A register-diversified corpus in the Stock Market domain: The case of idiomatic and support verb expressions

Eleni Tziafa
Aristotle University of Thessaloniki

1. Introduction

This paper reports on a corpus-based analysis of idiomatic and support verb expressions in different registers of the Stock Market domain. The research presented here is part of the research activities (building lexical resources and tools for Greek language) of the Laboratory of Translation and Language Processing, which are being carried out at the Aristotle University of Thessaloniki. This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) – Research Funding Program: "Heraclitus II - Investing in knowledge society through the European Social Fund". The ultimate goal of this research is to study the properties of the Greek stock market terms, in order to form part of a machine-readable dictionary of Greek language. Structural Linguistics (Harris 1951, Gross 1975) and register analysis (e.g. Biber 1988, 1993; Biber and Conrad 2009) are the theoretical backgrounds of this research. The corpus under study in this paper consists of Greek texts of the domain of Stock Market, divided into four register-diversified sub-corpora and comprises over 18 million running words. This work focuses on the identification and analysis of collocations for idiomatic (verb/noun) and support verb expressions of frequent words in this domain.

2. The corpus under study

2.1. The Greek Stock market

The language of business and finance has always attracted interest, since the movement of stock indexes can be an indicator, a "barometer" of the general trend in the economy. Of particular interest is the language of stock market which is often perceived as a secret code (Behr et al. 2007). English is the *lingua franca* of business and finance, and this is reflected in the language of every stock Exchange of the world and subsequently, in the language of the Greek Stock Exchange. Nevertheless, "the world of business is rapidly changing into a smaller place where individuals from different backgrounds and of different nationality develop common generic frames of reference for producing and understanding the communication of the marketplace" (Trosborg et al., 2005).

At these times of turbulence, as in an unexpected turn of events, Greece was found in the epicentre of global financial anxiety, "knowing the specifics of language and communication in business will help everyone to understand the deeper inner meaning implied in socio-economic, corporate and advertising discourse, to identify the manipulative mechanisms and techniques influencing public opinion" (Daniushina 2010). As in Greece there is a fertile ground for conspiracy theories, the decoding of the language of the stock exchange is considered essential. Since "the business environment is becoming increasingly more homogeneous" and the global economy "frontierless" (Trosborg et al. 2005), studying the language of Greek finance could be interesting even well beyond the country's frontiers. According to Kelly-Holmes & Mautner (2010) "new roles are emerging for minority languages particularly because of global, national and local market processes".

The Greek Stock Exchange is a peripheral stock exchange with 125 years of history. Therefore, besides common terms included in every stock market in the world, there are still certain new terms, neologisms,

simple or multi-word expressions, including idioms and metaphors, which were developed and have to be identified and analysed to be understood. Moreover, collecting and creating structured bodies of specialized text, poses a great challenge to language technology, especially for languages with small numbers of native speakers such as Greek.

2.2. Corpus balance and representativeness

While discussion on corpus representativeness has been a major and also controversial issue in corpus linguistics over the past 20 years, it has concentrated mainly on problems concerning the construction of large corpora, like Cobuild's Bank of English corpus and the BNC (Biber 1993, Atkins et al. 1992). Gavioli (2002) poses the problem of small corpus representativeness and of criteria used particularly in the design of small specialized corpora. Tony McEnery et al. (2006) observe that "while both general and specialized corpora should be representative of a language or language variety, the representativeness of the two types of corpora are measured in different ways. The representativeness of a general corpus depends heavily on sampling from a broad range of genres whereas the representativeness of a specialized corpus, at the lexical level at least, can be measured by the degree of 'closure' (McEnery and Wilson 2001: 166) or 'saturation' (Belica 1996: 61-74) of the corpus".

For this project we tried to combine both methods. The initial corpus of 50 million words, it was divided in the following four registers: written conversational (posts in e-forums), journalistic (newspaper articles), official business documents (stock market related documents) and academic (articles, dissertations, etc.). Biber (1993) states that "in statistical terms, the between-group variance is typically larger than within-group variance, and thus a sample that forces representation across identifiable groups will be more representative overall. Further, in the case of language corpora, proportional representation of texts is usually not desirable; rather, representation of the range of text types is required as a basis for linguistic analyses, making a stratified sample even more essential".

Nevertheless, extended sampling was beyond our scope, since this corpus was designed to be used for the expansion of a machine-readable specialized dictionary. Moreover, availability was limited, as these are specialized texts, and in Greek, a less resourced language and less represented in the web. Thus, after a limited random sampling and measuring lexical closure, until "the curve of lexical growth has become asymptotic" (Teubert 1999), the final corpus comprised about 19 million words. With the multi-billion corpora announced (Ferraresi et al. 2008, 2010, Baroni et al. 2009, Pomikalek 2009), any corpus seems relatively small, but, given the domain and the language, it is a relatively big corpus, considered that the Corpus of Greek Texts (Goutsos 2003) comprises of 30 million words and the Hellenic National Corpus 47 million words.

Above all, it is true though that "to achieve corpus representativeness along the lines of the Brown model requires knowledge of which genre is used how often by the language community in the sampling period", and "to determine the correlation of language production and reception among each register" (McEnery et al., 2006). As these were rather unrealistic, and because "corpus building is of necessity a marriage of perfection and pragmatism" (McEnery et al., 2006), we settled with a four-register specialized corpus, that gives us the opportunity to study and compare language varieties of formal and informal discourse from conversations in "trading slang" to academic prose.

2.3. Registers included in the Stock Exchange Corpus

The existence of different registers is even more justified as regards Stock Exchanges. It is a professional language and, at the same time it can be a scientific one. In many countries of the world, trading stocks was actually conducted in the streets, like rue Quincampoix in Paris, or in the Exchange Alley in London, in coffee shops. A coffee house was also the first Stock Exchange in Athens (named *Bella Grecia*). Thus, at the heart of this discourse there is the language of the market, a phrase that in Greek implies slang. Even today, the

language of the market remains a colloquial language. As time went by, though, the products sold in an organized stock market began to be more and more complicated, and so was the language used. Of course, every specialized language has its jargon, but in other domains the jargon is used between experts. In the case of stock exchange language, experts and non-experts mingle haphazardly, discussing the same issues. Moreover, due to the current crisis, all the unknown before words, considered to be specialized, now "invaded" the general language. Due to the current crisis, people are significantly more aware of financial terminology, as they are daily concerned about the *spread* of Greek bonds in financial markets, and the difference between *selective, temporary, controlled, orderly, disorderly, restricted or organized default*.

2.4. The structure of the corpus under study

The texts under study are instances of different domain-specific registers. The text corpus consists of four sub-corpora of Greek texts in the Stock Market domain and it comprises over 18 million words that have been automatically part-of-speech tagged by Unitex[1]. The time period covered by the corpus ranges from 1999 to 2010, a period marked by two major crises in Greece, a Stock Market crisis and a debt crisis. These turbulent years also give us the opportunity to study the language of the stock market in its high and its low, or in bearish and bullish times. It must be noted that texts are stored in their entirety, as full texts, where possible, and no translated texts are included (again, to the extent that such an exclusion is possible). Copyright, as predicted by McEnery (2006), Wilkinson (2006) and many others, involved much effort, trouble and frustration. The permissions obtained for the corpus under study are strictly for non commercial and non-profit use, as acquired by:

- ASE S.A. (Athens Stock Exchange S.A.), regarding data from the whole site, 32Gb, mostly press releases and annual reports, a part of which (2 million words of annual reports in Greek and English language) was aligned to be used as a parallel corpus.
- Professors, graduate and postgraduate students who donated their lecture notes and dissertations.
- Administrators of two major forums with public discussions on stock exchange.

Journalistic articles are also included, from two major newspapers in Greece, *Ta Nea* and *Kathimerini*, and also other, financial or stock market specialized newspapers.

Table 1 (next page) illustrates the text sources used in the compilation of the corpus under study.

[1] Unitex: http://www-igm.univ-mlv.fr/~unitex/index.php?page=3.

Sub-corpora	Source	Time Period	Number of texts	Words
A. Posts in public discussions with Stock Market as subject	http://www.neoforum.gr http://www.capital.gr	2009-2010 2009-2010	46,912 posts	4,459,377
B. Journalistic articles, in printed or electronic form	Financial newspapers Stock Market web-pages	1999–2000 2000-2010	12,407 articles	5,114,460
C. Press Releases, Annual Reports of the Athens Stock Exchange, official documents	Web-page of the Athens Stock Exchange (http://www.ase.gr)	2000-2010	18,199 press releases, 21 annual reports, 47 official documents	5,663,526
D. Academic texts	Lecture notes, PhD theses, dissertations) and on-line available articles	2002-2010	350 texts	3,567,942
			Total	**18,805,305**

Table 1: Analysis of the corpus

Sub-corpus A consists of posted messages in public discussions in two internet forums, both dedicated to the stock market. This kind of forum appeared in Greece the last three years. This corresponds to the discourse of *professional business communication* which, according to Daniushina (2010) performs "an instrumental-persuasive function".

Sub-corpus B comes from journalistic texts, scanned from newspapers for the period 1999-2000 and consequently complemented with articles in an electronic format from 2000 to 2010, which were written at the same literary level. This corresponds to *the discourse of business media* which performs "an informative-polemic function".

Sub-corpus C comes from the website of the Athens Stock Market, which contains announcements, annual reports and articles dating from the year 2000, corresponding to *document business discourse* (internal and external business correspondence, corporate documents, regulations and charters of companies and organizations, articles of incorporation, etc. – mainly, written discourse) and it performs "a regulative function".

Sub-corpus D contains academic texts whose main focus is Money Markets and Stock Market Derivatives, which were provided from University modules. Furthermore, postgraduate and doctorate dissertations were used, available on-line. This register corresponds to *training and academic business discourse* and it performs "an educational function".

Keywords were extracted using Wordsmith tools and having as a reference corpus the publications for a period of 6 years of the newspaper *Ta Nea* (about 115 million words), as the only large corpus available as a whole for Greek language. This is a single-source corpus, obviously not balanced, but as a journalistic corpus it represents sufficiently general language, as opposed to specialized language.

	A	Posts	B	Newspapers	C	Company's documents	D	Academic
1.	ΦΙΛΕ	**friend**	ΜΕΤΟΧΩΝ	**Shares**	ΕΥΡΩ	**Euro**	ΤΙΜΗ	**Price**
2.	ΈΓΡΑΨΕ	**Wrote**	ΑΝΑΚΟΊΝΩΣΗ	**Announcement**	ΕΤΑΙΡΊΑΣ	**Company**	ΑΓΟΡΆΣ	**Market**
3.	ΕΥΡΏ	**Euro**	ΕΥΡΏ	**Euro**	ΜΕΤΟΧΏΝ	**Shares**	ΜΕΤΟΧΏΝ	**Shares**
4.	ΕΤΕ	**National Bank of Greece**	ΑΓΟΡΆ	**Market**	ΤΡΑΠΕΖΑ	**Bank**	ΔΕΊΚΤΗ	**Index**
5.	SPREAD	**Spread**	ΙΔΊΩΝ	**Own (= "equity")**	**AE**	**SA**	ΕΤΑΙΡΊΑΣ	**Company**
6.	SHORT	**short**	ΜΕΤΟΧΉ	**Share**	ΔΙΟΙΚΗΤΙΚΟΎ	**Administrative (Board of Directors)**	ΕΛΈΓΧΟΥ	**Control**
7.	ΔΝΤ (IMF)	**IMF**	ΔΙΑΒΑΣΤΕ	**Read**	ΣΥΜΜΕΤΟΧΩΝ	**"holding company"**	ΠΑΡΑΓΏΓΩΝ	**Derivatives**
8.	ΤΡΆΠΕΖΕΣ	**Banks**	ΔΗΜΟΦΙΛΈΣΤΕΡΕΣ	**Most popular**	ΣΥΝΕΛΕΥΣΗ	**Meeting**	ΚΙΝΔΎΝΟΥ	**Risk**
9.	ΟΜΌΛΟΓΑ	**bonds**	ΕΤΑΙΡΊΑΣ	**Company**	EURO	**Euro**	ΕΠΕΝΔΥΤΉΣ	**Investor**
10.	ΕΥΧΑΡΙΣΤΩ	**thank you**	ΡΥΘΜΙΖΌΜΕΝΗΣ	**Regulated**	ΚΕΦΑΛΑΊΟΥ	**Capital**	ΧΑΡΤΟΦΥΛΑΚΊΟΥ	**Portfolio**
11.	ΑΜΚ	**equity increase**	ΕΠΙΦΆΝΕΙΑ	**Surface**	ΑΝΑΚΟΙΝΏΝΕΙ	**Announces**	ΕΠΙΧΕΊΡΗΣΗΣ	**Enterprise**
12.	ΓΔ	**General index**	ΚΈΡΔΗ	**Profit**	ΟΜΊΛΟΥ	**Group**	ΣΜΕ	**Futures**
13.	MARKET	**market**	ΓΝΩΣΤΟΠΟΊΗΣΗ	**Notification**	ΜΕΤΟΧΙΚΟΎ	**Shareholders**	ΑΞΊΑ	**Value**
14.	ΒΆΣΗΣ	**basement**	BANK	**Bank**	ΣΥΜΒΟΥΛΊΟΥ	**Board of directors**	ΣΥΝΑΛΛΑΓΏΝ	**Transaction**
15.	ΤΙΜΗ	**Price**	ΝΑΥΤΕΜΠΟΡΙΚΗ	**Naftemporiki**	ΆΡΘΡΟ	**Article**	ΕΣΩΤΕΡΙΚΟΎ	**Internal**
16.	ΚΑΛΗΜΈΡΑ	**Good morning**	ΠΛΗΡΟΦΟΡΊΑΣ	**Information**	ΨΉΦΟΥ	**vote**	ΚΕΦΆΛΑΙΟ	**Capital**
17.	ΕΤΑΙΡΊΑΣ	**Company**	ΕΚΑΤ	**Million**	ΜΕΡΊΣΜΑΤΟΣ	**Dividend**	GRANGER	**Granger**
18.	CAPITAL	**Capital**	ΟΙΚΟΝΟΜΙΑ	**Economy**	ΕΛΛΑΔΟΣ	**Greece**	ΑΝΆΛΥΣΗ	**Analysis**
19.	LONG	**Long**	ΣΥΝΈΛΕΥΣΗΣ	**Meeting**	ΧΡΉΣΗΣ	**Fiscal year**	ΑΠΟΔΟΣΕΙΣ	**Yield**
20.	ΦΙΛΟΙ	**friends**	WALL	**Wall Street**	ΣΥΝΕΛΕΥΣΗΣ	**meeting**	ΔΙΚΑΙΩΜΑΤΟΣ	**Right**

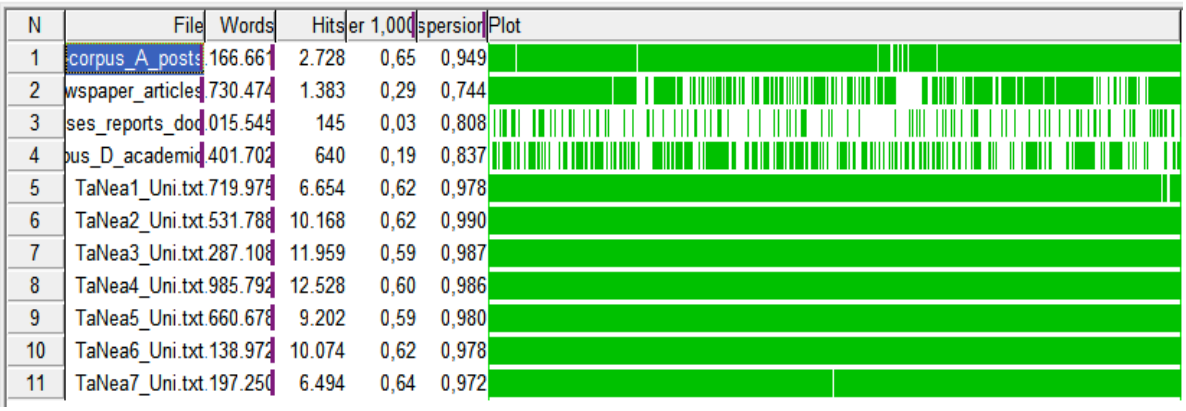
Table 2: Keywords

	A	Posts	B	Newspapers	C	Company's documents	D	Academic
1.	**Είναι**	**Is**	**ευρώ**	**Euro**	**ευρώ**	**Euro**	**Είναι**	**Is**
2.	Έχει	**Has**	Ανακοίνωση	**Announcement**	μετοχών	**Shares**	Έχει	**Has**
3.	**Ευρώ**	**Euro**	αγορά	**Market**	εταιρείας	**Company**	Τιμή	**Price**
4.	έγραψε	**Wrote**	εκατ	**Millions**	είναι	**Is**	Αγοράς	**Market**
5.	Ελλάδα	**Greece**	Αγορά	**Market**	μετοχές	**Shares**	Πρέπει	**Must**
6.	Έχουν	**(they) have**	εταιρεία	**Company**	εκατ	**Millions**	Μπορεί	**Can**
7.	Ήταν	**Was**	κέρδη	**Profit**	αύξηση	**Increase**	Μετοχών	**Shares**
8.	Πρέπει	**Must**	μετοχή	**Share**	Συμβουλίου	**Board**	Αγορά	**Market**
9.	αγορά	**Market**	έχουν	**Have**	Διοικητικού	**Board**	Περίπτωση	**Case**
10.	Μονάδες	**Units**	μετοχές	**Shares**	ΤΡΑΠΕΖΑ	**Bank**	Δείκτη	**Index**
11.	Μπορεί	**Can**	ιδίων	**Own**	Αθηνών	**Athens**	Αξία	**Value**
12.	Τράπεζες	**Banks**	Ελλάδα	**Greece**	Γενική	**General**	Συναλλαγών	**Transactions**
13.	Σήμερα	**Today**	δισ	**Billions**	Συνέλευση	**Meeting**	Σχέση	**Relation**
14.	Τώρα	**Now**	εργασίας	**Work**	έχει	**Has**	Επενδυτές	**Investors**
15.	Εκατ	**Millions**	Capital	**Capital**	Εταιρίας	**Company**	Στοιχεία	**Elements**
16.	Όχι	**No**	αύξηση	**Increase**	ΑΕ	**SA**	Βάση	**Base**
17.	ΕΤΕ	**NBG**	σήμερα	**Today**	κεφαλαίου	**Capital**	Ελέγχου	**Control**
18.	Φιλε	**Friend**	τιμή	**Price**	Αξίας	**Value**	Κινδύνου	**Risk**
19.	Μετοχές	**Shares**	αγοράς	**Market**	Ομίλου	**Group**	Επιχειρήσεις	**Enterprises**
20.	spread	**spread**	ήταν	**was**	Άρθρο	**Article**	**Ευρώ**	**Euro**

Table 3: Raw Frequencies

Raw frequency and keywords confirm that all the texts belong to the same domain, although there are obvious differences among registers, especially as regards the subcorpus A, where we can see some conversational conventions, like *goodmorning, my friend, or thank you*.

A significant difference between registers or genres in this domain is revealed by the plots (produced by Wordsmith Tools) of the light verb *κάνω [kano] (= to do/make)*. The singular 3rd person type "*kanei*" is the most frequent type of the verb kano, and in most cases, as measured, it forms support verb expressions. The difference in dispersion and frequency per 1,000 words is explained as in Greek the support verb expressions with kano are considered very colloquial and there is a clear difference between the four registers.



N	File	Words	Hits	er 1,000	spersion	Plot
1	corpus_A_posts	166.661	2.728	0,65	0,949	
2	wspaper_articles	730.474	1.383	0,29	0,744	
3	ses_reports_doc	015.545	145	0,03	0,808	
4	ous_D_academic	401.702	640	0,19	0,837	
5	TaNea1_Uni.txt	719.975	6.654	0,62	0,978	
6	TaNea2_Uni.txt	531.788	10.168	0,62	0,990	
7	TaNea3_Uni.txt	287.108	11.959	0,59	0,987	
8	TaNea4_Uni.txt	985.792	12.528	0,60	0,986	
9	TaNea5_Uni.txt	660.678	9.202	0,59	0,980	
10	TaNea6_Uni.txt	138.972	10.074	0,62	0,978	
11	TaNea7_Uni.txt	197.250	6.494	0,64	0,972	

Picture 1: Plot of the light verb *kano* (=to do/make)

Dispersion of *kanei* per 1,000 words				
A	B	C	D	Average dispersion *Ta Nea*
0.65	0.29	0.03	0.19	0.62

Table 4: Dispersion of *kanei* per 1,000 words

Subcorpus A is closer to the mean dispersion per 1,000 words in the reference corpus. The light verb kano is less used in subcorpus C, the one including official business documents.

2.5. Compiling and processing the corpus under study

A comprehensive review of all the tools used in order to build corpora is beyond the scope of this paper. Nevertheless, in order to obtain control over the texts and also to gain a deeper insight in the process involved, individual tools were used, freeware or open source, along with a few commercial products.
Each tool was selected among many others, depending on its function and the ability to support languages with other than latin alphabets, such as Greek. The biggest problem is the rapid expansion and change in this field of study, resulting in the unavailability of tools or their replacement by other, more sophisticated tools[2]. Moreover, tools were chosen to be used with big and batch files.
Therefore, texts were

- Extracted from the Web[3]
- Converted to plain text[4]

[2] The last time the tools were accessed was on June 2011.

[3] We used Httrack (available at http://www.httrack.com).

[4] We used HtmlasText (available at http://www.nirsoft.net/utils/htmlastext.html) and since the page contained besides html pages, word and pdf files, we used also Zilla Word to Text Converter for word files (available at http://www.pdfzilla.com/zilla_word_to_text_converter.html) PTconverter for pdf files (available at http://www.digitzone.com/ptconverter.html). For pdf files that included text in columns or containing tables or presentations, we used a commercial product such as ABBYY FineReader 10 Professional Edition (available, under a fee, at http://buy.abbyy.com/content/frpro/default.aspx). This was necessary in the case of scanning of newspaper articles in printed form for the early period of research (1999-2000).

- Converted to UTF-8[5]
- Merged in one document[6]
- Cleaned from duplicates[7]

Minor tools were also used in order to remove redundant carriage returns[8], and to change file names for batch text files[9]. Furthermore, we used some regular expressions[10] in order to delete English texts and their automatic translations, usually generated by Google translate, that were included in the subcorpus A (posts), and in order to anonymise posts. For some posts, we also converted Greenglish to Greek[11]. For keywords and statistical measures the program Wordsmith tools[12] was used. Finally, the corpus was marked up with textual information by means of Unitex[13]. All texts in the corpus were automatically annotated for part-of-speech and lemmatized through Unitex. This research actually forms part of the program for the development and gradual completion of the Greek version of Unitex - a text analysis system, already operating in many European languages.

2.6. What was not included?

- Spoken language

In general, a corpus is not considered balanced, unless it includes also speech data. This is not the case with specialized corpora. Nevertheless, it would be ideal to have three registers in the domain of spoken language as well, such as broadcast news, telephone conversations and conference meetings, but since in this domain all these are sensitive data, any kind of permission is unlikely. There are currently (June 2011) available online 214 videos (1 minute to 1 hour) with stock exchange news from the main news (www.ert-archives.gr), 265 videos (about half an hour each) from www.neoforum.gr, 728 in www.capital.gr, 21 live recordings from www.livestream.com for the Greek Stock Exchange. These videos are beyond the period under study (1999-2010) and permissions are granted only for the videos from *neoforum.gr*. The language used in videos is very close to this used in journalistic articles, and it isn't really a natural oral discourse. Nevertheless, the posts from the forums are very close to oral discourse, actually a new type of discourse, "*web discourse"*, according to Daniushina (2010) "that combines elements of both spoken and written discourses. Communication on the net requires a time contact (synchronization of information generating and perception) and provides a deep involvement in the situation with instant responding, typical of oral talk, although the "talk" is made in written (or quasi-written) form. Thus, a web business discourse is the reality of the 21st century, and we can anticipate it to be growing and ripe for research".

- European Union Documents

Although EU documents seem very useful, since they can be used as parallel corpora, they were not included due to the Eurospeak. Terms are biased; they are choices of translators and terminologists, and they are rarely found in authentic texts. Moreover, Greek language is not the source language. And "translated language is at best an unrepresentative special variant of the target language" (McEnery, 2006).

---

[5] All the txt files were converted to UTF-8, in order to retain as many characters as possible from the original files. Depending on the codification suitable for each tool, we used Ansi2Uni 1.4 or Simple Text Encoding Converter 1.0 or EditPad Lite 6.3.1 (http://www.all-freeware.com/results/convert/ansi/to/unicode) or Convert Ansi to Unicode (http://convert-ansi-to-unicode.software.informer.com/) or CpConverter_v0.1.4 (http://sourceforge.net/projects/cp-converter/files/cp-converter/CpConverter%20V%200.1.4.0/CpConverter_v0.1.4.zip/download)

[6] In order to make corrections as automatically as possible, we merged all files in one file, but keeping their original name and a separator between them, as in many cases it was necessary to split them again. So, we used Txtcollector (http://bluefive.pair.com/txtcollector.htm) and JS Text File Merger 1.0.0 (http://www.tucows.com/preview/373437). In order to split them back we used Gsplit 3 (available at http://www.gdgsoft.com/download/gsplit.aspx or Simpli File Split and Merge 1.4.0 (http://www.freedownloadscenter.com/Utilities/File_Splitting_Utilities/Simpli_File_Split_and_Merge_Download.html)

[7] For this we used mostly a commercial product like Noclone (http://noclone.net/download.aspx, where there is also available a free version), but we also tried freeware such as Duplicate Cleaner (http://www.digitalvolcano.co.uk/content/duplicate-cleaner/downloaddcmenu) or Duplicate File Finder 3.5 (http://www.snapfiles.com/downloads/dupfilefinder/dldupfilefinder.html) or duplicate File Finder (http://www.ashisoft.com/v 3.5.3).

[8] We used CleanHaven (http://www.holymackerelsoftware.com/MoreSoftware/CleanHaven/CleanHaven.html).

[9] We used Renamer (http://www.snapfiles.com/get/denrenamer.html) in order to change file name of temporary files and remove .tmp and also in order to facilitate random sampling.

[10] Regular expressions implemented via Edit plus (http://www.editplus.com) and Notepad ++ (http://notepad-plus-plus.org/download).

[11] Greeklish to Greek: http://services.innoetics.com/greeklish/Gadgets.aspx.

[12] http://www.lexically.net/wordsmith

[13] Unitex: http://www-igm.univ-mlv.fr/~unitex/index.php?page=3.

- Legal documents for the regulation of Stock Exchange

They were not included, although they were available without permission, because of their language which is *katharevousa*, an older form of Greek, and not Modern Greek.

3. Verbs in the Stock Market domain

Specialized dictionaries or bilingual glossaries mostly focus on nouns. When verbs are mentioned, they are usually described as derived from nouns, with no other information than part of speech. Bilingual glossaries are likely to give translations of verbs, but with very little syntactic and semantic information on how to use them in a sentence. The verb is a part of speech, just like adjectives and adverbs, which has long been discarded by terminologists, as shown repeatedly by L'Homme (1998, 2002, 2004) and Kübler (Kübler 2002, Kübler and Frérot 2003). Cowie (1989) also noted that in general language dictionaries, verbs are often the least well-treated part-of-speech, especially regarding syntax. Our hypothesis corresponds with Holtz (2007) "the co-occurrence of nouns and verbs is expected to vary among different domains, being evidence for register variation, as reflex of the situational parameter of field".

To decide on their status, we followed Frérot and Josselin-Leray (2007) who relied on the criteria used by L'Homme (2002) to include verbs in specialized dictionaries, i.e. "Normally, verbs [...] are included in specialized dictionaries if they are not used in other contexts" , e.g. the English verb terms *escrow (*a financial instrument held by a third party on behalf of the other two parties in a transaction)*, or avalize* (to give one's accord), or if they have a meaning that cannot be described using definitions found elsewhere (e.g. *to de-hedge*, in the field of finance, has a metaphorical meaning that cannot be captured using definitions found in other dictionaries, for instance, general language dictionaries). We also used the morphological criterion (when verbs are derived from nouns or vice-versa, e.g. *announce / announcement*). Another criterion was whether the verb combined with a term in noun form (e.g. *buy a spread, to take a report* - a slang phrase signifying that an individual's trade order has been executed). For this project we skimmed through the concordance lines and the context provided by Unitex and roughly established the frequencies of the verbs and their various constructions.

The verb / noun collocations in the four subcorpora under study are presented in Diagram 1. Their average number is about 30,000 per register and they decrease as discourse becomes more formal.

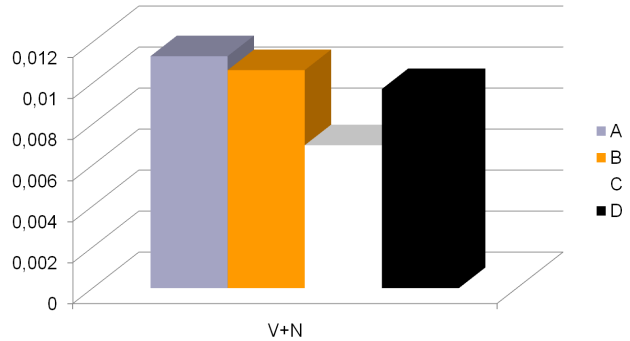


Diagram 1: V+N

The main verbs in Greek language that are used in support verb constructions are *kano* (to do/to make) and *eho* (to have), which is also used as auxiliary verb. As discussed previously (Picture 1), the frequency of support verb expressions is higher at the informal, conversational discourse. This is not the case though with *have*, which does not have the same connotation as *kano* as a very colloquial verb.

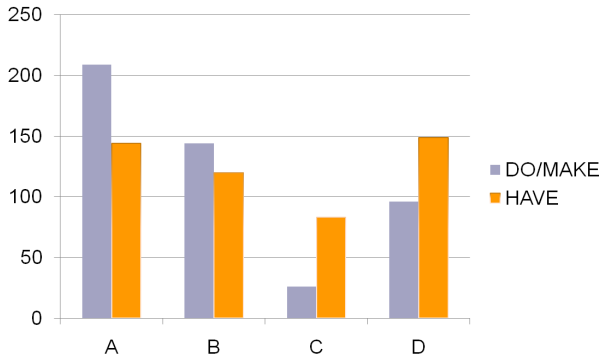


Diagram 2: verbs kano (to do make), eho (to have)

At the end of this procedure, there were extracted 185 predicative nouns, which they collocate with other 315 support verbs besides *kano* (to do/make) and *eho* (to have). The theoretical and methodological approach we used as a tool to analyse verb structures is based on the "lexique-grammaire" or lexicon-grammar tables[6], which describe each verb via a basic sentence and divides them into classes, according to their basic structures and their common transformational and distributional features. The extensive study of these nouns resulted to the building of 5 lexicon-grammar tables in order to describe their syntactic structure.

Another challenge were the idioms or fixed expressions, because their meaning, most of the times, is not derived by their components. As it was expected, idioms were abundant in the informal web discourse, and almost absent in the formal, academic discourse. As regards an automatic retrieval, based on the most unexpected words either semantically or syntactically, this would be a very difficult task, because the picture is obscured by many idioms from the general language.

The idioms reflecting an aspect of the stock market language had to do either with an upwards or downwards movement or an aggressive movement, e.g. *turn the stock into rocket, a plane, a jumbo, make a rally, or take a plunge* (literary translation from Greek).

	A	**B**	**C**	**D**
V+N	48815	53479	39105	34440
kano matches	8413	4174	466	1589
kano+N matches	2251	1582	183	434
(diff) collocates *kano*	1869	1228	229	619
collocates N of *kano*	806	620	94	249
- genitive	756	473	64	205
- nominative, adjectives, adverbs, name entities	559	358	43	164
- Certain semantic features (human, feelings , time)	237	163	33	108
Npred (a verb for each one)	**209**	**144**	**26**	**96**
Possible Idioms	**293**	**181**	**1**	**12**

Table 5: Extraction of predicative nouns and idioms

In Table 5, the procedure is following procedure is shown: after finding the concordances of the verb kano, nouns were extracted from its collocates. This was a more effective procedure than extracting collocations <kano> + <N>, since in Greek the complement of the verb is not always at the right part of it. Afterwards, all the nouns in genitive were excluded, as the support verb expressions are formed in Greek with accusative. Nominative was also excluded, in cases that there was no ambiguity, e.g. when the word was of male gender.

After extracting also name entities, and words with certain semantic features, we had manageable results. At the end, predicative nouns were extracted on the basis of the related verb for each one of them.

4. Perspectives of the corpus under study

Waiting for the appropriate interface, the Stock Exchange Corpus will soon be made available on line, at the site of the Laboratory of Translation and Language Processing of the Department of French Language and Literature in the Aristotle University of Thessaloniki (http://www.frl.auth.gr), in order to be used either as an autonomous resource for the study of a specialized vocabulary or as a part of a general corpus of Greek language. Our future plans include the completion of a comparable Stock Exchange Corpus in English.

References

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

Behr, I., (2007)"Langue, economie, entreprise : Le travail des mots", *Langue, discours et societe : Allemagne, Autriche, Pays-Bas,* no. 5-6, Paris : Presses Sorbonne nouvelle.

Benson, M. (1985) Collocations and Idioms, in R. Ilson (ed.) *Dictionaries, Lexicography and Language Learning* (ELT Documents 120), pp. 61–68. Oxford: Pergamon Press.

Biber, D. (1988) Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D. (1993) Using register diversified corpora for general language studies. Computational Linguistics, 2: 219–41.

Biber, D. (1993). "Representativeness in corpus design". *Literary and Linguistic Computing*, 8, 1-15.

Biber, D. (1995) *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (2006) *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., & Conrad, S. (2009) *Register, genre, and style.* Cambridge: Cambridge University Press.

Biber, D., U. Connor & T.A. Upton (2007) *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.

BNC [The British National Corpus] (2001) Oxford: Oxford University Computing Services.

Collins COBUILD English Collocations on CD-ROM (1995) London: HarperCollins.Collins Dictionary: Español-Inglés, English-Spanish (2003) Barcelona/London: Grijalbo/HarperCollins.

Cowie, A.P. (1988) Stable and Creative Aspects of Vocabulary Use, in R. Carter and M. McCarthy (eds) Vocabulary and Language Teaching, pp. 126–39. London/New York: Longman.

Daniushina Y. V. (2010) "*Business linguistics and business discourse"*, in *Calidoscopio, Vol. 8, No 3.* Available at: http://www.unisinos.br/revistas/index.php/calidoscopio/article/viewArticle/294.

Ferraresi, A., E. Zanchetta, M. Baroni and S. Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgarriff and S. Sharoff (eds.) *Proceedings of the 4th Web as Corpus Workshop (WAC-4) – Can we beat Google?*, Marrakech, 1 June 2008.

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

Frérot, C. and A. Josselin-Leray (2007), Enriching Bilingual Dictionaries with Corpus-Based Data: First Steps Towards an Improved Description of Verbs in General Bilingual Dictionaries Thanks to a Popular-Science Corpus, Talk at *4th Corpus Linguistics Conference*, 27th - 30th July 2007, University of Birmingham, Birmingham, UK. Electronically published http://ucrel.lancs.ac.uk/publications/CL2007/paper/84_Paper.pdf

Gavioli, L. (2002) "Some thoughts on the problem of representing ESP through Small Corpora" in Ketterman, B. And Marko, G. (eds.) *Teaching and learning by doing corpus analysis : proceedings of the Fourth International Conference on Teaching and Language Corpora*, Graz 19-24 July, 2000.

Goutsos, D. (2003) "Corpus of Greek Texts: Design and implementation" [In Greek]. *Proceedings of the 6th International Conference of Greek Linguistics, University of Crete, 18-21 September 2003*. CD-Rom publication.

Gross, M. (1975) *Méthodes en syntaxe*, Paris: Hermann.

Harris, Z. (1951) *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Holtz, M. (2007) "Corpus-based analysis of verb/noun collocations in interdisciplinary registers". Talk at *4th Corpus Linguistics Conference*, 27th - 30th July 2007, University of Birmingham, Birmingham, UK. Electronically published http://www.corpus.bham.ac.uk/corplingproceedings07/paper/14_Paper.pdf

Kelly-Holmes, H. & G. Mautner (2010) *Language and the market*, Basingstoke: Palgrave-Macmillan.

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

L'Homme, M.-C. (1998) 'Le statut du verbe en langue de spécialité et sa description lexicographique'. *Cahiers de lexicologie*, 73–2, 61–84.

L'Homme, M.-C. (2002) What can Verbs and Adjectives Tell us about Terms?, *Sixth International Conference on Terminology and Knowledge Engineering,* Nancy, 28–30 août 2002, 65–70.

L'Homme, M.-C. (2004) *La Terminologie : principes et techniques*, Montréal, Presses de l'Université de Montréal.

McEnery, T., R. Xiao, Y. Tono, (2006) *Corpus-based language studies: an advanced resource book*. Routledge Applied Linguistics.

Pomikalek, J., P. Rychly and A. Kilgarriff, (2009) Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics. Special Issue of Research in Computing Science Vol 41, Mexico City.

Trosborg, A.; Jorgensen, P.E. (eds.) (2005) *Business discourse. Texts and contexts*. Bern, Peter Lang.

Unitex, http://www-igm.univ-mlv.fr/~unitex/index.php

Wilkinson, M. (2006). Legal aspects of compiling corpora to be used as translation resources: questions of copyright. *Accurapid* Vol 10 (2). Online at: http://www.accurapid.com/journal/36corpus.htm (June 2011)