# Extraction terminologique et corpus alignés Anglais Grec.

Tita Kyriacopoulou, Claude Martineau, Eleni Tziafa

**Extraction terminologique et corpus alignés Anglais Grec.**

Tita Kyriacopoulou[1], Claude Martineau[2], Eleni Tziafa[3]

Université Paris-Est Marne-la-Vallée Université Aristote de Thessalonique

**Abstract**

The objective of this research is the extraction of bilingual terminology from aligned texts. The alignment was performed using the program UNITEX, on texts from the domain of telecommunications in Greek (GR) and English language. In order to track the terms in the texts, local grammars have been used, taking into account the immediate context. These grammars contain several graphs that have been constructed at the Institut Gaspard Monge and the Aristotle University of Thessaloniki.

**Résumé**

L'objectif de notre recherche est d'extraire de la terminologie bilingue à parti des textes alignés. L'alignement s'opère avec le logiciel UNITEX à partir de textes du domaine des télécommunications grecs (GR) et anglais (EN). Pour le repérage des termes dans les textes nous avons utilisé des grammaires locales avec l'exploitation du contexte immédiat. Les grammaires utilisées comportent plusieurs graphes qui ont été élaborés au sein de l'Institut Gaspard Monge et de l'Université Aristote de Thessalonique.

**Mots clés :** extraction, terminologie, corpus alignés, grammaires locales

## 1. Introduction

Notre travail s'inscrit dans le cadre d'un projet de recherche ayant comme objectif le traitement automatique des langues et en particulier du grec et s'intéresse à l'extraction de terminologie bilingue dans un corpus technique. Notre objectif est d'explorer les nouvelles techniques de filtrage / formatage / recherche, en langue naturelle, pour faciliter l'accès à

[1] Institut Gaspard-Monge, Université Paris-Est Marne-La-Vallée tita@frl.auth.gr
[2] Institut Gaspard-Monge, Université Paris-Est Marne-La-Vallée Claude.Martineau@univ-mlv.fr
[3] Université Aristote de Thessalonique eltziafa@yahoo.gr

l'information multilingue. Pour cela nous nous servons des outils et des ressources linguistiques existantes ainsi que des corpus du domaine des télécommunications grecs (GR) et anglais (EN).

Après une présentation de la problématique d'extraction de la terminologie, nous nous penchons sur les problèmes liés à la typologie des termes techniques. Nous exposons aussi notre méthodologie ainsi que quelques résultats obtenus. En conclusion nous signalons quelques perspectives qu'ouvre ce travail.

**2. Le problème de la terminologie**

En TALN, nous avons besoin, en fonction des applications, de dictionnaires (généraux et terminologiques, monolingues et bilingues) formalisés, renseignés par des informations linguistiques (morphologiques, syntaxiques, sémantiques) et des grammaires (monolingues et de transfert). De plus, les dictionnaires doivent être aussi complets que possible et codés de façon systématique.

Les dictionnaires généraux ainsi que les grammaires sont faits par des linguistes qui en revanche rencontrent des difficultés pour constituer des dictionnaires terminologiques. 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.

De ce fait, toutes les tentatives de récupérer automatiquement les bases de données terminologiques ont échoué ou, dans le cas d'une récupération partielle, elles ont demandé un effort considérable de corrections manuelles. Or, il s'agit d'un problème majeur puisque les textes qu'on est censé de traiter entre autres en TALN sont des textes techniques et en plus les termes techniques posent énormément des problèmes de traduction (c'est concrètement la principale difficulté de la traduction technique).

Il faut noter aussi que les termes techniques sont souvent des acronymes, ou des symboles et touchent « la problématique » des entités nommées dans la mesure où il s'agit de segments « semi figés » qui incluent des protocoles, des noms de réseaux, des noms de société ou d'organisme nationaux et internationaux et il est parfois difficile de faire des listes exhaustives. Examinons un terme grec avec son acronyme et son équivalent en anglais :

| | | **Terme** | **Acronyme** |
|---|---|---|---|
| (i) | GR | ασύμμετρου ρυθμού σύνδεση | ΑΡΥΣ |
| (ii) | EN | Asymmetric Digital Subscriber Line | ADSL |

L'acronyme anglais ADSL est aussi souvent utilisé dans les textes grecs et par conséquent il faudra l'inclure dans les dictionnaires du grec. De plus les termes techniques présentent

beaucoup de variations. On trouve en effet le terme grec en question sous la forme : *ασύμμετρη ψηφιακή συνδρομητική γραμμή*.

**3. Recherche expérimentale**

Notre objectif est d'explorer les nouvelles technologies de filtrage / formatage / recherche, en langue naturelle, pour faciliter l'accès à l'information monolingue et multilingue tout en enrichissant des ressources lexicales existantes. 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 effet les linguistes ont du mal à connaître tous les emplois des termes techniques pour pouvoir les décrire et les formaliser ; de plus ces mots présentent des problèmes particuliers et ils sont en évolution constante. Notre recherche vise à tester l'aligneur intégré récemment dans UNITEX (cf. section 6) afin de constituer des textes parallèles, qui seront d'une grande utilité dans le domaine des systèmes de mémoire de traduction, et expérimenter le repérage automatique des termes techniques dans les documents ainsi exploités.

**4. Corpus**

Pour notre recherche nous avons vérifié les outils et les ressources linguistiques sur un corpus. Ce corpus a été publié par la Commission Nationale de Télécommunications et de Services Postales Helléniques (EETT). EETT est l'Autorité Nationale Hellénique qui régule et supervise le marché des communications électroniques. Il s'agit d'un corpus bilingue, écrit en grec et en anglais, de 449 kilooctets, comportant 16 706 +17 165 = 33 871 occurrences de mots. Il est constitué d'une vue d'ensemble du marché de télécommunications et de services postales pour l'année 2007. Il convient ici de distinguer les corpus réunissant des textes originaux dans plusieurs langues (corpus comparables) des corpus comportant des textes sources avec leur traduction dans une ou plusieurs langues (corpus parallèles). Notre corpus est conçu comme un corpus parallèle, mais grâce aux textes sources, d'une part, et aux textes cibles, d'autre part, il peut en principe faire office de corpus comparables.

La taille du corpus présente un inconvénient non négligeable, car le nombre d'occurrences des différents termes est trop réduit pour mener des études quantitatives. En effet, leur fréquence et leur emploi pragmatique varient sensiblement en fonction du type de texte. Pour ce qui est de la fréquence, le corpus est trop petit pour tirer des conclusions générales. Soulignons quand même ici que dans la majorité des cas les « ressources techniques et scientifiques » ainsi que la documentation technique monolingue et bilingue sont « propriétaires » et inaccessibles pour des raisons de confidentialité.

**5. Typologie des termes de télécommunications**

Les termes sont des mots simples, comme *ευρυζωνικότητα (broadband)* et *ευρυζωνικά (with broadband)* ou des mots composés, comme *κινητή τηλεφωνία (mobile telephony)*. Les termes composés sont soit des noms complexes, par exemple *ενεργοποιημένη γραμμή προεπιλογής φορέα (Activated Line of Carrier Pre-Selection),* soit des adverbes, *με κάρτα (by card), με συμβόλαιο (by contract), από σταθερό προς κινητό (from fixed to mobile phone),* soit des acronymes, comme *ΑΡΥΣ (ADSL)* et des combinaisons de noms, symboles et acronymes, comme *διείσδυση PSTN γραμμών και καναλιών ISDN (penetration of PSTN lines and ISDN channels), γραμμή PSTN (PSTN line), γραμμή ISDN BRA (ISDN BRA line), γραμμή xDSL εναλλακτικών παρόχων μέσω ΑΠΤΒ* (xDSL Line of Other Local Operators via LLU). Ils peuvent être aussi des combinaisons d'acronymes et de chiffres, comme ΑΡΥΣ *8Mbps (ADSL 8Mbps).* Nous présentons ci-dessous quelques exemples de termes composés sans les présenter exhaustivement :

**Adj + Nom**

*καρτοκινητή τηλεφωνία (prepaid mobile telephony), κινητή τηλεφωνία (mobile telephony), σταθερή τηλεφωνία (fixed telephony), φωνητικός τηλεφωνητής (voice mail), γραπτό μήνυμα (SMS)*

**Nom + Nom (au génitif)**

*καλάθι κλήσεων, ζώνη χρέωσης, συνδρομητής συμβολαίου, βήμα χρέωσης*

**Nom +Adj +Nom**

*συνδρομητής καρτοκινητής τηλεφωνίας (prepaid mobile subscriber), καλάθι χαμηλής χρήσης (low usage basket), καλάθι υψηλής χρήσης (high usage basket)*

**Nom + Nom + Adj + Nom**

*φορείς παροχής καθολικής υπηρεσίας (ΦΠΚΥ) (universal service providers – USP), τέλος εγκατάστασης νέας σύνδεσης (installation charge for a new connection)*

**Adj + Nom + Nom + Nom**

*ενεργοποιημένη γραμμή προεπιλογής φορέα (Activated Line of Carrier Pre-Selection)*

**Adj + Nom + Adj + Nom**

*αδεσμοποίητη πρόσβαση στον τοπικό βρόχο (Local Loop Unbundling, LLU)*

**Adj + Nom + Nom**

*ασύμμετρου ρυθμού σύνδεση* (ΑΡΥΣ, adaptation en grec du terme ADSL, *Asymmetric Digital Subscriber Line, ou quelquefois ασύμμετρη ψηφιακή συνδρομητική γραμμή*)

**6. Méthode et ressources utilisées**

Le but de notre travail est de développer un système d'extraction et de filtrage de termes techniques bilingues à partir des textes alignés GR-EN en réutilisant autant que possible les

ressources et outils existants, en particulier, ceux développés par l'équipe d'Informatique Linguistique de l'IGM[4], à l'Université de Marne-la-Vallée. Nous avons donc adopté une approche symbolique qui nous a également permis de valider la couverture et la pertinence de nos ressources, ainsi que de les adapter pour passer de la simple reconnaissance à l'extraction. Cette méthode a déjà été expérimentée dans le domaine des entités nommées et s'est révélée suffisamment efficace. Pour le repérage des termes techniques, nous avons fait appel au système Unitex (Paumier 2003). Unitex est un environnement de développement qui permet de construire des descriptions formalisées de grammaires et d'utiliser des ressources telles que des dictionnaires généraux et spécialisés. Tous les objets traités par Unitex sont ou peuvent être transformés en des transducteurs à nombre fini d'états (*RTN* en anglais).

Le corpus présenté ci-dessus était en état « brut ». Il a fallu le traiter et l'analyser avec UNITEX. L'alignement s'appuie sur des textes « prétraités » et découpés en phrase. Cette tâche difficile est effectuée grâce aux grammaires locales de « découpage » intégrées dans le logiciel. Nous présentons (figure1) la grammaire du grec moderne :

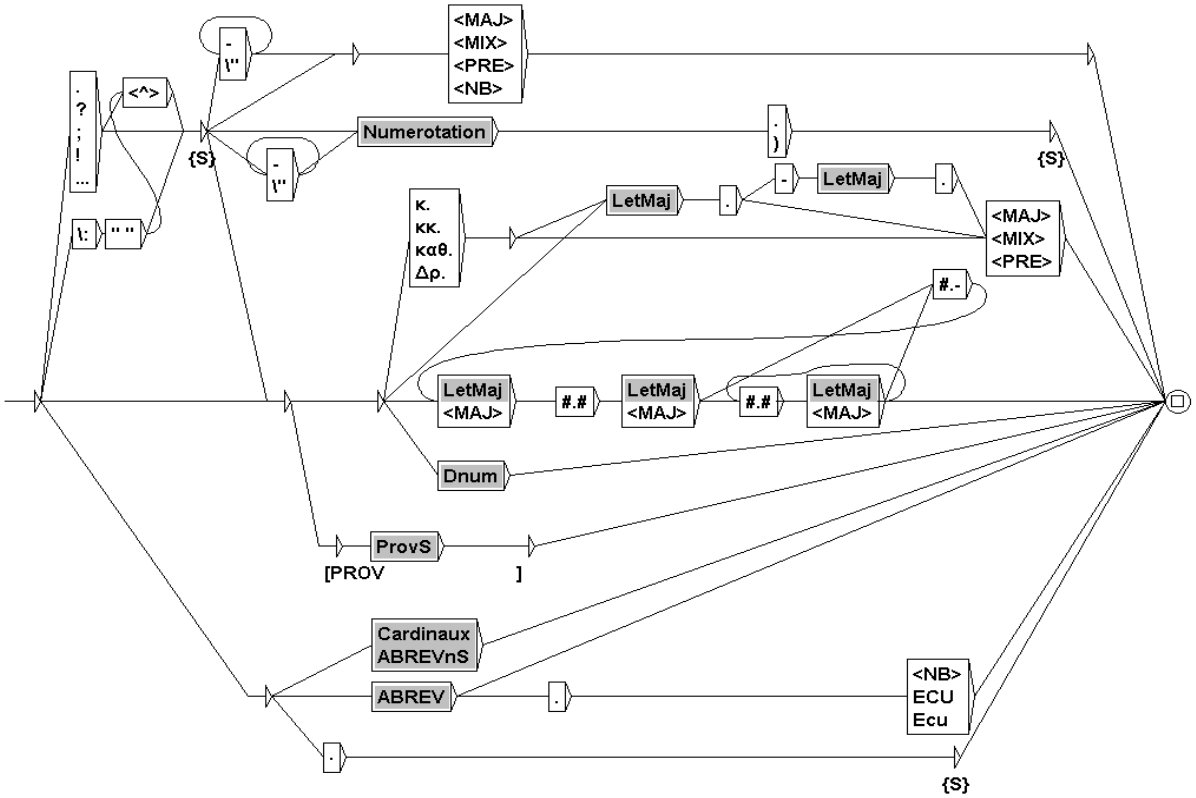


**Figure 1**

Ensuite l'alignement est effectué, mais un autre facteur important à prendre en compte est la mise en page des documents. Notre corpus, comme la majorité des corpus « non littéraires », avait une mise en page particulière avec des tables de matières, des schémas et des graphiques

[4] Voir : http://infolingu.univ-mlv.fr/.

intégrés. Malgré les nettoyages manuels opérés sur les corpus originaux l'alignement pose des problèmes

Pour le repérage des termes dans le texte nous avons utilisé des grammaires locales avec l'exploitation du contexte immédiat. Les grammaires locales que nous avons utilisées comportent plusieurs graphes. Ces graphes ont été élaborés au sein de l'Institut Gaspard Monge et l'Université Aristote de Thessalonique depuis de nombreuses années et sont rassemblés et accessibles grâce au système Graalweb (Constant 2004). Initialement prévus pour effectuer la reconnaissance de patterns morphosyntaxiques, ils ont dû être adaptés pour l'extraction des termes. Ceci a consisté à transformer ces automates en transducteurs qui produisent les balises initiales et finales délimitant chaque terme. De plus, des variables y ont été ajoutées afin d'extraire non seulement le terme mais aussi les attributs le concernant. La création de grammaires locales est amplement utilisée pour l'extraction des entités nommées par exemple. Cette méthode consiste à utiliser la présence de mots appelés « mots déclencheurs » dans le contexte immédiat (droit ou gauche) d'une entité nommée potentielle. Or cela apparaît inopérant dans le cadre des termes techniques. Plutôt que de nous servir de « déclencheur » nous pouvons parfois nous servir d'une proximité de forme commune au deux langues. Nous donnons (figure 2) un exemple de graphe permettant de reconnaître dans un texte un nom (par exemple d'organisation) suivi de son sigle entre parenthèses. Nous présentons ensuite (figure 3) les concordances obtenus sur un corpus parallèle grec et anglais.

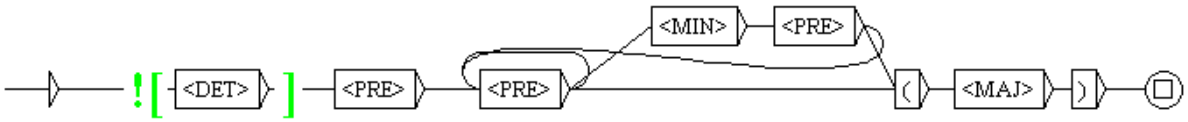


**Figure 2**

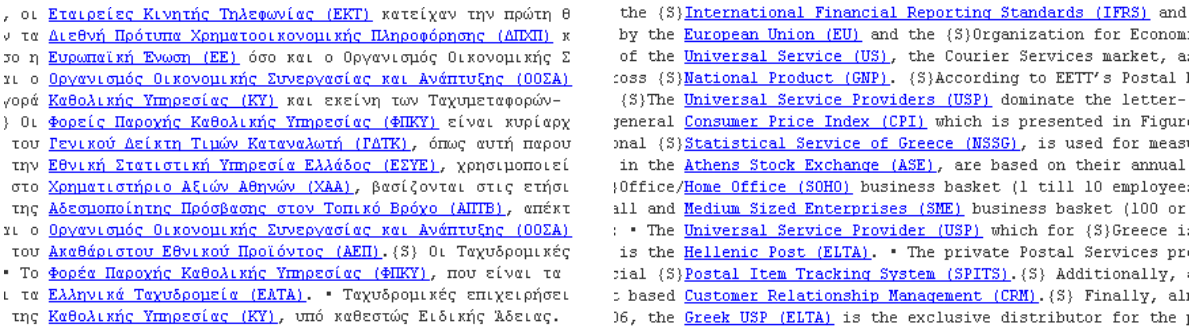


**Figure 3**

## 7. Conclusion

Cette étude qui se poursuit nous a tout d'abord permis de valider les ressources linguistiques dont nous disposons. Si l'extraction de termes peut s'apparenter dans une certaine mesure à

celle d'entités nommées (certains termes pouvant en être), cette approche utilisée seule est insuffisante de par la non présence systématique de « déclencheurs ». Il faut en plus se servir de la typologie et du vocabulaire de base du domaine concerné (ici les télécommunications) pour étudier non seulement un contexte immédiat mais aussi un contexte (gauche et droit) plus étendu. Ceci peut améliorer à la foi la qualité et la quantité de termes extraits ainsi que l'alignement de textes.

**Bibliographie**


Chinchor N., 1998. « MUC-7 Named Entity Task Definition (version 3.5) », in *Proceedings of the 7th Message Understanding Conference (MUC-7), 19 April-1 May* 1998, Fairfax, VA.

Constant M., 2004. « GRAAL, une bibliothèque de graphes : mode d'emploi », in Muller C., Royeauté J. et Silberztein M. (éds), *Cahiers de la MSH Ledoux 1, INTEX pour la linguistique et le traitement automatique des langues*, Presse Universitaire de Franche-Comté, Besançon : 321-330.

Courtois B.,1990. « Un système de dictionnaires électroniques pour les mots simples du français », in Courtois B. et Silberztein M. (éds), *Dictionnaires électroniques du français*, *Langue Française*, n° 87, Larousse, Paris : 11-22.

Gross M., 1981.,« Les bases empiriques de la notion de prédicat sémantique », in *Langages*, n° 63, Larousse, Paris : 7-52.

Hobbs J., Appelt D., Bear J., Israel D., Kameyama M., Stickel M. et Tyson M., 1996. « FASTUS : a cascaded finite-state transducer for extracting information from natural-language text», in Roche E. et Schabes Y. (éds), *Finite State Devices for Natural Language Processing*, MIT Press, Cambridge, USA : 383-406.

Li H., SRIHARI R, Niu C et Li W., 2002. « Location normalization for information extraction », in *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, Association for Computational Linguistics, Taipei, Taiwan : 1-7.

Paumier S., 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Thèse de doctorat, Université de Marne-la-Vallée.

Roche E. et Schabes Y., 1997. *Finite-State Language Processing*, Roche E. et Schabes Y. (éds), MIT Press, Cambridge, Mass./London (Language, Speech and Communication), 464 p.

Sekine S. et Nobata C.,1998. « An Information Extraction System and a Customization Tool », in *Proceedings of the New Challenges in Natural Language Processing and its Application*, *25-26 May* 1998, Tokyo, Japan.

Watrin P.,2006. *Une approche hybride de l'extraction d'information : sous-langages et lexique-grammaire*, Cental, Université de Louvain-La-Neuve, Belgique [Thèse de doctorat]