (voir phonétiques) (*χρεοκοπία-χρεωκοπία* / *défaut*), soit flexionnelles. Dans ce dernier cas, il s'agit de variantes soit d'accentuation (*αντιπρόσωπος* / représentant – *αντιπρόσωπου*$_{génitif}$ + *αντιπροσώπου*$_{génitif}$), soit de terminaison (*αποτελεσματικότητα* / efficacité - *αποτελεσματικότητας*$_{génitif}$ + *αποτελεσματικότητος*$_{génitif}$), soit des deux (*κεφαλαιοποίηση* / capitalisation - *κεφαλαιοποίησης*$_{géniif}$ + *κεφαλαιοποιήσεως*$_{génitif}$).

Pour la flexion des noms composés, nous nous sommes basées sur les noms simples et les adjectifs et nous avons défini des filtres pour éliminer les formes non acceptables ou agrammaticales. Par ailleurs, signalons que certains composants peuvent apparaître entre guillemets (*εταιρεία-"φάντασμα"* / société « fantôme ») et d'autres encore proviennent de mots étrangers (ils sont d'ailleurs écrits en caractères latins) : *offshore εταιρεία* / société offshore.

4.1 Les termes boursiers

Les termes du domaine boursier étudiés sont soit, des mots simples soit, des expressions polylexicales. Sous le terme « expressions polylexicales », on entend les expressions figées (*kick the*

bucket / casser sa pipe), les noms composés (*annual general meeting / assemblée générale anuelle*), les expressions à verbes supports (*make a decision / prendre une décision*), etc. (Sag et al. 2002). L'interprétation de ces unités lexicales n'est pas du tout évidente à cause de leur opacité ni pour les locuteurs natifs ni pour les apprenants d'une langue étrangère et encore moins pour une machine. À noter que dans le dictionnaire du vocabulaire boursier que nous avons élaboré au cours de notre recherche, les unités polylexicales représentent 90% des entrées.

Parmi les 9.634 termes que nous avons recensé, il y a des mots simples, comme *αποϋλοποίηση /* dématérialisation, des mots composés, comme *λευκός ιππότης /* chevalier blanc, des expressions figées, comme *κλείνω τον αέρα /* arrêter les ventes à découvert, et des collocations (semi- figées) comme *αθέτηση συμφωνίας /* rupture du contrat.

La complexité de la structure interne est variable. Elle peut être composée de deux mots (« *μετοχή δώρο* » bonus stock / action donnée en prime, « *αγορά καλάθι* » basket purchase / achat à un prix global) ou encore de cinq mots (« *διαδικασία αίτησης σύγκλησης συνέλευσης μετόχων* » / procédure de l'assemblée générale des actionnaires) et elle peut comporter des lettres ou des symboles (« *A&O* » / April & October, « *P/E* » / Price per Earnings, « *T+1* », « *T+2* », « *ημέρα T+1* » / jourT+1).

Les structures morphosyntaxiques des termes sélectionnés dépassent les 200 (cf. figure 1 : Liste de structures). Elles se composent des verbes, noms, déterminants, adverbes, prépositions et particules. Voici quelques exemples chiffrés :

Termes	Structure	Pourcentage
2325	Adjectif+Nom (AN)	30%
1721	Nom + Nom (NN)	22%
643	Nom + Adjectif + Nom (NAN)	8,4%

Tableau 4: Les structures des termes

4.2 Dictionnaire électronique grec des termes simples et composés

Le dictionnaire électronique est une base de données morphologique. Chaque entrée est identifiée par une graphie unique. La structure de l'entrée consiste en :

- un symbole de partie du discours,
- une classe flexionnelle pour chaque composant,
- un filtre de restriction de formes.

Les exemples ci-dessous illustrent la structure des entrées du dictionnaire :

split.N3051,N+[Eco]
spread.N3051,N+[Eco]
άγγι"γμα.N365,N+[Eco]
stock.N305 option.N305,N+[Eco]
stock.N205 option.N205,N+[Eco]
A.A5 B.A5 μετοχή.N251,N+[Eco]
A.A5 ομόλο"γο.N301,N+[Eco]
δημοσιονομικός.A10,A
reverse.A5 repo.N3051 επί.PREP μετοχών.N251,N,-GP4

où N3051, N365, N305, N205, N251, N301 indiquent des noms, respectivement de classe flexionnelle 3051, 365, 305, 205, 251, 301 ; A5 et A10 indiquent des adjectifs de classe flexionnelle 5 et 10 et PREP une préposition ; -GP4 désigne le filtre de restriction de formes. Sur ces exemples, on voit que les entrées figurent comme dans les dictionnaires usuels, sous leur forme canonique attestée :

- noms au nominatif singulier, masculin, féminin ou neutre selon leur genre,
- adjectifs au nominatif masculin singulier.

La base de données que représente le dictionnaire électronique permet d'obtenir le dictionnaire des formes fléchies grâce à des traitements informatiques. Le programme élaboré par S. Mrabti (cf. Kyriacopoulou et al. 2003) fléchit les mots simples et les mots composés. Le programme tient compte des déplacements de l'accent dans les mots ainsi que de l'existence de variantes pour telle ou telle forme.

Le dictionnaire des mots fléchis (45.830 formes) présente le résultat de la flexion au format des dictionnaires Unitex (Paumier 2002). À chaque forme du mot fléchi sont associées ses caractéristiques grammaticales, comme la catégorie grammaticale du mot composé (nom, adjectif, etc.), le genre (m pour masculin, f pour féminin et n pour neutre) de celui-ci et son cas (N pour Nominatif, G pour Génitif, A pour Accusatif, V pour Vocatif). Le programme détermine automatiquement les caractéristiques du mot composé grâce à celles de ses composants et grâce aux filtres utilisés.

Voici un extrait du dictionnaire électronique grec des termes simples et composés :

splits,split.N+[Eco]:Nnp:Gnp:Anp:Vnp
spread,spread.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
spreads,spread.N+[Eco]:Nnp:Gnp:Anp:Vnp
άγγιγμα,άγγιγμα.N+[Eco]:Nns:Ans:Vns
αγγίγματος,άγγιγμα.N+[Eco]:Gns

αγγίγματα,άγγιγμα.N+[Eco]:Nnp:Anp:Vnp
αγγιγμάτων,άγγιγμα.N+[Eco]:Gnp
stock option,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
stock option,.N+[Eco]:Nfs:Gfs:Afs:Vfs:Nfp:Gfp:Afp:Vfp
time sharing,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
time-sharing,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
venture capital,.N+[Eco]:Nns:Gns:Ans:Vns:Nnp:Gnp:Anp:Vnp
AB μετοχή,.N+[Eco]:Nfs:Afs:Vfs
AB μετοχής,AB μετοχή.N+[Eco]:Gfs
AB μετοχές,AB μετοχή.N+[Eco]:Nfp:Afp:Vfp
AB μετοχών,AB μετοχή.N+[Eco]:Gfp
A B μετοχή,.N+[Eco]:Nfs:Afs:Vfs
A B μετοχής,A B μετοχή.N+[Eco]:Gfs
A B μετοχές,A B μετοχή.N+[Eco]:Nfp:Afp:Vfp
A B μετοχών,A B μετοχή.N+[Eco]:Gfp
A ομόλογο,.N+[Eco]:Nns:Ans:Vns
A ομολόγου,A ομόλογο.N+[Eco]:Gns

Conclusion

Nous venons de présenter le premier dictionnaire électronique du grec moderne dans le domaine boursier. Il contient environ 10.000 termes. Il a été construit manuellement dans un premier temps et enrichi ensuite de façon semi-automatique. Les termes recensés ont été extraits d'un corpus de textes authentiques ce qui nous a permis de disposer en plus d'un corpus grec monolingue de 19.000.000 de mots. Les ressources linguistiques ainsi constituées peuvent enrichir considérablement les systèmes en TAL (Traitement Automatique des Langues) concernant la langue grecque. Mais il faut aussi considérer que la terminologie touche les nouvelles technologies et cela a des conséquences sur son évolution trop rapide parfois. Ainsi des termes utilisés il y a encore quelques années comme *ανταλλαγή επιτοκίου απέριττης βανίλιας* (plain vanilla interest rate swap) ont disparu. Cela pose bien évidemment des problèmes de maintenance et de mises à jour des dictionnaires existants. Il y a aussi des termes qui sont entrés dans notre vie quotidienne comme le terme *Σοφοκλέους/Sofokleous* (Sofokleous est le nom de la rue où se situait le bâtiment de la bourse d'Athènes) qui désigne « la bourse d'Athènes » et qui est encore utilisé par les employés, même si les locaux de la Bourse ont déménagé.

Bibliographie

Baroni (M.), Bernardini (S.), Ferraresi (A.) et Zanchetta (E.), 2009 : « The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora », *Language Resources and Evaluation 43(3)* , pp. 209-226.

Behr (I.), Hentschel (D.), Kauffmann (M.) et Kern (A.) (éds.), 2008: *Langue, économie, entreprise. Le travail des mots*, 27-29 mars 2008, Paris, Presses Sorbonne Nouvelle.

Burnard (L.); Aston, (G.), 1998 : *The BNC handbook: exploring the British National Corpus*, Edinburgh, Edinburgh University Press.

Cowie (A. P.), 1998 : *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Daniushina (Y. V.), 2010 : « Business linguistics and business discourse », dans *Calidoscopio*, Vol. 8, No 3, derniere mise en jour le 15 decembre 2011, http://www.unisinos.br/revistas/index.php/calidoscopio/article/viewArticle/294.

Dunning (T.), 1993 : « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, Vol 19, No. 1, pp. 61-74.

Ferraresi (A.), Bernardini (S.), Picci (G.) et Baroni (M.), 2010 : « Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation » dans Xiao (R.) éd., *Using Corpora in Contrastive and Translation Studies*, Newcastle, Cambridge Scholars Publishing.

Ferraresi (A.), Zanchetta (E.), Baroni (M.) et Bernardini (S.) 2008 : « Introducing and evaluating ukWaC, a very large web-derived corpus of English », dans Evert (S.), Kilgarriff (A.) et Sharoff (S.), éd., *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?,* LREC 2008, Marrakech, pp. 45-54.

Gavioli (L.), 2002 : « Some thoughts on the problem of representing ESP through Small Corpora », dans Ketterman (B.) et Marko (G.), éd., *Teaching and learning by doing corpus analysis : proceedings of the Fourth International Conference on Teaching and Language Corpora,* Graz.

Giry-Schneider (J.), 1978 : *Les nominalisations en français. L'operateur « faire » dans le lexique*, Genéve, Droz.

Goutsos (D.), 2003: « Greek Text Corpus : Design and materialization », dans *Proceedings of 6th International Conference on Greek Linguistics*, University of Grete.

Gross (G.), 2003 : « Trois applications de la notion de verbe support ». *L'Information grammaticale* 59, 1993, 2003, pp. 16-22.

Hatzigeorgiou (N.), Spiliotopoulou (A.), Vacalopoulou (A.), Papakostopoulou (A.), Piperidis (S.), Gavriilidou (M.) and Karayanis (G.), 2000 : « Hellenic National Corpus (HNC): a Modern Greek corpus on the internet », dans *Proceedings of the 21st Annual Meeting of the Linguistics*

Department, School of Philology, Faculty of Philosophy, Aristotle University of Thessaloniki, May 2000, Thessaloniki, pp. 812-821.

Holton (D.), Mackridge (P.), Filippaki-Warburton (I.) 1997 : *Greek: A Comprehensive Grammar of the Modern Language,* London, Routledge.

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

Kyriacopoulou (T.), 2005 : *L'analyse automatique des textes écrits: Les cas du grec moderne*, Thessaloniki, University Studio Press.

Kyriacopoulou (T.), Mrabti (S.) et Yannacopoulou (A.), 2003 : «Le dictionnaire électronique des noms composés en grec moderne», *Lingvisticæ Investigationes 25:1,* Amsterdam/Philadelphia, John Benjamins, pp. 7-28.

Mejri (S.), 2008 : « Vers un dictionnaire électronique des séquences figées », dans Dotoli (G.), Papoff (G.), éd., dans *Du sens des mots. Le réseau sémantique du dictionnaire, Biblioteca della Ricerca,* pp. 117-129.

Paumier (S.), 2002 : *Manuel d'utilisation du logiciel Unitex*, IGM, Universit´e de Marne-la-Vallée, dernier mise en jour le 15 decembre 2011, http://www-igm.univ-mlv.fr/_unitex/manuelunitex.pdf.

Pomikalek (J.), Rychly (P.) et Kilgarriff (A.), 2009 : « Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics », dans *Special Issue of Research in Computing Science Vol 41*, derniere mise en jour le 15 decembre 2011, http://pics.cicling.org/2009/RCS-41/003-014.pdf.

Sag (I. A.), Baldwin (T.), Bond (F.), Copestake (A.), Flickinger (D.), 2002 : « Multiword Expressions: A Pain in the Neck for NLP », dans *Lecture notes in computer science*, Vol. 2276, pp. 1-15.

Scott (M.), 2008 : *WordSmith Tools version 5*, Liverpool, Lexical Analysis Software.

Sinclair (J. M.), 1991: *Corpus, concordance, collocation*, Oxford, Oxford University Press.

Triandaphyllidis (M.), 2005 : *Grammaire* du Grec *Moderne (langue démotique),* Thessaloniki, Institut d'études néo-helléniques.