Challenges in Extracting Terminology from Modern Greek Texts

Aristomenis Thanopoulos and **Katia Kermanidis and Nikos Fakotakis**[1]

Abstract. This paper describes the automatic extraction of economic terminology from Modern Greek texts as a first step towards creating an ontological thesaurus of economic concepts. Unlike previous approaches, the domain-specific corpus utilized is varying in genre, and therefore rich in vocabulary and linguistic structure, while the pre-processing level is relatively low (basic morphological tagging, the detection of elementary, non-overlapping chunks) and fully automatic. The idiosyncratic properties of Modern Greek noun phrases are taken into account: the freedom in word ordering, the richness in morphology. Also, the peculiarity of the available corpora is dealt with: the large size of the economic compared to the balanced corpus. A combination of statistical filters (relative frequency ratios and log likelihood) and smoothing is employed in order to deal with the afore-mentioned challenges when filtering out non-terms.

1 INTRODUCTION

Terms are the linguistic expression of concepts. Domain-specific terms capture the knowledge of a given domain and reflect it in the form of words that are commonly acceptable by the members of the domain community, enabling the latter to interact and exchange information. In contrast to the use of static dictionaries, acquiring terminology automatically from domain texts leads to a list of extracted terms that may be dynamically updated and ranked according to usage. Term extraction is a first step towards acquiring a domain ontology. An ontology is a thesaurus that provides the relationships among the terms, and sorts them in a hierarchical structure, based on their semantic specificity and their properties.

Several methods have been employed for the extraction of domain terms. Regarding the linguistic pre-processing of the text corpora, approaches vary from simple tokenization and part-of-speech tagging ([1],[2]), to the use of shallow parsers and higher-level linguistic processors ([4],[8]). The latter aim at identifying syntactic patterns, like noun phrases, and their structure (e.g. head-modifier), in order to rule out tokens that are grammatically impossible to constitute terms (e.g. adverbs, verbs, pronouns, articles, etc).

Regarding the statistical filters, that have been employed in previous work to filter out non-terms , they also vary. Using corpus comparison, the techniques try to identify words/phrases that present a different statistical behavior in the corpus of the target domain, compared to their behavior in the rest of the corpora. Such words/phrases are considered to be terms of the domain in question. In the most simple case, the observed frequencies of the candidate terms are compared ([1]). Kilgarriff in [6] experiments with various other metrics, like the χ^2 score, the t-test, mutual information, the Mann-Whitney rank test, the Log Likelihood, Fisher's exact test and the TF.IDF (term frequency-inverse document frequency). Frantzi et al. in [2] present a metric that combines statistical (frequencies of compound terms and their nested sub-terms) and linguistic (context words are assigned a weight of importance) information.

In this paper we present the first phase of the ongoing work towards the creation of an ontology hierarchy of economic concepts. This phase includes the extraction of economic terms automatically from a Modern Greek phrase-analyzed corpus by corpora comparison in combination to applying a threshold to the relative frequency ratios.

An important aspect of the present approach is the stylistic nature of the domain-specific (economic) corpus. In most of the previous work, the domain corpus is to a large extent restricted in the vocabulary it contains and in the variety of syntactic structures it presents. Our economic corpus does not consist of syntactically standardized taglines of economic news. On the contrary, it presents a very rich variety in vocabulary, syntactic formulations, idiomatic expressions, sentence length, making the process of term extraction an interesting challenge.

In addition to this, the employed pre-processing tools (shallow phrase chunker) make use of limited resources (see section 2.2) and the question arises whether the resulting low-level information is sufficient to deal with the linguistic complexity of the corpus.

Another challenge that has been faced by the present work is the language itself. In Modern Greek the ordering of the constituents of a sentence or a phrase is loose and determined primarily by the rich morphology. As a result, the extraction of compound terms, as well as the identification of nested terms, are not straightforward and cannot be treated as cases of simple string concatenation, as in English. Section 2.3 describes an approach for extracting the counts of candidate terms, which takes into account the freedom in word ordering.

Finally, a peculiar trait of the current work is the corpora that are available to us. While the economic corpus is sufficiently large, the balanced corpus is relatively small. As a result, the terms (especially bi-grams) that occur in both corpora are few, while many valid terms appear in the domain specific corpus alone. This makes it impossible to use the traditional methodology of corpora comparison alone (that presupposes the appearance of a candidate term in both corpora) in order to filter out non-terms. A smoothing technique is applied to overcome this problem, which is described in section 3.

[1] Wire Communications Laboratory, University of Patras, Greece. Email: {aristom, kerman, fakotaki}@wcl.ee.upatras.gr

2 LINGUISTIC PROCESSING

A set of linguistic processing tools have been employed in order to parse the textual corpora. The first goal is to detect nouns (e.g. *τράπεζα* - bank), nominal compounds (*αύξηση κεφαλαίου* - capital increase) and named entities (*Τράπεζα της Ελλάδος* - Bank of Greece). All the above structures appear in the noun and prepositional phrases in a sentence. These types of phrases need to be detected, non-content words that appear in them have to be disregarded, and the candidate economic terms need to be formed. This process is described in detail in the rest of this section.

2.1 Modern Greek

Regarding the properties of the language that are strongly related to the current task, it has to be taken into account that Modern Greek is highly inflectional. The rich morphology allows for a larger degree of freedom in the ordering of the constituents of a phrase (headword and modifiers), compared to other languages such as English or German. More specifically, modifiers like adjectives, numerals and pronouns may precede or follow the head noun.

Another common property of noun phrases is the presence of nominal modifiers in the genitive case that denote possession, quality, quantity or origin. They are nouns and usually follow the head noun they modify.

The following two examples show the afore-mentioned freedom. The two phrases have exactly the same meaning (*bank account*). The first phrase is an adjective-noun construction, while the second is a noun-genitive modifier construction.

τραπεζικός λογαριασμός	$\text{bank}_{[\text{ADJECTIVE}]}$ $\text{account}_{[\text{NOUN}]}$
λογαριασμός τράπεζας	$\text{account}_{[\text{NOUN}]}$ $\text{bank}_{[\text{NOUN-GENITIVE}]}$

2.2 Corpora and processing tools

The corpora used in our experiments were:

1. The ILSP/ELEFTHEROTYPIA ([3]) and ESPRIT 860 ([9]) Corpora (a total of 300,000 words). Both these corpora are balanced and manually annotated with complete morphological information. Further (phrase structure) information is obtained automatically.

2. The DELOS Corpus, [5], is a collection of economic domain texts of approximately five million words and of varying genre. It has been automatically annotated from the ground up. Morphological tagging on DELOS was performed by the analyzer of [10]. Accuracy in part-of-speech and case tagging reaches 98% and 94% accuracy respectively. Further (phrase structure) information is again obtained automatically.

All of the above corpora (including DELOS) are collections of newspaper and journal articles. More specifically, regarding DELOS, the collection consists of texts taken from the financial newspaper EXPRESS, reports from the Foundation for Economic and Industrial Research, research papers from the Athens University of Economics and several reports from the Bank of Greece. The documents are of varying genre like press reportage, news, articles, interviews and scientific studies and cover all the basic areas of the economic domain, i.e. microeconomics, macroeconomics, international economics, finance, business administration, economic history, economic law, public economics etc. Therefore, it presents a richness in vocabulary, in linguistic structure, in the use of idiomatic expressions and colloquialisms, which is not encountered in the highly domain- and language-restricted texts used normally for term extraction (e.g. medical records, technical articles, tourist site descriptions). To indicate the linguistic complexity of the corpus, we mention that the length of noun phrases varies from 1 to 53 word tokens.

All the corpora have been phrase-analyzed by the chunker described in detail in [11]. Noun, verb, prepositional, adverbial phrases and conjunctions are detected via multi-pass parsing. From the above phrases, noun and prepositional phrases only are taken into account for the present task, as they are the only types of phrases that may include terms. Regarding the phrases of interest, precision and recall reach 85.6% and 94.5% for noun phrases, and 99.1% and 93.9% for prepositional phrases respectively. The robustness of the chunker and its independence on extravagant information makes it suitable to deal with a style-varying and complicated in linguistic structure corpus like DELOS.

It should be noted that phrases are non-overlapping. Embedded phrased are flatly split into distinct phrases. Nominal modifiers in the genitive case are included in the same phrase with the noun they modify; nouns joined by a coordinating conjunction are grouped into one phrase. The chunker identifies basic phrase constructions during the first passes (e.g. adjective-nouns, article nouns), and combines smaller phrases into longer ones in later passes (e.g. coordination, inclusion of genitive modifiers, compound phrases). As a result, named entities, proper nouns, compound nominal constructions are identified during chunking among the rest of the noun phrases.

The most significant sources of error during the automatic chunking process, which also affect the performance of the term extraction process, are:

1. Excessive phrase cut-up, usually due to erroneous part-of-speech tagging of a word (the word *πλήρες* - full - in the following example is erroneously tagged as a noun and not as an adjective)

NP[*Το πλήρες****]*** ***NP[****κείμενο της ανακοίνωσης****]*** instead of

NP[*Το πλήρες κείμενο της ανακοίνωσης****]***

*(****NP[****the full text of the announcement****]****)*

2. Erroneous NP tagging (unidentifiable adverbs, like *όντως* – in fact – in the following example, are marked as nouns)

NP[*όντως****]*** instead of ***ADV[****όντως****]***

In order to detect simple phrases inside larger coordination constructions, we applied the following simple empirical grammar to every noun and prepositional phrase extracted by the chunker. The grammar, which directly identifies conjunctive expressions and produces a list of simple noun phrases, employs the following rules:

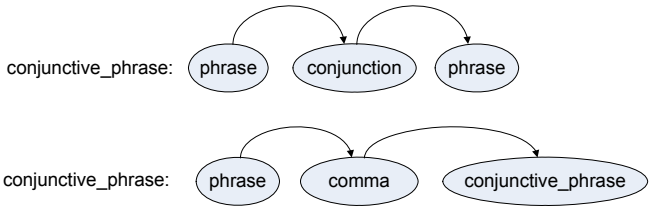

Figure 1. The rules for splitting coordinated phrases.

2.3 Candidate terms

As mentioned before, the noun and prepositional phrases of the two corpora are selected, as only these phrases are likely to contain

terms. Words of no semantic content (i.e. introductory articles, adverbs, prepositions, punctuation marks and symbols) are removed from the phrases.

Coordination schemes are detected within the phrases, and the latter are split into smaller phrases respectively according to the grammar depicted in Figure 1. The occurrences of words and N-grams, pure as well as nested, are counted. Longer candidate terms are split into smaller units (tri-grams into bi-grams and uni-grams, bi-grams into uni-grams).

Regarding the bi-grams, in order to overcome the freedom in the word ordering, as discussed in section 2.1, we considered bi-gram *A B* (*A* and *B* being the two lemmata forming the bi-gram) to be identical to bi-gram *B A*, if the bi-gram is not a named entity. Their joint count in the corpora is calculated and taken into account. The resulting uni-grams and bi-grams are the candidate terms. The candidate term counts in the corpora are then used in the statistical filters described in the next section.

Figure 2 shows the count calculation for the nested candidate terms. The two tri-grams, *A B C* and *B C D* occur in a corpus three and four times respectively. The accumulative counts of the nested terms are shown in parentheses.

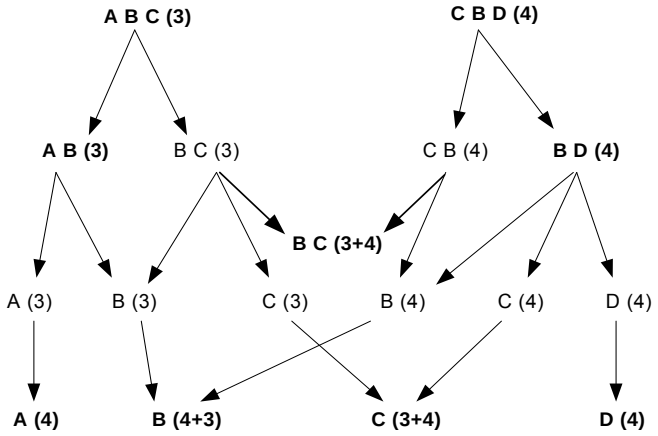

Figure 2. Calculation of n-gram frequencies, given the phrase-chunked corpus. The finally extracted n-gram frequencies are indicated in bold.

3 TERM FILTERING

In this section we describe the statistical filters that have been used to filter out non-terms. With *D* we denote Delos and with *B* the balanced corpus. As a first step, the occurrences of each candidate term *w* ($c_w(D)$ and $c_w(B)$) are counted in the two corpora separately.

A particularity of the present work is that, unlike in most previous approaches to term extraction, the domain-specific corpus available to us is quite large compared to the balanced corpus. As a result, several terms that appear in DELOS do not appear in the balanced corpus, making it impossible for the LLR statistic to detect them. In other words, these terms cannot be identified by traditional corpora comparison.

In order to deal with this phenomenon, we applied a smoothing technique to take into account terms that do not appear in the balanced corpus. More specifically, we applied Lidstone's law ([7]) to our candidate terms, i.e. we augmented each candidate term count by a value of λ=0.5 in both corpora. Thereby, terms that actually do not appear in the balanced corpus at all, end up having $c_w(B)=0.5$.This value was chosen for λ because, due to the small size of the balanced corpus, the probability of coming across a previously unseen word is significant.

Filtering was then performed in two stages: First the relative frequencies are calculated for each candidate term *w*, as

$$RF_w = f_w(D)/f_w(B), \quad (1)$$
$$f_w(D) = c_w(D)/N \quad (2)$$
$$f_w(B) = c_w(B)/M \quad (3)$$

N and *M* denote the counts of all candidate terms in *D* and *B* respectively.

In the next step, for those candidate terms that present an $RF_w>1$, LLR is calculated (according to the formula of [6]) as

$$\begin{aligned} LLR_w = 2\cdot(&c_w(D)\cdot\log(c_w(D)) + c_w(B)\cdot\log(c_w(B)) + \\ &(N-c_w(D))\cdot\log(N-c_w(D)) + (M-c_w(B))\cdot\log(M-c_w(B)) - \\ &(c_w(D)+c_w(B))\cdot\log(c_w(D)+c_w(B)) - M\cdot\log M - N\log N - \\ &(N+M-c_w(D)-c_w(B))\cdot\log(N+M-c_w(D)-c_w(B)) + \\ &(N+M)\cdot\log(N+M)\ ) \end{aligned} \quad (4)$$

The LLR metric detects how surprising (or not) it is for a candidate term to appear in DELOS or in the balanced corpus (compared to its expected appearance count), and therefore constitute an economic domain term (or not). Unlike other statistics (like the χ^2 and mutual information), it is an accurate measure even for rare candidate terms, and for this reason it was selected for the present task. It is asymptotically χ^2 distributed. So, for one degree of freedom, candidate terms that present an LLR value greater than 7.88 (critical value) can be considered as valid terms with a confidence level of 0.005.

4 EXPERIMENTAL RESULTS

The final list of extracted terms was evaluated by a group of three experts in economics and finance. The evaluators were in constant contact to agree upon ambiguous cases of terms. The most important factor for this ambiguity is the lack of context information, especially for uni-grams. In other words, there are several cases of words that may or may not be economic terms depending on the context in which they appear.

Table 1 lists a window from the list of the candidate terms, selected by chance. Their counts in both corpora are also shown (original counts, prior to smoothing), along with their RF value, and the tags that were given to them by the experts. These are terms with either RF<<1 or RF >>1, i.e. terms that present a significant difference between their frequencies in the two corpora, and so they vary from strongly economic (e.g. *tax-related*) to non-economic (*island*).

As the LLR threshold value decreases (the N-best number increases), the number of non-economic and mostly non-economic terms that enters into the N-best terms also increases causing the precision to drop.

The results cannot be easily compared to those of previous approaches, due to the many differences in resources and pre-processing. Merely as an indication, these results are comparable to the ones reported in [1] (73% to 86% precision, using a threshold on term frequencies in technical corpora on fiber optic networks, depending on the specific domain corpus and the size of the extracted list of candidate terms, which is similar to the list size in the current work).

Figure 3 shows the percentage of terms that have been correctly labeled as valid terms (y-axis) when taking into account the N-best labeled terms (x-axis) (i.e. for different LLR thresholds). This graph refers to terms that appear in both corpora and for which $RF_w>1$. *Strongly economic* are terms that are characteristic of the

Table 1. The 24 terms with the highest LLR scores along with their counts and their domain relevance.

word	translation	DILOS Freq.	IEL Freq.	Relative Freq. Ratio	LLR	Important to the Domain	Possibly Important to Domain	Unimportant to Domain
φορολογικός	tax-related	352	13	4,63	49,0	✓	-	-
παρών	present	13	24	0,09	48,5	-	-	✓
γλώσσα	language	13	24	0,09	48,5	-	-	✓
αριστερός	left, leftist	7	20	0,06	48,3	-	✓	-
εσωκομματικός	intra-party (political)	10	22	0,08	48,1	-	✓	-
διάλογος	dialog	131	68	0,33	47,4	-	-	✓
πετρέλαιο	oil (petrol)	213	3	12,14	47,2	✓	-	-
κερδοφορία	profitability	164	0	-	47,1	✓	-	-
πρόβλεψη	prediction	283	8	6,05	46,9	✓	-	-
νησί	island	14	24	0,10	46,8	-	-	✓
άγκυρα	anchor	4	17	0,04	46,2	-	-	✓
γιεν	yen	161	0	-	46,1	✓	-	-
στόχος	target	821	64	2,19	46,1	✓	-	-
αστυνομία	police	45	38	0,20	46,0	-	✓	-
εργάτης	factory worker	3	16	0,03	45,9	-	✓	-
προοπτική	prospect	446	23	3,32	45,8	✓	-	-
ΟΤΕ	ΗΤΟ (company)	149	0	-	45,8	✓	-	-
συμφωνία	agreement	654	45	2,49	45,8	✓	-	-
γερμανικός	German	238	5	8,14	45,7	-	✓	-
πολιτισμός	culture	31	32	0,17	45,6	-	✓	-
δουλειά	job, work	38	35	0,19	45,6	-	✓	-
διευθύνων	chief (executive)	199	3	11,43	45,6	✓	-	-
διοικητικός	administrative	278	8	5,94	45,6	✓	-	-
ισοτιμία	currency	182	2	15,68	45,4	✓	-	-

domain and necessary for understanding domain texts. *Economic* are terms that function as economic within a context of this domain, but may also have a different meaning outside this domain. As regards to the aforementioned labeling, this category includes terms connected both directly and indirectly to the domain. *Mostly non-economic* are words that are connected to the specific domain only indirectly, or more general terms that normally appear outside the economic domain, but may carry an economic sense in certain limited cases. *Non-economic* are terms that never appear in an economic sense or can be related to the domain in any way. For example, referring to Table 1, "φορολογικός" ("tax" [adjective]) is considered as a strongly economic term, while "πολιτισμός" ("culture") is characterised as possibly important to the domain of economics, since it often involves a financial level.

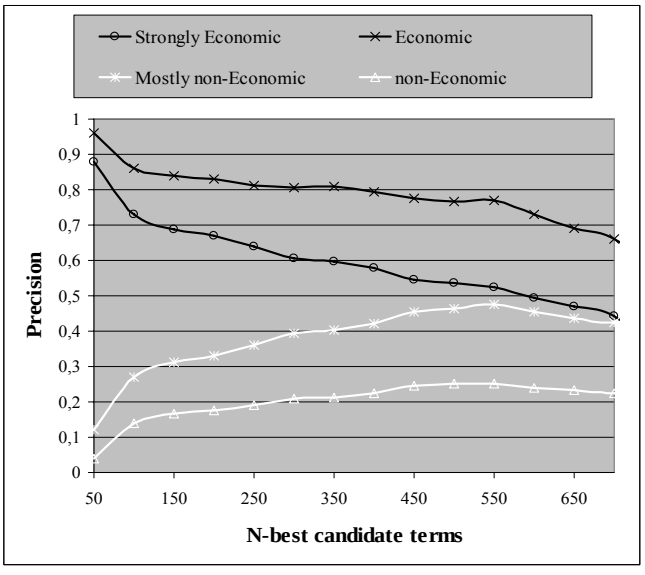

Figure 3. Precision (y-axis) for the N-best candidate terms (x-axis) that appear in both corpora and that present RF>1.

Figure 4 shows the precision achieved for the terms appearing in both corpora that present an RF<1. It is an interesting graph to observe, in combination with Figure 3, as it shows how the method performs for the terms that are more frequent in the balanced corpus in comparison to DELOS.

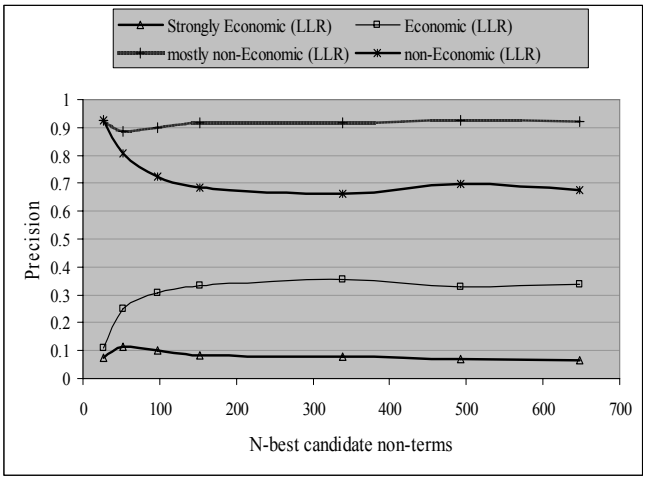

Figure 4. Precision (y-axis) for the N-best terms (x-axis) that appear in both corpora and that present RF<1.

Figure 5 depicts comparative results between LLR and term extraction based on simple frequency counts on DELOS only. This experiment was performed to show the importance of corpora comparison for term extraction, compared to using only a domain-specific corpus and applying simple frequencies to the candidate terms appearing in it. As expected, corpus comparison (LLR) leads to better results as it is concluded by the increased distance between the Economic term curves and the non-Economic term ones. Simple frequency counts tend to include many undesired N-grams among the candidate terms with the highest ranks, simply because these N-grams appear frequently in the corpus. As a result, the precision values with frequencies on one corpus only, inevitably drop.

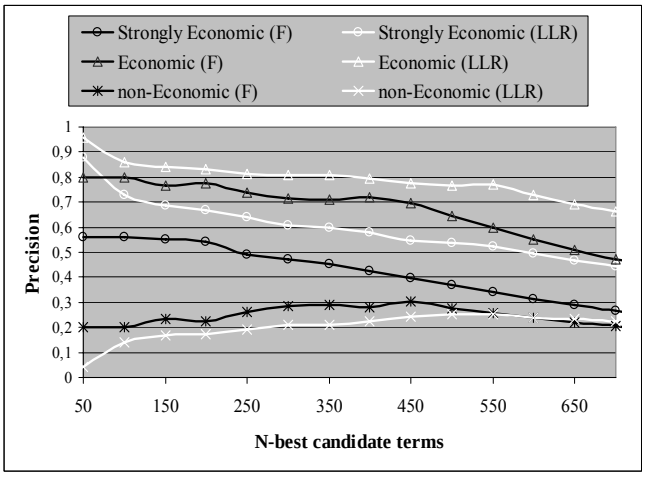

Figure 5. Comparative precision between LLR and simple frequency counts on DELOS.

Table 2 shows the RF and LLR scores of the 20 most highly ranked economic terms, ordered by their LLR value. The depicted counts are the original ones, prior to smoothing. An interesting term is "υψηλός", the ancient Greek form for "high", used today almost exclusively in the context of the degree of performance, growth, rise, profit, cost, drop (i.e. the appropriate form in economic context), as opposed to its modern form "ψηλός", which is used in the concept of the degree of actual height.

Table 2. The 20 most highly ranked economic terms

Rank	word	translation	Cw(D)	Cw(B)	RFw	LLR
1	εταιρία	company	5396	0	1845,9	852,0
2	δρχ	drachma	3003	1	342,5	465,5
3	μετοχή	stock	2827	6	74,4	414,0
4	αγορά	buy	2330	33	11,9	257,2
5	αύξηση	growth, rise	2746	66	7,1	247,6
6	κέρδος	profit	1820	15	20,1	228,2
7	τράπεζα	bank	1367	11	20,3	171,8
8	επιχείρηση	enterprise	1969	56	6,0	162,1
9	κεφάλαιο	capital	1325	14	15,6	157,3
10	σημαντικός	important	1872	56	5,7	149,3
11	πώληση	sell	1203	11	17,9	147,3
12	προϊόν	product	1282	16	13,3	146,0
13	όμιλος	(company) group	1036	5	32,2	140,0
14	Α.Ε.	INC	820	0	280,7	126,4
15	μετοχικός	stocking	790	2	54,1	112,8
16	τιμή	price	1722	70	4,2	110,9
17	επιτόκιο	interest (financ.)	821	4	31,2	110,0
18	υψηλός	high (old form)	711	0	243,4	109,2
19	κόστος	cost	1031	19	9,0	103,4
20	κλάδος	branch	833	7	19,0	103,2

Figure 6 shows the difference in precision with LLR for the N-best terms with and without the application of smoothing. When smoothing is not applied, the drop in performance is significant (around 20%). The expected performance improvement due to the smoothing process is further enhanced, because the terms that appear only in DELOS (and not in the balanced corpus) are not taken into account when smoothing is not performed.

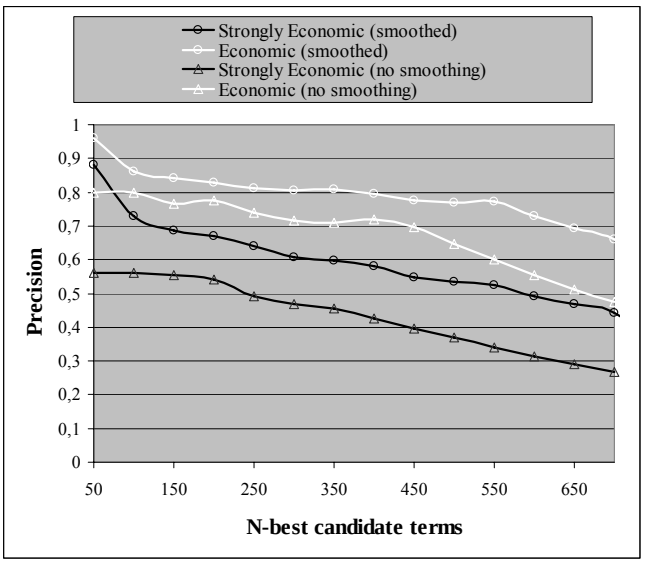

Figure 6. Comparative precision using the LLR metric with and without smoothing.

5 CONCLUSION

In this paper we have presented the process of automatically extracting economic terminology from Modern Greek texts. The properties of the language are taken into account by utilizing appropriate pre-processing tools. The linguistic complexity of the domain-specific corpus is addressed by adjusting the traditional candidate term formation methodology to deal with the freedom in word ordering. Finally, the unusual size difference between the two corpora (domain-specific and general) leads to a sparse data problem, which is dealt with satisfactorily by applying Lidstone's smoothing law.

ACKNOWLEDGEMENTS

We thank the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II), and particularly the Program PYTHAGORAS II, for funding the above work.

REFERENCES

[1] P. Drouin, '*Detection of Domain Specific Terminology Using Corpora Comparison*', 4th International Conference on Language Resources and Evaluation (LREC), 79–82, Lisbon, (2004).

[2] K.Frantzi, S. Ananiadou, and H. Mima, '*Automatic Recognition of Multi-word Terms: the C-value/NC-value Method*', International Journal on Digital Libraries, **3** (2), 117–132, (2000).

[3] N. Hatzigeorgiu, M. Gavrilidou, S. Piperidis, G. Carayannis, A. Papakostopoulou, A. Spiliotopoulou, A. Vacalopoulou, P. Labropoulou, E. Mantzari, H. Papageorgiou, and I. Demiros, '*Design and Implementation of the online ILSP Greek Corpus*', 2nd International Conference on Language Resources and Evaluation (LREC), Athens, 1737–1742, (2000).

[4] A. Hulth, '*Improved Automatic Keyword Extraction Given More Linguistic Knowledge'*, International Conference on Empirical Methods in Natural Language Processing (EMNLP), Sapporo, 216-223, (2003).

[5] K. Kermanidis, N. Fakotakis and G. Kokkinakis, '*DELOS: An Automatically Tagged Economic Corpus for Modern Greek*', 3rd International Conference on Language Resources and Evaluation (LREC), Las Palmas de Gran Canaria, 93-100, (2002).

[6] Kilgarriff, '*Comparing Corpora*', International Journal of Corpus Linguistics, **6** (1), 1-37, (2001).

[7] C. Manning and H. Schuetze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[8] R. Navigli and P. Velardi, '*Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites*', Computational Linguistics, **30** (2), 151–179, (2004).

[9] Partners of ESPRIT-291/860, *Unification of the Word Classes of the ESPRIT Project 860*, Internal Report BU-WKL-0376, (1986).

[10] K. Sgarbas, N. Fakotakis and G. Kokkinakis, '*A Straightforward Approach to Morphological Analysis and Synthesis*', Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries (COMLEX), Kato Achaia, Greece, 31–34, (2000).

[11] E. Stamatatos, N. Fakotakis and G. Kokkinakis, '*A practical chunker for unrestricted text*', Proceedings of the Conference on Natural Language Processing (NLP), Patras, Greece, 139–150, (2000).