EXTRACTING TERMS FROM AN ENGLISH-GREEK POPULAR SCIENCE PARALLEL CORPUS FOR TRANSLATION TEACHING PURPOSES

by

Maria Matsira

A thesis submitted to the School of Humanities of the University of Birmingham for the degree of MPhil(B) in Corpus Linguistics

Department of English
School of Humanities
The University of Birmingham
2007

ABSTRACT

This thesis is focused on the compilation and analysis of a parallel corpus of popular science texts, i.e. articles appearing in a wide circulation popular science magazine and their translations. The stimulus of the research is translation teachers' regular practice of using articles of this genre as teaching material. The goal of this study is to introduce a methodology for extracting terminology for translation teaching purposes, which can be easily understood and implemented by both translation teachers and students using readily available commercial software.

Drawing on the fields of Corpus Linguistics, Translation Studies and Terminology on a theoretical level, this thesis follows the steps of 1) the creation of a translational English-Greek popular science corpus 2) its subdivision to smaller thematic sub-corpora and 3) its analysis (quantitative and qualitative) towards the extraction of candidate terms which, after being filtered through technical dictionaries, form single and multi-word term lists.

Overall, this thesis outlines the procedure of decision-making steps taken to derive the keywords and the criteria employed for regarding them as terms.

ACKNOWLEDGEMENTS

I am totally indebted to my supervisor Mr Philip King for his assistance throughout the research and writing of this thesis.

I am also grateful to my parents for their financial but mainly psychological support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background

Corpora have been extensively used so far and have served, with great success, a whole range of linguistic applications, from confirming or disproving grammatical rules, indicating frequent language patterns and verifying standardized language uses to bringing to light extra-linguistic factors, concerning stylistics, pragmatics, etc. that are language, genre or even text-specific.

One of these linguistic applications is also translation; a misjudged and for many years ignored practice (Baker 1993: 234), despite its great importance to the intercultural bonding among people. Only recently, have corpora been introduced to the field of translation studies and there are many potential applications. Translation's affiliation with corpora has, in fact, many things to offer, since investigating the language comparatively always gives to researchers more material to look for and new things to explore.

More especially, comparable but mainly parallel (or translation) corpora (Baker 1995, Hatim 2001, Olohan 2004) investigate the various translation phenomena, and attempt to give answers to questions like: are there specific translation strategies which are usually followed by translators? Are there translation norms or translation universals (Baker 1993: 243-245)? Or what does translation language look like?

An area which has been frequently investigated by the use of translation corpora is that of terminology. Language for special purposes –henceforth LSP– texts are quite often in translation, representing different linguistic genres and being of great interest to translators, and to a further extent to translation students, in terms of decoding their language's special nature; since one of translators' first steps towards the accomplishment of the translation task is the finding of the text's specific terminology which is often regarded as one of the main

obstacles to the understanding and especially to the rendering of the original text into another language:

> "Terminology is all the words I don't know and need to find out." (Bononno 2000: 648)

All of the above are research points of this study, whose topic is the extraction of terms from a parallel English-Greek corpus consisted of Scientific American recent articles and their translations which have appeared in the Greek version of the same magazine. The corpus was originally compiled to serve the process of terminology extraction. The analysis can be used by translation teachers as a pilot method to teach translation students how to extract terms from translation corpora by using commercially available software programs (Wordsmith 3.0 and Multiconc). To the purpose of this study, keyword frequency lists for the whole corpus, each sub-corpus and each text for both languages have been extracted and they have been compared to each other: each text against the whole corpus, each text against its relevant sub-corpus and each text against each other text of the same sub-corpus, to get lists with candidate terms which will be further investigated whether they constitute or not technical terms.

To support and scientifically ground the findings, four (one English monolingual and three bilingual: two English-Greek, Greek-English and one English-Greek) technical dictionaries have been used along with a wide range of internet resources, to compare the terms in question.

In addition to the method of terms validation in dictionaries, a range of arguments, mainly based on the analysis of the concordance lines and the words that the candidate terms collocate with, was built upon the methodology of this thesis to complement the writer's

hypothesis concerning the ability of popular science parallel corpora to yield terms that can be used as valuable material to translation teaching process.

1.2 Hypothesis of the thesis project

The hypothesis upon which this study is built can be outlined as following: "using parallel texts and available commercial software, is it possible to create a plausible list of technical terms, which can match technical terms generally recognized in the literature, or extracted by some other means, such as dictionaries?"

Due to the lack of previous background material, –existence of collections of popular science articles and especially from Scientific American past issues (for similar study see Pearson 2003:15-24) but not for Greek language– the task of this study was first to compile an appropriate corpus and then to analyze it.

No taggers or lemmatizers were used to analyze the corpus, due to the fact that there were not any readily available for Modern Greek. However, a detailed quantitative and qualitative analysis has been made for every sub-corpus, in order to show the depth of the analysis. The filtering of the results from technical dictionaries and web sources has been done to reinforce the scientific nature of the results.

1.3 Structure of the thesis

The paper is divided in three main parts: theory, methodology, and analysis.

The theory part contains the following sections: corpus linguistics, translation studies and corpora, and terminology, which constitute the basis upon which the whole analysis is grounded. Corpus linguistics because it is the methodological approach we use in this study to achieve our goal. Translation studies because it constitutes the stimulation and the reason for this study, and terminology because it is the centre point of this study.

The methodology part constitutes an analytical explanation of the path we followed to accomplish this task, as well as an extensive justification of the decisions we made and the reasons that led us to make them. The methodology part contains: the know-how of the corpus compilation, including the articles collection, the articles editing, the bilingual alignment, the division into sub-corpora, the criteria for the division into sub-corpora and the criteria for the analysis.

Finally, the analysis part includes the analysis of every sub-corpus which is divided into four parts:

- an overview which gives general information of every sub-corpus, including the number of articles, the number of words and the relevant topic, and also the reasons for which they were selected for this sub-corpus and some general issues, problems or differences that emerged from the fact that the corpus is parallel and two different languages are involved.
- the quantitative analysis which is actually the analysis of the keyword frequency lists taken by Wordsmith 3.0. This includes the comments on the lists as they are formed by the program itself; the comparison of the positions of the candidate terms within the two different language lists; their frequency within the article; their frequency within the reference corpus and their keyness as a result of the comparison of the two above-mentioned frequency lists. Multiconc parallel concordances have also been used at this stage of analysis, when this has been regarded as necessary.
- the qualitative analysis that is mainly based on the results which emerged from the quantitative analysis. This includes a detailed analysis of the candidate terms and focuses on the most troublesome ones; and it also includes the looking up at the concordances with the help of Multiconc and/or Wordsmith 3.0.

- the dictionary verification which involves the checking of the candidate terms in technical dictionaries as well as online glossaries and encyclopedic information, in order to discover which of them appear in there, as a result of standardization. Here, we have to clarify that this is not about a comparison between corpora and dictionaries or an evaluation of the already existing dictionaries with the view to criticising them. We simply use the dictionaries to support (verify or reject) our introspection about some words being terms.
- the summary and remarks section which contains the comments on the findings emerging from the total of the analyses and some further remarks on what has been previously analyzed during the three stages of the analysis.

2. THEORY

2.1 Corpus Linguistics

2.1.1 Defining a corpus

> "One of the principle uses of a corpus is to identify what is central and typical in the language." (Sinclair 1991: 17)

Before starting outlining the nature and the special characteristics of Corpus Linguistics, we would like to quote here definitions of 'corpus', given by scholars over the years. Some of the most representative ones are the following, which are cited in chronological order:

> "A corpus is a collection of naturally-occurring language text, chosen to characterize a state of variety of a language." (Sinclair 1991: 171)

> "A corpus is a large and principled collection of natural texts (Biber 1998: 12)

> "A corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description." (Kennedy 1998: 1)

> "A corpus is a body of texts assembled in some principled way." (Kenny 2001: 22)

> "Corpus is a text collection which has been designed for linguistic research, in order to represent some aspect of language." (Stubbs 2001: 25)

> "A corpus can be defined as a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis." (Tognini-Bonelli 2001: 2)

Common to all the above definitions is the term 'text' either in the phrase 'body of texts' or in the phrase 'collection of texts'. The two first definitions, those of Sinclair and Biber, refer also to natural language processing, a trend in linguistics that shifted scholars interest from language structure to language use (Biber 1998: 1) and converted it from a pure rationalistic to a more empirical view of language (McEnery & Wilson 1996: 4).

2.1.2 Corpus Linguistics: branch of linguistics, methodology or discipline?

Corpus Linguistics, as emerges from above, is the branch of linguistics, albeit in a non-conventional sense (McEnery & Wilson 1996: 2), that investigates corpora. It is more concerned on how natural language works beyond grammatical rules and syntactic limitations and it is, as we mentioned above, an empirical study of language (Tognini-Bonelli 2001: 2).

However, a corpus is not an object of study. Corpus provides a way to investigate the inner nature of natural language. Consequently, corpus linguistics is not a branch of study but a methodology to the service of linguistics, and maybe something even more than that (McEnery & Wilson 1996: 2; Tognini-Bonelli 2001: 1).

> "Corpus linguistics is not an end in itself but is one source of evidence for improving descriptions of the structure and use of languages, and for various applications, including the processing of natural language by machine and understanding how to learn or teaching language." (Kennedy 1998: 1)

As Tognini-Bonelli (2001: 64, 84) also confirms, corpus is an evidence of what is there in the language and it can shed light to something new or well-hidden in language. She also

supports that corpus linguistics is a discipline on its own and it has its own field of study (2001: 49).

The above argumentation on corpus linguistics' nature brings to mind Saussure's dualism on langue and parole (Stubbs 1996: 41), Chomsky's argument for competence and against performance in language (Kennedy 1998:7; McEnery & Wilson 1996:5), and the eternal debate between quantitative and qualitative analysis of language. Scholars' dilemma was about being focused on introspection or on empirical data for the study of language. Corpus linguistics, however, can verify humans' introspections about language through real facts that appear in it.

2.1.3 Historical Background

Corpus Linguistics appeared in the 1950's but at that time, the availability of computers was limited and their computing power small. The image that Corpus linguistics presented in the 1950's was far enough from what Chomsky was thinking of language. He protested against the utility of corpora in the linguistic research, claiming that they cannot be representative of the language they examine, since they are finite and error-prone. As a pure pragmatist, he prioritized human's competence in language over human's performance in it. But one thing is certain, that Chomsky could not predict the evolution that corpora would have nowadays by the penetration of computers in their study (McEnery & Wilson 1996: 4-10).

2.1.4 The role of computers

> "A corpus is a collection of texts, selected and compiled according to specific criteria. The texts are held in electronic format, i.e. as computer files, so that various kinds of computer tools, i.e. software, can be used to carry out analysis on them." (Olohan 2004: 1)

Olohan's corpus definition introduces the aspect of computers in the study of language through corpora. Today, no corpus analysis can be conducted without computers due to the size of corpora as well as to time and money restrictions. Thus, various software programs are being constantly designed or improved to deal with the fast progress of corpus linguistics. To the same direction, Tognini-Bonelli in her book "Corpus Linguistics at Work" (2001), gives another definition of corpora, prioritizing this time the contribution of computers to the evolution of linguistics:

> "A corpus is taken to be a computerized collection of authentic texts, amenable to automatic or semi-automatic processing or analysis. The texts are selected according to explicit criteria in order to capture the regularities of a language, a language variety or a sub-language (Tognini-Bonelli 2001: 55)

Nowadays, anyone who owns a PC and the appropriate software can conduct a research on corpora. Computers have the power to handle large numbers of texts and they can process them quickly. However, the results the computer exports are nothing but numbers that need to be interpreted and that is where human brain comes into play. Only the researcher can, actually, relate percentages to grammar patters and see what frequencies reveal about what is typical in language (Kennedy 1998: 5).

2.1.5 Types of corpora

According to Hunston's classification of corpora (2002: 14-16), we have the following corpus categories:

- Specialized corpora: these corpora represent a specific kind of language, e.g. corpora that study language diachronically; corpora that investigate the degree of idiomaticity of certain languages; or corpora that treat a specific topic of language, like environment and others.

- General corpora: these corpora include as many different types of texts as possible. Their special characteristic is their considerably big size. Such corpora are the British National Corpus and the Bank of English.
- Learner corpora: these corpora are consisted of pieces of language produced by students as opposed to language produced by native speakers.
- Pedagogic corpora: these are 'teacher's corpora" designed and compiled to deal with the needs the students may face during a language course.
- Historical or diachronic corpora: are the corpora which investigate language over time and they are mostly interested to capture any significant changes that a language undergoes during its evolution.
- Monitor corpora: are the corpora that are made to "track current changes in a language" and to be enriched every now and then. They are always 'balanced' (the notion of balance will be explained in the next section, Corpus design and compilation). The Monitor corpus is also named dynamic corpus (Sinclair 1991: 24) as opposed to the notion of static corpus whose size is finite and the genres to be included, already fixed (Kennedy 1998:60).
- Comparable corpora: two or more corpora in different languages or even in the same language that are built in order to compare the languages or the "different varieties of one language".
- Parallel corpora: two or more corpora in different languages that are connected to each other with translational relations, i.e. the one corpus contains the translations of the texts which consist the other (a more thorough description of comparable and parallel corpora will be given in the Translation studies and corpora section).

2.1.6 Corpus design and compilation

Nowadays, there are many already made and also free or limited access corpora, like Brown corpus, LOB corpus, BNC and the Bank of English, However, individual researchers may need to compile their own corpus to use it to the purposes of their own research.

In general, there are guidelines to help somebody to compile a corpus, but anyone is more or less free to design it as long as it fits to their own study. These guidelines concern the aspects of size, balance and representativeness (Hunston 2002; Kennedy 1998; McEnery & Wilson 1996; Tognini-Bonelli 2001).

> "The feasible size of a corpus is not limited so much by the capacity of a computer to store it, as by the speed and efficiency of the access software." (Hunston 2002: 25)

However, as we previously said, the size depends on the purpose of use. The issues of balance and representativeness are related to the issue of corpus size; since a corpus should contain samples or full texts of more or less the same size, so as to be balanced and should include more or less all genres or all kinds of a specific genre to be representative. In some cases, when the language is being investigated diachronically, the corpus should also be representative of all periods of a language.

> "A corpus in modern linguistics […] might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration." (McEnery & Wilson 1996: 24)

These are the factors a corpus compiler should take into account in order to design an adequate and accurate corpus. However there are further steps to be taken when one wishes to conduct a more detailed analysis. These steps are part of a process known as corpus

annotation and it includes word-class tagging and parsing (Kennedy 1998; McEnery & Wilson 1996). Roughly speaking tagging refers to the marking up of each word of a corpus for its grammatical category (noun, adjective, verb, etc.) and parsing to the syntactic marking up of the already tagged words (Hunston 2002: 18-19). To these two, we could also add the process of lemmatization (Kennedy 1998: 206), which is very useful in a corpus of highly inflected languages and the semantic annotation (Hunston 2002: 88, Kennedy 1998: 225), which has not been made completely feasible due to the fact that languages are dynamic, not static, among other things, and its descriptions are always trying to catch up with its constant evolution.

2.1.7 Corpus analysis: Methods and procedures

The analysis of a corpus is the most important step in a corpus enterprise since it is the stage where results are extracted and conclusions are drawn. After that, corpus researchers have accomplished their task and reached their target; however in order to achieve that, a range of methods has been followed and various corpus tools have been used.

These techniques can be divided into three categories: the frequency lists, the concordance lines, which their analysis focus on the study of collocations, and the statistics, i.e. the mathematical representation of the findings emerged from the two previous methods. Today, there are software programs which are able to do all the above.

> "A corpus does not contain new information about language, but the software offers us a new perspective on the familiar." (Hunston 2002: 3)

These programs based on special algorithms can get wordlists or keyword lists from a corpus, extract concordances lines and lists with the collocates, as well as calculate MI and T-scores to ground the results statistically (McEnery & Wilson 1996) (see also Oakes 1998 for a more extensive and detailed analysis).

2.1.8 Corpus-based or corpus-driven?

Generally, in corpus linguistics there are two approaches that are applied nowadays. These are: the corpus-driven approach (CDA) and the corpus-based approach (CBA) (Tognini-Bonelli 2001). In the corpus-driven approach, the corpus serves as the starting point of the research. By analyzing and observing it, researchers "detect linguistic phenomena without prior assumptions and expectations" (Storjohann 2005:4); whereas in the corpus-based approach the corpus serves as "an additional supporting material" (Storjohann 2005:6) and researchers use it to prove or disprove introspections they have about language.

Due to the fact that it is more or less believed that all language special traits have been mapped corpus-driven approach has not been frequently preferred so far by researchers. However, nowadays, among the corpus linguists there is a turn of interest to the corpus-driven approach because it appears to be "holistic and systematic" (Storjohann 2005: 8).

2.2 Translation Studies and Corpora

2.2.1 Translation Studies: An Overview

Translation Studies is a wide academic field, which, according to Baker (1998b, cited in Olohan 2004:1), apart from translation, "incorporates also interpreting, dubbing and subtitling". It can be considered as a relatively new discipline, as it has only been established as an academic subject for approximately fifty years (Munday 2001:5).

Nonetheless, translation has always been a matter of discussion and controversy. Its "artificial" character, along with the idea that translation is a distortion and "betrayal" of an original text (Baker 1993:233), hampered translation for many years in its development as a separate field of study.

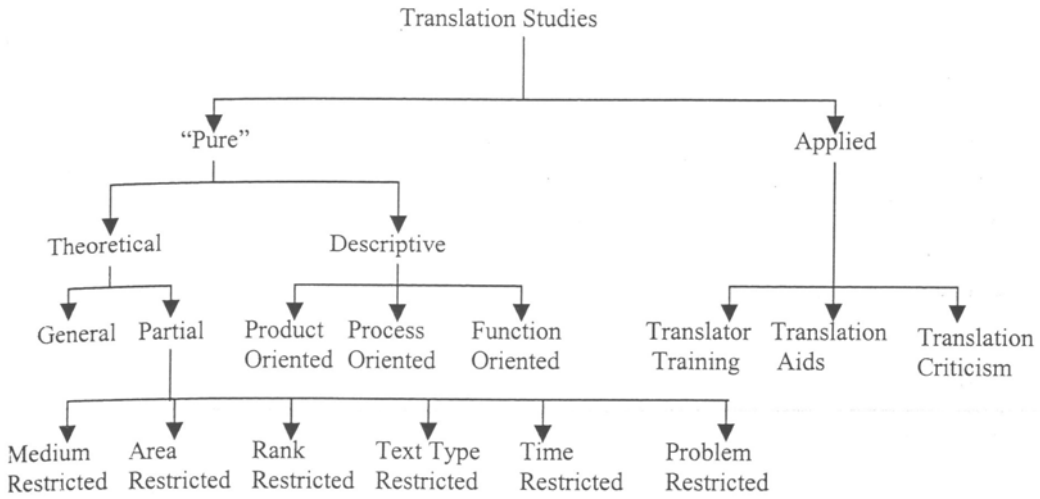
Nowadays, translation studies –James Holmes was the first to introduce this term– seeks its place among the other disciplines of language, having its own object of study, its own

methodology, as well as its own research tools. Translation Studies' object of study is given in Holmes' "The Name and the Nature of Translation Studies" (1988b/2000:181):

> "Translation Studies is concerned with the complex of problems clustered round the phenomenon of translating and translations."

Holmes presents translation studies in the form of a map, which divides into two big areas: the "Pure" translation studies and the Applied translation studies. "Pure" translation studies encompasses the Theoretical and the Descriptive Translation Studies (DTS) and its objective is, on the one hand, to "describe the phenomena of translating and translations" and, on the other hand, to "establish the principles that can describe and explain such phenomena" (1988b/2000:184).

Figure 1 Holmes map of Translation Studies



Descriptive Translation Studies is "a) product-oriented, b) function-oriented and c) process-oriented". In other words, it is an empirical discipline, which -unlike others- is not based on a theory and seeks for its application, but it has as its starting point the undeniable existence of a product (translated text), the inscrutable nature of a process (translation) and its rules, which tries to unfold.

Theoretical Translation Studies can be General (when including a general translation theory) or Partial (when including partial translation theories). These Partial theories may be: Medium-restricted theories (human vs machine translation; written translation vs interpreting), Area-restricted theories (emerging in different languages or different cultures), Rank-restricted theories (reflecting various linguistic ranks/levels), Text-type restricted theories (dependent on the type of text or the genre), Time-restricted theories (concerning diachronic vs synchronic study of translations), Problem-restricted theories (assigned by the solution of a specific translation problem, e.g. metaphors, proper names, etc.).

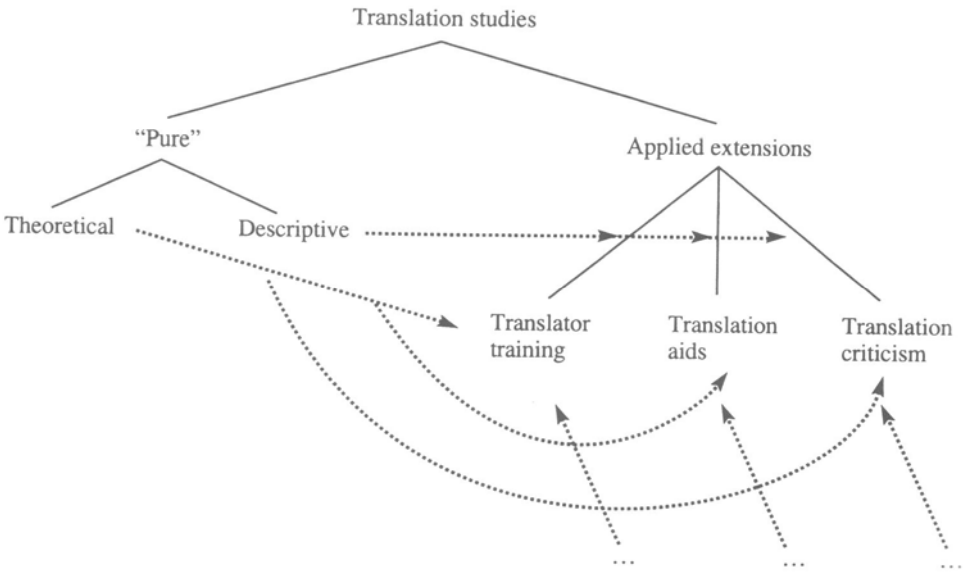
Finally, the area of Applied Translation Studies –to return to the first main division of Translation Studies- comprises the following categories: Translator training (translating as a method in the process of foreign-language acquisition or translating as a method in translators' training), Translation Aids (dictionaries, termbanks, grammars, etc.), Translation policy (translators should consult their colleagues on issues that concern translations as well as the role of translator) and Translation criticism.

In "The Name and Nature of Translation Studies" (1988b/2000: 191), Holmes introduces briefly two other, very important types of research: the study of Translation Studies (historical study of translation theory, translation description and study of applied translation studies) and the methodological/metatheoretical study of Translation Studies (study of the methods and models used in the discipline).

The point in Holmes' study from which Toury starts his argument in DTS is his comment on the type of relationship among the three main branches: Theoretical, Descriptive and Applied Translation Studies. This relationship is dialectical in terms of "each branch both providing insights for and using insights from the other two" (Baker 1998: 279).

The difference between Holmes' DTS and Toury's DTS, as we previously said, is seen in the distinction between the three main areas of Translation Studies. Toury claims that the relationship between Theoretical, Descriptive Translation Studies and Applied Translation Studies –which he called Applied Extensions– is unidirectional, since in every branch there are some "bridging rules" which prevent the automatic transition from the one to the other (Toury 1995:18).

Figure 2 Toury's map of relation between Translation Studies and its applied extensions



Toury's Descriptive Translation Studies (DTS) is the first conscious attempt to examine translation systematically by:

1. "looking at the significance or acceptability of the target culture system."
2. "searching for similarities and difference between ST and TT."
3. "using the previous two, in order to help translator to make the right decisions in future translations."

(Toury 1995: 36-39)

Toury's investigation of translation behaviour led him to the ascertainment that in translation, there are norms. Norms are the inter-level between translator's constraints of rules (how the

translation should be) and idiosyncrasies (what choices the translators finally make with regard to their translation). However, norms are socially and culturally determined and change over time (Toury 1995: 56, 62). There are three types of norms in translation: initial norms, preliminary norms and operational norms (1995:56-59).

"Initial norm" refers to the translators' initial choice as to what norms they will adopt in their translation. If they choose to adhere to the source text, then this can be considered as the translated text's adequacy with regards to the source text and if they choose to adhere to the target text, this shows the translated text's acceptability by target culture.

Preliminary norms concern the policy that the translator has to make concerning the text to be translated and the directness of translation –always with regard to the target language audience.

Operational norms, finally, concern the decisions the translator makes during the translation process and 1) affect the matrix of the text (matricial norms): how the text material is distributed, the amount of the text that will be translated, the changes that will take place in text segmentation (omissions, additions, etc.) and 2) determine the material –textual and linguistic– that is going to be used to the creation of the target text (textual-linguistic norms).

Overall, norms can be found both in the textual and extra-textual level and they constitute the basis upon which the laws of translational behaviour will be grounded (Toury 1995:65).

Two laws in translation that are suggested by Toury (1995:267-279) are the law of growing standardization and the law of interference. The former can be recognized by the following facts: when the source language "textemes" become "repertoremes" (Toury 1995:268) in the target language/culture, that is, the textual relations of the source text are modified leading to a completely different text which is representative of the target language/culture; when the items that are chosen in the translated text are of lower lever than those of the source text;

when translation assumes a more peripheral and conservative position in the target language; when the source text textemes are replaced by equivalent combinations in the target text, so as to serve the same purpose (1995:267-274). The latter can be recognized by facts like: in general there is a tendency of the target text to maintain the make-up of the source text, especially when the translator has decided to use source text as a crucial factor; the degree of interference depends on the "prestige" of the source language but also on the different linguistic and textual levels in which the translated text is being realised (1995:274-279).

The term laws in translation, though, is characterized by Kenny (2001:54) as an "unfortunate" choice, since they can better be regarded as "hypotheses", having a prescriptive character rather than a binding obligation, as the term "laws" implies in natural sciences. (Kenny 2001:54).

Toury, in a previous work of his, uses the term "universals" to refer to "general tendencies" that appear frequently in translations, but he denies labelling them as norms (Toury 1978 cited in Kenny 2001:52-53). Practically, both "norms" and "universals" explain recurring patterns in translation. In reality, not even Toury himself can clearly distinguish those two.

A lot of researchers have investigated the universal features of translation, but, here, we will adopt Baker's description as it is presented in her work, "Corpus Linguistics and Translation Studies Implications and Applications" (1993: 243-245).

Therefore, as types of universals of translation, we distinguish the following:

1. explicitation: the translated text presents a higher level of explicitness in comparison with the source text.
2. simplification: it can be lexical, syntactic and stylistic (Laviosa 1998:288). This stands for the tendency of the translator to simplify in the translation, complex structures found in the source text.

3. normalization: represented by translator's tendency to adopt in his/her translation conventional grammatical patterns.
4. avoidance of repetitions: the translator tends to avoid in the translated text repeating utterances of the source text.
5. Toury's universal of growing standardisation: "translations overrepresent features of their host environment in order to make up for the fact they were not originally meant to function in that environment" (Vanderauwera 1985:11 cited in Baker 1993:245).
6. Toury's universal of interference: the translation tends to maintain recurring patterns of the source text to the extent that the cohesion of the new text reveals that it is a translation (Baker 1993:243).

According to various researchers, norms, laws and universals constitute all distinctive features of translation and the best way for these to be examined is within their natural environment, the translated text. A valuable and reliable source of information about texts is corpus and as we shall see below it has already contributed much to the enrichment of the knowledge on translation research.

2.2.2 Corpus Linguistics and Translation Studies: when the method met the discipline

The use of corpus techniques for translation purposes has a short history, despite the fact that corpus linguistics investigates language for more than fifty years (McEnery and Wilson 1996:1). This can be explained by two facts:

1. "the negative image of mainstream linguistics that was developed within translation studies during the 80s and 90s, according to which translation was related neither with the linguistic patterns translators used, nor with its social and ideological context."

2. “the traditional attitude of corpus linguists to translated text. They considered it as non representative of the language being studied.”

(Baker 1999:282)

Nowadays, however, there is a considerable amount of literature on corpus-based translation studies, as it seems that translation has found a valuable “research tool which enables it to be studied in a number of ways and through a variety of methods” (Olohan 2004:1).

Language studies, as we previously mentioned, in the case of translation, are empirical studies. The intuitions of language scholars, however, can form hypotheses, which can be tested (and can be either verified or demolished) by the systematic study of a corpus (Olohan 2004:14-15).

All in all, Baker’s statement reflects adequately what corpora are to translation:

> “The profound effect that corpora will have on translation studies, in my view, will be a consequence of their enabling us to identify features of translated text which will help us understand what translation is and how it works.” (Baker 1993: 243).

2.2.3 Types of corpora used in translation studies research

At this point, I will present the types of corpora used for the description and analysis of translation, following Kenny’s – and not only his- typology (2001:58-65).

Monolingual single corpora

Monolingual single corpora are the “corpora that contain texts in one language only”. They include either “original texts in one language” (non-translational) or “translations in one language” (translational). Both are used for translation purposes; even though only translational include translated texts. The non-translationals promote a better knowledge of

the source language (suggested to translation trainees), while the translational ones identify features that are only representative in translated language (Kenny 2001:58-59).

Comparable corpora

Comparable is a corpus that is composed by two (or more) monolingual, non-translational corpora or two (or more) monolingual, translational corpora or, still, a monolingual corpus of original texts and a monolingual corpus of translated texts, which are connected on the basis of comparison; they are all considered to be comparable corpora (ibid.). The use of the later is mainly focused on the "investigation of features that are particularly characteristic of translated language as opposed to source language" (ibid.).

Parallel corpora

The corpora that contain the original texts in one language and their translations in another language are called "parallel corpora". These corpora need to be aligned, with the help of either machine-aided translation programmes or with corpus software, specially designed to align corpora (to a word, sentence or paragraph level), as well as to analyse them (concordance lines, word-lists, key-word-lists, statistics, clusters).

Parallel corpora interest is exclusively focused on the examination of particular features in translations as opposed to their originals; therefore, they are compiled "according to some principles: translator, school of translators, period, text-type, text-linguistic phenomenon, or any other principle which could be given a justification" (Toury 1995:38).

2.3 Terminology

2.3.1 Terminology: definition, historic background and aspects

Let us now turn to consider what people mean when they refer to terminology.

According to Sager (1990/1996:2), "terminology is the study of and the field of activity concerned with the collection, description, processing and presentation of terms, i.e. lexical items belonging to specialized areas of usage of one or more languages."

However terminology as a word is used to denote three different concepts (Cabre 1999: 32):

a. "The principles and conceptual bases that govern the study of terms"

b. "The guidelines used in terminographic work"

c. "The set of terms of a particular special subject"

Terminology, as an activity is very akin to lexicography (Sager 1990/1996:2), but they differ basically on their approaches. That is to say, lexicography is mainly interested in the lexical representation of the concept, the word and the environment the word appears in, i.e. its context. On the other hand, terminology is mainly focused on the concept itself, without being so much concerned in the name that represents the term. In other words, terminology is mostly interested in the synchronic aspect of a specific term (Cabre 1999:33), i.e. its specific function and use within a text, which usually appears to be specialized.

The interest in terminology is closely related to the technological progress. Not only because the latter revolutionized terminology's extraction techniques but mainly because it made its establishment important and necessary. The starting point for terminology's scientific development is 1930's, when its theoretical foundations were grounded almost simultaneously by three schools: the Austrian School, the Soviet School and the Czech School (Cabre 1999:7).

Terminology's diachronic course can be divided, according to Cabre (1999:5) into four periods:

1. the origins (1930-1960)

2. the structuring field (1960-1975)
3. the boom (1975-1985)
4. the expansion (1985-present)

During the last three periods, terminology has been developed gradually along with computer progress, allowing us today to talk about automatic terminology extraction.

Terminology's most characteristic feature is its interdisciplinary nature. Being the study of terms, terminology (and terminography) is at the service of all sciences, including physics, chemistry, biology but also business and social sciences, providing its means to the processing and the classification of the old terms, as well as to the identification of new ones. As an independent study, however, it is mostly related to disciplines like lexicography/lexicology, logic, ontology, computer science and information science (Cabre 1999:8; Sager 1990/1996: 3-7), sharing with them a common ground, either because their object of study is similar, or because they use the same means.

From the above, we conclude that terminology cannot be described completely by the notion of discipline nor by that of methodology. As Sager points out, "we see terminology as a number of practices that have evolved around the creation of terms, their collection and explication and finally their presentation in various printed and electronic media." (1990/1996:1) To Sager's previous statement Cabre (1999:10) adds that "terminology is not an end in itself but addresses social needs and attempts to optimize communication among specialists by providing assistance either directly to translators or to committees concerned with the standardization of language."

2.3.2 Special languages, terms and standardization

> "Language is a complex, heterogeneous system made up of interrelated subsystems, each of which can be described at the phonological, morphological, lexical, syntactic and discourse levels." (Cabre 1999:56)

Both general and special languages constitute part of natural language as opposed to artificial languages, but their main difference is that in general language the codes that humans share to communicate are common to almost all language users; whereas in special (or specialized) languages, as variants of the general language (Cabre 1999:61), only a few people can understand and share its codes.

Special languages are related to terminology in that their lexicon consists of terms. "The items which are characterized by special reference within a discipline are the 'terms' of that discipline, and collectively form its 'terminology'; those which function in general reference over a variety of sublanguages are simply called 'words', and their totality the 'vocabulary'." (Sager 1990/1996:19) As the number of words in a language is finite, we may have different concepts being represented by the same names. In the case of terms, however, and because they are determined by the user of the special language as to what role they will play in it and what concept they will represent, we usually have one term for one concept. Nevertheless, this is not fixed, since it depends on "the conceptual properties of the discipline and on the goodwill and good intentions of users" (Sager 1990/1996:20) and that is where the importance of standardization lies.

The International Organization for Standardization (ISO) gives for standardization the following definition:

> "The process of formulating and applying rules for an orderly approach to a specific activity for the benefit and with the co-operation of all concerned, and in

particular for the promotion of optimum overall economy taking due account of functional conditions and safety requirements." (cited in Cabre 1999:195)

The reasons for which standardization is essential are cited in Sager (1990/1996:115):

1. for economy reasons: the choice of a term instead of another, which proved to be more trivial
2. for precision reasons: the choice of a term which represents more adequately and with bigger clarity a concept than another
3. for appropriateness reasons: the choice of a term which with respect to connotations is more appropriate than another

However, terminological standardization has a slightly different meaning. It may not refer to the naming of any commercial products, but it maintains the necessity of an approval by an authoritative body, on a term's predominance over the others as regards a certain concept.

Terminological standardization has, according to Cabre (1999:199) three different meanings:

1. the institutional standardization: it is the standardization made by a body
2. the international standardization: it is the standardization made by an international body
3. the non-interventionist standardization: it is the standardization made by the mutual accord between the terminological system monitors and the end-users of a specific field.

Reversing Guespin's and Laroussi's statement (cited in Cabre 1999:201) about what terminological standardization can describe, we would like to make our argument about what terminological standardization depends on. Thus, terminological standardization is

determined by a. the conditions that form the scientific discourse and the type of relationships between sciences and technical fields, b. the conditions in communication during knowledge and technology transfer and c. the forms that emerge from language conflict, which reflect the social and political context of the countries concerned.

2.3.3 Terminology users and translation

Sager, in his book "A practical course in terminology processing" distinguishes seven types of terminology users, after he makes a first discrimination between the user of terminology and the learner of terminology (1990/1996:197-199). The criterion for distinguishing them into those seven users' types is the kind of information they retrieve from term banks. Consequently we have:

1. the subject specialists who create term banks in order to look for the meaning or the spelling of a specialized term
2. the professional communication mediators, among them technical writers, translators and interpreters, who use term banks to extract accurate terms that can exist outside their contextual environment
3. the specialist lexicographers and terminologists who are the main creators and at the same time administrators of the term banks
4. the information and documentation specialists, like librarians and indexers who use term banks to identify and describe specialist documents
5. the language planners who work for the maintenance and the development of the natural language and they also take care of all the standardization matters that emerge in the language.

6. the professional language users such as publishers, language teachers, researchers in applied linguistics, etc. whose lexicological needs can be dealt by a flexible and reliable lexical database.

7. the general users of the language who, although, may not have a daily interest on specialized terms, when they express it, it is an urgent demand for the finding of the most accurate and appropriate term which can serve their needs.

Another distinction among terminology users is also made by Cabre (1999:11-12). She divides terminology users in two big categories: the direct users and the intermediaries. To the first category belong the specialists in each subject field and to the second belong the so-called "language professionals", like technical writers, translators and interpreters. Cabre also adds another category –not to the previous two but to that of users– that of terminologists, which includes also terminographers, neologists, language planners and information scientists. Their job is exclusively the processing of terms.

Translators, as we saw above, are one type of users of terminology. However, the notion of terminology in translation stands for two different things: a. the set of the specialized terms (usually unknown) of the source text, which the translator looks up in terminological banks and dictionaries, in order to unfold and understand their special meaning and b. the set of terms of the target language, which the translator looks up in bilingual term banks and bilingual dictionaries, so as to use them as translation equivalents to the terms of the source text. As Cabre points out: "terminological equivalence is the key to multilingual terminology." (1999:45)

2.3.4 Terminology extraction and corpora

As we previously mentioned computers have revolutionized language research in terms of automatic analysis of text and processing of large quantities of data. Terminology compilation has also been affected by this trend and nowadays a range of computer tools are at terminologists' disposal to assist them to accomplish their task.

Corpora constitute one of the most important sources to extract terminology, since they can handle large amounts of data as well as be analyzed automatically.

> "Text analysis of large corpora can be used to isolate new terms, and therefore new concepts, to discover the possible obsolescence of terms and their concepts and to highlight other changes in conceptual systems." (Cabre 1999:132)

According to many researchers, there are some standard methods for automatic terminology extraction (Penas,Verdejo, Gonzalo 2001: 2):

i. Term extraction via morphological analysis: POS tagging and shallow parsing

ii. Term weighting with statistical information.

iii. Term extraction via syntactical analysis, which is primarily based on the first method and requires beforehand POS tagging in order to be accomplished.

The automatisation of the terminological extraction, however, still faces serious problems, like (1) recognition and identification of complex terms (2) identification of the terminological nature of a lexical unit (3) appropriateness of a terminological unit to a specific domain. (Cabre; Bagot; Palatresi 2001: 54)

Another question that arises within terminology extraction is the issue of single-word terms (mono-lexical terms) and multi-word terms (poly-lexical terms) (Lemay, L'Homme, Drouin 2005: 227-255; Vintar 2001: 121-132). Terminologists think preferentially of nouns when they consider domain-specific concepts (Cmerjek, Curin 2001: 3; Lemay, L'Homme, Drouin

2005:227). However, these nouns can sometimes be noun phrases (NPs) that are constituted by several part of speech combinations, such as Noun-Noun collocations or Noun-Adjective collocations (Heid 1999). These can also be terminologically relevant, since “in general language, many collocates in noun-verb or noun-adjective collocations have a collocational meaning, i.e. are not understood in the same meaning as in contexts outside the collocation.” (Heid 1999)

All these are issues that are going to be directly or indirectly investigated in this thesis. But let us now see what methodological steps we took in order to achieve our goal.

3. METHODOLOGY

3.1 Corpus Compilation

The corpus used in this study was initially compiled for the needs of the present paper. The material for the building up of our corpus came from six past issues of Scientific American (November 2006–April 2007), and included the original version in English and the translated version in Greek. The selection of the articles was made on the basis of the Greek translations, since the problem was the difficulty of collecting data for the Greek corpus. The whole corpus consists of 90 articles (45 English and 45 Greek). The size of the English corpus is approximately 132.813 words (tokens) while the size of the Greek corpus is approximately 139.782 words (tokens).

The material was collected in two ways: the English part through the Internet and the Greek part through the laborious task of scanning, since there was no way to get access to the electronic issues of the Greek Scientific American. For the scanning a Greek OCR (Abbyy Fine Reader 8.0) was used; some of the editing had to be done manually.

After the data collection and editing, the next step was the alignment of the two corpora, which would enable us first to compare and then to attempt to extract candidate terms. For the alignment, a software program: Multiconc, created in the University of Birmingham, was used. Minmark 2.0 (a Multiconc tool) aligned the texts on paragraph level:

> "It is difficult to employ this approach at sentence level since a skilled translator may well translate one sentence by two, or two by one, three by two, and so on. This is the central problem of text alignment." (see Multiconc manual website)

During the alignment, we came across phenomena, such as omission or adaptation in the translated text, which were either translator's decisions or the moderator's or maybe even the

editor's, always with regard to the target audience. However, Multiconc can provide parallel concordances at sentence level, or, when no match appears at sentence level; the user can select a paragraph-level alignment.

The corpus was divided into sub-corpora according to topic. The topics and sub-topics are indicated in the header information accompanying each article in Scientific American. Therefore, taking that into account, we ended up with 7 sub-corpora: Biology/Anthropology, Energy/Environment/Geology, Medicine, Physics Planetology/Cosmology, Psychology and Technology. However, we should mention here that the sub-corpora contain different numbers of articles, since the collection of the material was made by only criterion their appearance in the issues of Scientific American (Greek edition) between November 2006 and April 2007.

3.2 Criteria for dividing the sub-corpora

This division into sub-corpora was made in order to facilitate terminology extraction. In other words, all articles dealing with a given area were gathered into one sub-corpus, so as to help researcher to collect terms that belong to the same or similar scientific field and organize them accordingly, afterwards. However, as one could notice, in some cases the topics of one sub-corpus may vary a lot, whereas in some other cases the topic of the sub-corpus is one, even if this entails very few articles involved and a considerably limited sub-corpus size.

3.3 Corpus Analysis

3.3.1 The use of Wordsmith 3.0 lists

In the analysis of the corpus, Wordsmith 3.0 was used to extract wordlists, keyword lists and concordances. Wordsmith 3.0 is not the latest version (the latest is Wordsmith 4.0) of this

software, however this is the only one that works with Greek, which is why it was used in this study.

The methods we followed here, however, are not purely automatic (maybe one could call them semi-automatic) since the means we had at our disposal for Greek were somewhat limited. For minor languages like Greek (Vintar 2001: 130), taggers are limited, and consequently the statistical analysis can only be done in terms of frequency.

As some researchers pointed out:

> "The relative frequency of a lexical unit in two different corpora is strongly linked to the importance of the unit in the corpora. The more frequently it appears in a corpus, the more likely it is to be significant in this corpus" (Lemay, L'Homme, Drouin 2005: 232); however, "alone, the frequency is not a robust metric to assess the terminological property of a candidate, but it does carry useful information, as does also the length of terms"(Patry; Langlais 2005:4).

First of all, we created wordlists and then keyword lists for every article as well as for its translations. This promotes the comparative analysis of the original and its translation as well the analysis across the articles of every sub-corpus. This helped us to get some reliable results about what is domain-specific within the corpora.

For the extraction of keyword lists, we used as reference corpora, the wordlist of the entire English corpus which we compiled for the purposes of this study (for our English analysis corpus), and the wordlist of the entire Greek corpus that emerged from the collection of the Scientific American articles (for our Greek analysis corpus). The keyword lists we retrieved represent also every article of each sub-corpus separately and the whole of the articles included in a sub-corpus.

The way keywords are calculated is, according to Wordsmith 3.0 manual (see website), the following: the frequency of each word in the smaller of the two wordlists is compared with the frequency of the same word in the reference wordlist. All words which appear in the smaller wordlist are included in the analysis, unless they are in a stop list. The keyness is a very important element of the Wordsmith tool because it computes one item's frequency in the small wordlist, the number of running words in the small wordlist, the item's frequency in the reference corpus, the number of the running words in the reference corpus and finally cross-tabulates all these. The element of keyness was used extensively in this study and a part of the results was actually based on it.

3.3.2 The use of Multiconc parallel concordances

Multiconc, apart from an alignment tool, was also used at the stage of quantitative analysis, in order to shed light to the obscure cases of poor matching across languages.

Overall, this method revealed cases of omission, adaptation, mismatching, errors in editing, translator's mistakes and others. This tool promoted significantly the comparative analysis of the corpora, with its useful method of the alignment "on the fly". The option of viewing parallel texts at paragraph level as well as sentence level was useful as it enabled certain ambiguous cases of correspondences to be sorted quite easily.

3.3.3 The use of Wordsmith 3.0 concordances

Wordsmith 3.0 concordance tool was also used, mainly during the qualitative analysis. The easy transition from keyword lists to concordance lines and then to the counting of collocates was the main advantage of this tool. The existence also of the Viewer & Aligner tool took over in cases Multiconc could not cope with.

3.4 Criteria for the analysis

Because it is impossible, due to lack of space and time, to analyze every sub-corpus and even more every article, out of 90, in this study, we had to make some decisions beforehand which would facilitate our way through this long and detailed study. Therefore, we decided to set the following criteria:

- no taggers or lemmatizers were used in this study, since there were not any readily available for Modern Greek. As a consequence, no stop lists with function words for both languages were created, except the one the software program itself provided. This was taken into account but we regarded it as not necessary. All word forms of a lemma were checked. For English these were the singular and plural form as we were only interested in nouns (see two criteria below); for Greek both numbers and many different cases, since Greek is highly inflectional.
- from the collocates we searched within the safe distance of 5 left and 5 right collocates (Sinclair 1991:170, Vintar 2001:126), despite the fact that we did not expect to find any multi-word terms consisting of more than three single-word terms.
- to create our list of candidate terms we decided to exclude:
  – single letters, like *N, S, R, D*, etc. which appeared in the keyword lists as a result of the program's tendency to include all tokens in a text.
  – function words but also, in general, articles, verbs, adverbs, past particles and even adjectives, because there were more chances these to collocate with a noun and contribute to the formation of a multi-word term, to being a term themselves. Therefore, the words we looked for were only nouns.
  – proper names, place names and animals, like *Lucy, Tehuacan, Rover, Spirit, Apin, Mul, dogs, monkeys* etc. To this category, we added people's professions, like *scientists, doctors*, or capacities like *patients, programmers*, etc.

- acronyms already standardized and existing in dictionaries or in specialized glossaries have been first matched with their Greek equivalent, in cases where there was one, and then they have been included in the final candidate terms lists.
- the words that were not obviously nouns in the keyword lists, but actually adjectives or determiners in formed clusters, will not be mentioned in the single-word candidate terms lists. However if they form significant collocations with other nouns, which lead to the creation of multi-word terms, they will appear in the multi-word candidate terms lists, along with their fixed collocates.
- all collocations of a term, in the keyword list, in patterns: noun+noun, adjective+noun and noun+gerund, that appear at least twice have been examined; then their number of occurrences has been divided by the frequency of this specific term in the keyword list. All cases with a probability rate above 10% have been considered as multi-word terms and have been included in the multi-word term list.
- we will not comment on all candidate terms we are planning to include in our lists of terms. In other words the appearance in the final terms lists of terms which have not been analyzed in the present study is not an arbitrary decision, but one should keep in mind that the methodology developed here –but only partially exhibited– is going to be applied to a full extent to all candidate terms which originally appeared in the keyword frequency lists.
- our initial intention was to verify both English and Greek candidate terms in the dictionaries. However, this sometimes has not been made possible due to the lack of a reliable Greek monolingual dictionary. Consequently, and given the fact that we had five dictionaries at our disposal (two English monolingual and three English-Greek,

Greek-English bilingual), we made use of all these and also Internet, which offered us free access to specialized glossaries to verify our options.

4. ANALYSIS

4.1 Biology–Anthropology Sub-corpus

4.1.1 Overview

The Biology–Anthropology sub-corpus consists of 6 English articles (16642 words) and their translations into Greek (19003 words), which make them 12 in total (35645 words). We are aware that the topic of the articles belonging to the same sub-corpus may vary to an extent, but as was explained in the methodology section, the merging of some sub-corpora was made for reasons of economy, since we could not have corpora consisting of one or maximum two articles each. Therefore, it was decided that articles of similar topic should appear in the same sub-corpus, taking the risk of getting, in the end of the computer analysis, results that would not be very close to each other. However, this decision had to be made, like many others, in order to move on with our research.

The title Biology–Anthropology of this sub-corpus may not reflect the majority of the magazine sections in which these articles appear; however it was preferred, because it represents adequately the field of the majority of the articles included.

Taking all these factors into consideration, it is expected that the degree of technicality may vary among texts, and consequently, this may lead to an uneven distribution of technical terms within the sub-corpus. Therefore, our initial intention was to check each article individually from the Biology-Anthropology sub-corpus, as well as the entire Biology-Anthropology sub-corpus in both languages to see if we could retrieve from them potential terms.

4.1.2 Quantitative Analysis

Let us now begin with the analysis of the entire Biology-Anthropology sub-corpus and see what the keyword lists tell us about it and how we could fish out of it candidate terms. For the English sub-corpus we get a keyword list of 34 words, whereas our list for the Greek sub-corpus does not exceed the 24 words, and if we exclude the stop list that the program itself makes, then our list is limited to 31 and 22 words respectively.

The two language lists are a bit different. Below, we shall see why and to what extent these are different.

Table 1 Biology-Anthropology Sub-corpus English Keyword List

N	Word	Freq.	Biolen.lst %	Freq.	Sciamen.lst %	Keyness	P
1	Pseudogenes	65	0,39	65	0,05	123,7	0,000000
2	RNA	52	0,31	52	0,04	98,9	0,000000
3	Protein	43	0,26	57	0,04	67,6	0,000000
4	Genes	54	0,32	95	0,07	66,7	0,000000
5	Gene	40	0,24	57	0,04	59,4	0,000000
6	Water	64	0,38	149	0,11	58,3	0,000000
7	Canals	30	0,18	30	0,02	57,0	0,000000
8	Molecules	36	0,21	50	0,04	54,6	0,000000
9	m RNA	28	0,17	28	0,02	53,2	0,000000
10	Bones	23	0,14	24	0,02	42,6	0,000000
11	Afarensis	21	0,13	21	0,02	39,9	0,000000
12	DNA	23	0,14	27	0,02	39,4	0,000000
13	Molecule	25	0,15	33	0,02	39,4	0,000000
14	Genome	23	0,14	29	0,02	37,5	0,000000
15	Canal	20	0,12	21	0,02	36,9	0,000000
16	Sequences	22	0,13	28	0,02	35,6	0,000000
17	Sequence	26	0,16	42	0,03	34,7	0,000000
18	Riboswitches	18	0,11	18	0,01	34,2	0,000000
19	Site	24	0,14	37	0,03	33,4	0,000000
20	Plastic	17	0,10	18	0,01	31,2	0,000000
21	Riboswitch	16	0,10	16	0,01	30,4	0,000000
22	Functional	18	0,11	22	0,02	30,0	0,000000
23	Enzymes	17	0,10	20	0,01	29,1	0,000000
24	MIPS	15	0,09	15	0,01	28,5	0,000000
25	Aptamer	15	0,09	15	0,01	28,5	0,000000
26	Pseudogene	15	0,09	15	0,01	28,5	0,000000
27	Soil	18	0,11	24	0,02	28,1	0,000000
28	Genomes	15	0,09	16	0,01	27,4	0,000000

29	We	98	0,58	433	0,31	27,1	0,000000
30	Preserved	15	0,09	17	0,01	26,3	0,000000
31	Imprints	13	0,08	13		24,7	0,000001
32	Carbon	3	0,02	180	0,13	24,0	0,000001
33	Gas	5	0,03	227	0,16	26,1	0,000000
34	Will	14	0,08	386	0,28	29,5	0,000000

Table 2 Biology-Anthropology Greek Sub-corpus Keyword List

N	Word	Freq.	Biolgr.Lst %	Freq.	Sciamgr.Lst %	Keyness	P
1	RNA	57	0,30	57	0,04	107,7	0,000000
2	Ψευδογονίδια	36	0,19	36	0,02	68,0	0,000000
3	mRNA	29	0,15	29	0,02	54,8	0,000000
4	Μόρια	33	0,17	44	0,03	51,2	0,000000
5	Ψευδογονιδίων	25	0,13	25	0,02	47,2	0,000000
6	MIP	21	0,11	21	0,01	39,7	0,000000
7	Afarensis	21	0,11	21	0,01	39,7	0,000000
8	Γονιδίων	28	0,15	42	0,03	39,5	0,000000
9	Νερό	37	0,19	78	0,05	37,5	0,000000
10	Μόριο	24	0,13	32	0,02	37,2	0,000000
11	Οστά	21	0,11	25	0,02	35,4	0,000000
12	Γονιδίου	22	0,12	28	0,02	35,3	0,000000
13	Αλληλουχίες	19	0,10	21	0,01	33,7	0,000000
14	Αλληλουχία	20	0,11	25	0,02	32,6	0,000000
15	Γονίδια	30	0,16	60	0,04	32,2	0,000000
16	Τα	363	1,91	2.124	1,37	32,2	0,000000
17	Ριβοδιακόπτες	17	0,09	17	0,01	32,1	0,000000
18	DNA	19	0,10	23	0,01	31,6	0,000000
19	Καναλιών	18	0,09	21	0,01	30,8	0,000000
20	Α	22	0,12	36	0,02	28,8	0,000000
21	Νερού	27	0,14	57	0,04	27,3	0,000000
22	Θέση	31	0,16	74	0,05	27,0	0,000000
23	Θα	96	0,50	1.317	0,85	28,3	0,000000
24	Οι	168	0,88	2.133	1,37	35,0	0,000000

At first sight, we notice the problem that the lack of lemmatization causes, in the English but mainly in the Greek list. The word *pseudogene* appears both in plural and singular form at the top and the bottom of the list respectively; whereas its Greek equivalent *ψευδογονίδιο*, appears in the Greek list only in plural form in both the nominative/accusative (*ψευδογονίδια*)

and the genitive case (*ψευδογονιδίων*). The same thing happens to the nouns *gene* (5th), *molecule* (13th), *genome* (14th), *canal* (15th), *sequence* (17th) *and riboswitch* (21st). In the Greek list, however, the things seem more complicated as the equivalents appear in the following positions: *μόρια* (equivalent for *molecules*; plural; nominative/accusative) 4th, *μόριο* (equivalent for *molecule*; singular; nominative/accusative) 10th, *γονιδίων* (equivalent for *genes*; plural; genitive) 8th, *γονιδίου* (equivalent for *gene*; singular; genitive) 12th, *γονίδια* (equivalent for *genes*; plural; nominative/accusative) 15th, *αλληλουχίες* (equivalent for *sequences*; plural; nominative/accusative) 13th, *αλληλουχία* (equivalent for *sequence*; singular; nominative/accusative) 14th.

As can be seen from above, there is a difference in the position of words in the two lists but that is something that will be discussed in the qualitative analysis, after looking more closely at the collocations.

Another issue that arises here is the lack of matching of some English words with a Greek equivalent word. One reason for that is the highly inflectional nature of Greek. That is to say, the Greek list is so full with multiple word forms of the most frequent words that are unable to fit in all those which appear in the English list. Nonetheless, there might be exactly the opposite reason. The non inflectional nature of English makes impossible the matching with Greek, because the limited forms that an English word can take, are subdivided among the multiple forms that a Greek word can take.

To check the above assumption, we used Multiconc parallel concordances. From the English keyword list, we took four words which appeared not to have an equivalent in the Greek list. These were: *protein, genome, enzymes, aptamer*. For *protein* and *aptamer*, we saw that the case was the non-inflectional nature of English, i.e. the single form of these words in English,

and their appearance in the keyword list, do not show whether they are functioning as adjective or as noun. In Greek, however, adjectives and nouns require different word-forms.

Table 3 Multiconc Parallel Concordances

H:\multconc\biol.en P12 S1 Bacteria typically employ a number of **proteins** that constantly check the current stocks of various raw materials and adjust the number of transporters and enzymes allocated to different production lines. H:\multconc\biol.gr P12 Τα βακτήρια τυπικά διαθέτουν και χρησιμοποιούν έναν αριθμό **πρωτεϊνών** των οποίων ο ρόλος συνίσταται στα εξής: παρακολουθούν σταθερά τα τρεχούμενα αποθέματα των πρώτων υλών και ρυθμίζουν σε αριθμητικό επίπεδο τους μεταφορείς και τα ένζυμα που διατίθενται στις διαφορετικές γραμμές παραγωγής.

H:\multconc\biol.en P13 S2 In the soil bacterium Bacillus subtilis, a **protein** complex with the acronym TRAP controls one operon encoding enzymes for synthesizing the amino acid tryptophan and another describing a tryptophan transporter. <s>When TRAP senses that these proteins are not needed, it wraps the leading end of their mRNA instructions tightly around itself. H:\multconc\biol.gr P13 Στο βακτήριο του εδάφους Bacillus subtilis, ένα **πρωτεϊνικό** συμπλοκο με τα αρχικά TRAP ελέγχει ένα οπερόνιο, το οποίο κωδικεύει ένζυμα για τη βιοσύνθεση του αμινοξέος τρυπτοφάνη, και ένα άλλο οπερόνιο που εμπεριέχει τις οδηγίες για την κατασκευή του μεταφορέα της τρυπτοφάνης.

H:\multconc\biol.en P25 S2 The plan was to create an **aptamer** capable of recognizing a target molecule by binding to it and to join that to a second RNA segment that could signal the event with a visible readout. <s>For the latter role, we chose the "hammerhead" ribozyme.

H:\multconc\biol.gr P25 Το σχέδιο αφορούσε αφενός τη δημιουργία ενός **απταμερούς** ικανού να αναγνωρίζει κάποιο μόριο-στόχο στο οποίο και θα προσδένεται, αφετέρου τη σύνδεση αυτού με ένα δεύτερο τεμάχιο RNA που θα μπορούσε να σηματοδοτήσει το γεγονός της πρόσδεσης μέσω κάποιας ορατής ένδειξης.

H:\multconc\biol.en P25 S6 Once the **aptamer** end of our apparatus found and bound the target molecule, self-cleavage by the hammerhead would separate the quencher group from the fluorescent tag, and the molecule would light up as if a lampshade had been removed. H:\multconc\biol.gr P25 Από τη στιγμή που το **απταμερικό** άκρο της συσκευής μας εντόπιζε και προσέδενε το μόριο-στόχο, η αυτοδιάσπαση του σφυροκέφαλου θα διαχώριζε την αποσβένουσα ομάδα από τη φθορίζουσα ετικέτα, και το μόριο θα εξέπεμπε φως σαν να αφαιρούσαμε το αμπαζούρ μιας λάμπας.

From the examples emerges that *genome* and *enzymes* seem to be nouns. The lack of a Greek equivalent to these terms, though, seems to be a result of the various and different forms a word can take in Greek and the justifiable inability of the computer program to recognize all these as the same lemma, without the aid of a designated lemmatizer.

4.1.3 Qualitative Analysis

At this point we will analyze what, in the previous stage of the quantitative analysis seemed problematic and remained unexplained and then, we will try to create a first draft of our candidate terms lists by looking at their concordances.

For our analysis, we started from the top, i.e. the words of the highest keyness and, most of the times, also the highest frequency, and we went downwards. Therefore, for the English list of the Biology-Anthropology sub-corpus, we first checked the concordances of *pseudogenes* and *RNA*. For the former, Wordsmith did not present anything particular regarding the collocates, and for the latter, in a total of 52 instances, the collocates list showed us that *RNA* co-occurs with the noun *molecules* only 7 times.

The Greek equivalents, of the above two, adopted the same unfriendly behavior towards their neighbors. In other words, *RNA* collocated in the text only with the Greek equivalent of *RNA molecules*, i.e. with the cluster *μόρια RNA*.

By the collocations of the noun *genes*, but only in plural, we found out that sometimes (8 times out of 54) *genes* collocated with the adjective *functional.* However, interestingly, its singular form *gene* collocated with the noun *expression* giving to it a "biological" sense. For the Greek equivalent in plural, *genes*, which in the list appeared in two grammatical forms, the nominative/accusative case *γονίδια* and the genitive case *γονιδίων*, only the nominative/accusative case *γονίδια* collocated with the Greek equivalent of *functional, λειτουργικά*. Wordsmith 3.0 demonstrated that for the Greek equivalent of singular *gene* - which only appeared in the Greek list in the genitive case (*γονιδίου*) there was no collocate equivalent of *expression*.

On the next term in question, *molecules*, we will not comment, since its collocates (7 in the English and 6 in the Greek list) coincide with the previously analysed word *RNA*, hence, we assume that it is about the same case.

The word *mRNA*, which appeared on the 9th position in the English list and on the 3rd in the Greek list, collocated also with *μόρια* only half as often as *RNA* in the Greek list. To our

surprise, in the English list *mRNA* had no particular collocates. This practice could be seen as a translator's attempt to make its text more 'explicit' (Baker 1993: 243-244) and thus more easily understood to the target audience.

The next candidate term that appeared in both language lists was the word *sequence* (*αλληλουχία*) in both singular and plural form. However, neither the English nor the Greek term appeared to have any particular collocates in the same or in different cases or numbers.

Finally, the last candidate term to comment on in this sub-corpus is the word *riboswitches* and its Greek equivalent *ριβοδιακόπτες*. Although this word did not collocate with another noun, it might constitute a technical term. Finding it in the dictionary would lend support to the idea of it being a term, but its absence from the dictionary is not proof of the contrary, since dictionaries have size constraints, and are never totally up-to-date.

4.1.4 Dictionary Verification

Thus, from the checking in the dictionary, we concluded that all terms in question appeared in it; except one (*riboswitches/ριβοδιακόπτες.)* which has been found and verified on the Internet (see web resource).

4.1.5 Summary and Remarks

After having completed the threefold analysis of the Biology-Anthropology sub-corpus, we would like to comment on two main things:

The Biology-Anthropology sub-corpus, despite its quite patched nature, constitutes one of the most technical sub-corpora included in this study, as it is demonstrated in the final technical terms list.

The existence of all candidate terms in the Chambers Dictionary of Science and Technology (1999) is another index of their high level of technicality that reinforces our argument on the suggested terms.

Table 4 Single-word candidate terms

1. pseudogenes	**1. RNA**
2. RNA	**2. ψευδογονίδια**
3. protein	**3. m RNA**
4. genes	**4. μόρια**
5. gene	**5. ψευδογονιδίων**
6. water	**6. MIP**
7. canals	**7. γονιδίων**
8. molecules	**8. νερό**
9. m RNA	**9. μόριο**
10. bones	**10. οστά**
11. DNA	**11. γονιδίου**
12. molecule	**12. αλληλουχίες**
13. genome	**13. αλληλουχία**
14. canal	**14. γονίδια**
15. sequences	**15. ριβοδιακόπτες**
16. sequence	**16. DNA**
17. riboswitches	**17. καναλιών**
18. site	**18. νερού**
19. riboswitch	**19. θέση**
20. enzymes	
21. MIPS	
22. aptamer	
23. pseudogene	
24. soil	
25. genomes	
26. imprints	

Table 5 Multiword candidate terms

1. RNA molecules	**1. μόρια RNA**
2. m RNA transcript	**2. μετάγραφο m RNA**
3. protein machinery	**3. μόρια m RNA**
4. small molecules	**4. μόριο-στόχος**
5. ribosome binding	**5. οικογένειες γονιδίων**
6. gene expression	**6. οδηγο-αλληλουχία**
7. aptamer domain	**7. αλληλουχία DNA**
8. aptamer structure	**8. αλληλουχία RNA**
9. irrigation water	**9. αλληλουχία m RNA**
10. terraced irrigation	**10. λειτουργικά γονίδια**
11. pseudogene copies	**11. γνήσια γονίδια**
12. functional genes	**12. γονιδιωματικό DNA**
13. plastic imprints	
14. human genome	
15. mouse genome	
16. DNA sequences	

4.2 Energy–Environment–Geology Sub-corpus

4.2.1 Overview

This sub-corpus constitutes the biggest that exists in our parallel corpus. It contains 12 articles and their translations and its size in words is approximately 34976 words (the English version) and 40646 words (the Greek version).

The three words of the title cover almost the whole spectrum of the articles included in the sub-corpus. However, what one could notice is that they are thematically more closely related than the Biology–Anthropology articles.

4.2.2 Quantitative Analysis

The first lists we took from Wordsmith for the Energy–Environment–Geology sub-corpus contain 42 and 34 words respectively. As previously mentioned, these lists contain all parts-of-speech, including also acronyms and single letters.

Table 6 Energy-Environment-Climate Sub-corpus English Keyword List

N	Word	Freq.	Energen.lst %	Freq.	Sciamen.lst %	Keyness	P
1	Carbon	165	0,47	180	0,13	128,8	0,000000
2	Fuel	174	0,49	222	0,16	111,1	0,000000
3	Emissions	125	0,35	130	0,09	103,1	0,000000
4	Power	148	0,42	189	0,14	94,3	0,000000
5	Ethanol	102	0,29	102	0,07	87,8	0,000000
6	Plants	109	0,31	117	0,08	86,7	0,000000
7	Energy	245	0,69	455	0,33	79,9	0,000000
8	Hydrogen	121	0,34	162	0,12	72,0	0,000000
9	Nuclear	88	0,25	104	0,08	62,2	0,000000
10	Plant	68	0,19	69	0,05	57,6	0,000000
11	Methane	65	0,18	65	0,05	55,9	0,000000
12	Coal	65	0,18	65	0,05	55,9	0,000000
13	Oil	67	0,19	70	0,05	54,9	0,000000
14	Global	74	0,21	86	0,06	53,4	0,000000
15	Electricity	64	0,18	67	0,05	52,3	0,000000
16	S	89	0,25	122	0,09	51,1	0,000000
17	Per	68	0,19	78	0,06	49,9	0,000000
18	Gasoline	59	0,17	60	0,04	49,8	0,000000

19	Percent	89	0,25	128	0,09	47,4	0,000000
20	Greenhouse	50	0,14	50	0,04	43,0	0,000000
21	Vehicles	51	0,14	55	0,04	40,3	0,000000
22	Cost	57	0,16	71	0,05	37,5	0,000000
23	U	75	0,21	113	0,08	36,9	0,000000
24	Countries	44	0,12	48	0,03	34,3	0,000000
25	Wind	48	0,14	57	0,04	33,7	0,000000
26	Fuels	45	0,13	51	0,04	33,5	0,000000
27	Dioxide	44	0,12	49	0,04	33,5	0,000000
28	Efficiency	43	0,12	47	0,03	33,4	0,000000
29	Production	47	0,13	58	0,04	31,4	0,000000
30	Sea	37	0,10	40	0,03	29,1	0,000000
31	Climate	40	0,11	47	0,03	28,5	0,000000
32	Waste	36	0,10	39	0,03	28,3	0,000000
33	Will	165	0,47	386	0,28	27,8	0,000000
34	Warming	32	0,09	32	0,02	27,5	0,000000
35	Renewable	32	0,09	32	0,02	27,5	0,000000
36	Cars	34	0,10	36	0,03	27,4	0,000000
37	Oxygen	41	0,12	51	0,04	27,1	0,000000
38	Mw	31	0,09	31	0,02	26,7	0,000000
39	Corn	33	0,09	35	0,03	26,6	0,000000
40	Year	58	0,16	92	0,07	26,1	0,000000
41	Gas	108	0,30	227	0,16	25,6	0,000000
42	Atmosphere	37	0,10	46	0,03	24,4	0,000001
43	Light	19	0,05	224	0,16	29,5	0,000000

Table 7 Energy-Environment-Climate Sub-corpus Greek Keyword List

N	Word	Freq.	Energgr.Lst %	Freq.	Sciamgr.Lst %	Keyness	P
1	Άνθρακα	163	0,40	182	0,12	120,9	0,000000
2	Ενέργειας	171	0,42	235	0,15	95,2	0,000000
3	Ηλεκτροπαραγωγή	104	0,26	104	0,07	87,6	0,000000
4	Θα	530	1,31	1.317	0,85	67,0	0,000000
5	Καυσίμου	90	0,22	109	0,07	60,1	0,000000
6	Εκπομπές	74	0,18	82	0,05	55,3	0,000000
7	Εκπομπών	59	0,15	61	0,04	47,9	0,000000
8	Αιθανόλη	56	0,14	56	0,04	47,2	0,000000
9	Θερμοκηπίου	54	0,13	54	0,03	45,5	0,000000
10	Κυψέλες	54	0,13	56	0,04	43,7	0,000000
11	Μεθανίου	50	0,12	50	0,03	42,1	0,000000
12	Πετρελαίου	51	0,13	53	0,03	41,1	0,000000
13	Μεγαβάτ	47	0,12	47	0,03	39,6	0,000000
14	Αιθανόλης	45	0,11	45	0,03	37,9	0,000000
15	ΗΠΑ	73	0,18	107	0,07	36,6	0,000000
16	Υδρογόνου	59	0,15	76	0,05	36,1	0,000000
17	Γαιάνθρακα	42	0,10	42	0,03	35,4	0,000000
18	Χώρες	48	0,12	54	0,03	35,2	0,000000

19	Παραγωγή	65	0,16	91	0,06	35,1	0,000000
20	Ανά	75	0,18	115	0,07	34,7	0,000000
21	Κόστος	52	0,13	63	0,04	34,7	0,000000
22	Καύσιμα	47	0,12	54	0,03	33,6	0,000000
23	Καυσίμων	42	0,10	47	0,03	31,0	0,000000
24	Υδρογόνο	56	0,14	78	0,05	30,5	0,000000
25	Οχήματα	39	0,10	43	0,03	29,3	0,000000
26	Παραγωγής	42	0,10	50	0,03	28,7	0,000000
27	Οχημάτων	35	0,09	36	0,02	28,6	0,000000
28	Καύσιμο	52	0,13	75	0,05	26,7	0,000000
29	Παγκόσμια	38	0,09	46	0,03	25,4	0,000000
30	Βενζίνη	30	0,07	30	0,02	25,3	0,000001
31	Φυτά	36	0,09	42	0,03	25,2	0,000001
32	Σταθμούς	33	0,08	36	0,02	25,1	0,000001
33	Ατμόσφαιρα	40	0,10	51	0,03	24,9	0,000001
34	Αερίων	32	0,08	35	0,02	24,3	0,000001
35	Φωτός	3		130	0,08	41,1	0,000000
36	Κύτταρα	3		139	0,09	44,9	0,000000

At the beginning, the two lists look almost the same, and especially the top ten words, which with only a few deviations are the same in both lists. The first word, *carbon* is identical in both language lists (*άνθρακας* in Greek). Then in the second position we have the word *fuel* which does not match in position to its Greek equivalent, since this is separated in two forms -in genitive and nominative/accusative case, which appear on the 5th (*καυσίμου*) and the 28th position (*καύσιμο*) respectively. Their sum is 142 instances and we observe a significant difference between the two numbers. One of the reasons that cause this problem might be explained by translator's tendency to use singular and plural interchangeably, depending on the effect he/she wants to produce in his/her text. An attempt to interpret the data by only counting the total of the frequencies of all singular and plural forms of this specific word in both lists would not be sufficient; hence the looking at the concordances which is described in the qualitative analysis plays an equally important role.

Third in the list is *emissions.* The Greek equivalent forms that stand for *emissions,* *εκπομπές* and *εκπομπών,* are on the 6th and 7th position respectively and their sum (74+59=133) is not that far from the English number of occurrences.

The word *power*, which appears to be 4th in the list, along with the words *electricity* (15th), and the Greek *ενέργειας* (3rd) and *ηλεκτροπαραγωγή* (4th) became a major problem for the matching, the quantitative and the qualitative analysis. In reality, there is a connection among these four words on the level of correspondences, in the way that the last two constitute translations of the first two, and we suspect that this is closely related to the translator's strategies, given the purpose of the translation and the limitations the translator had to face. Nonetheless, during the quantitative analysis this was unclear, since at this stage we are restricted to the interpretation of such tendencies into numbers.

So, the word *power* appears 148 times in the whole sub-corpus, whereas in the Greek list no word is equal, or maybe close to this number; and its potential equivalent *ηλεκτροπαραγωγή* appears one position above it and differs from it by 23 instances -a difference that has not been considered significant in a previous example. However, this is not the same case because there are no other word forms of the same lemma to be added to and so far we are not sure if this specific word functions as a noun or as a determiner.

The problem with the Greek *ενέργεια* can only be explained if we look it comparatively with its presumable English equivalent *energy*. The frequency of the first outnumbers the frequency of the second and no other forms of these words appear in the lists. Here, we would like also to refer to *electricity* for which we get no single-word equivalent for Greek in the list. However, we know that this is not true because the translator has found other ways to render this word into Greek.

To test some of our assumptions, we made use of Multiconc parallel concordances and that is what we got from them:

Table 8 Multiconc Parallel Concordances

H:\multconc\energy.en P15 S1 BASED ON PREVIOUS EXPERIENCE, **electricity** from new **nuclear power** plants is currently more expensive than that from new coal- or gas-powered plants.
H:\multconc\energy.gr P15 ΜΕ ΒΑΣΗ ΤΗΝ ΠΡΟΗΓΟΥΜΕΝΗ ΕΜΠΕΙΡΙΑ, η **ηλεκτρική ενέργεια** από νέους πυρηνικούς σταθμούς **ηλεκτροπαραγωγής** κοστίζει επί του παρόντος ακριβότερα από εκείνη η οποία παράγεται σε νέους σταθμούς που καίνε γαιάνθρακα ή φυσικό αέριο.
H:\multconc\energy.en P118 S1 At a 2004 workshop, experts sketched out designs for a "SuperGrid" that would simultaneously transport **electricity** and hydrogen.
H:\multconc\energy.gr P118 Σε ένα συνέδριο το 2004, οι ειδικοί παρουσίασαν σχέδια για ένα «Υπερδίκτυο» που θα μεταφέρει ταυτόχρονα **ηλεκτρικό ρεύμα** και υδρογόνο.
H:\multconc\energy.en P162 S6 But the fertilizer, water, and natural gas and **electricity** currently expended in ethanol production from corn will need to be substantially decreased.
H:\multconc\energy.gr P162 Αλλά τα λιπάσματα, το νερό, το φυσικό αέριο και ο **ηλεκτρισμός** που τώρα χρησιμοποιούνται στην παραγωγή αιθανόλης από καλαμπόκι θα χρειαστούν να μειωθούν σημαντικά.
H:\multconc\energy.en P206 S5 The northern German state of Schleswig-Holstein currently meets one quarter of its annual **electricity** demand with more than 2,400 wind turbines, and in certain months **wind power** provides more than half the state's **electricity**.
H:\multconc\energy.gr P206 Το βόρειο γερμανικό κρατίδιο του Σλέσβιγκ Χολστάιν καλύπτει προς το παρόν το 1/4 των ετησίων αναγκών του σε **ηλεκτρική ισχύ** με περισσότερες από 2. ανεμογεννήτριες, και ορισμένους μήνες περισσότερη από τη μισή **ηλεκτρική ενέργεια** του κρατιδίου προκύπτει από την **αιολική ενέργεια**.

The above examples are representative of these cases and demonstrate roughly the problem with multiple equivalences.

4.2.3 Qualitative Analysis

The purpose of this analysis is twofold: first we tried to unfold the problem of the mismatch between English and Greek candidate terms, during the quantitative analysis, and then we committed ourselves to the process of extracting single-word as well as multi-word terms from the concordances.

Thus, *carbon* occured 165 times in the sub-corpus, of which 44 collocated with the noun *dioxide* forming the well-known chemical term, *carbon dioxide*. The same

happened in Greek with *άνθρακα* (163 instances) collocating 46 times with *διοξείδιο*, forming the term *διοξείδιο του άνθρακα*.

Another important collocate of carbon was the word *emissions*. It co-occurred with it 29 times in the English sub-corpus (*carbon emissions*) and so it did in the Greek sub-corpus, with only exception the sum of the three cases (nominative/accusative: εκπομπές 74 and genitive: εκπομπών 59) that gave approximately the same total (εκπομπές/εκπομπών άνθρακα: 32).

The next word we commented on during the quantitative analysis was *fuel*. As we said, this word appeared in the list both in singular and plural form and for the sake of the analysis we examined them both. This case turned out to be very interesting, since for different numbers we took different collocates. For instance, *fuel* in our sub-corpus collocated 28 times with *cell* and 12 times with *cells*, making a total of 40 instances out of 174, while in Greek *καυσίμου* (*fuel*) collocates with *κυψέλες* (*cells* in nominative/accusative) 29 times and with *κυψελών* (*cells* in genitive) 9 times, giving a total of 38 instances –almost the same as the English one.

Fuels (45), however, collocated with *fossil* 16 times out of 45 forming the term *fossil fuels*. In the Greek sub-corpus, the case has been the same with the plural forms *καύσιμα* (47), *καυσίμων* (42) collocating with *ορυκτά* 27 times in total, forming the collocation *ορυκτά καύσιμα/καυσίμων*.

Problems however arose in the case of four terms: *power*, *electricity*, *ενέργεια* and *ηλεκτροπαραγωγή*, which we came across at an early stage of the analysis.

Of the 148 times that appeared in the sub-corpus, in 64, the word *power* formed clusters. More specifically, the word *power* collocated with the adjective *nuclear* 38 times from which 10 times it collocated also with the noun *plants* forming the phrase

nuclear power plants. The remaining 16 clusters were also made up from adjectives such as *solar* and *electric* forming two other types of *power*, *solar power* and *electric power*.

In Greek, however, as it emerged from the parallel concordances, *nuclear power* has been translated mainly by *ηλεκτροπαραγωγή* (104) (*electricity production*) but sometimes also by the synecdoche *πυρηνική ενέργεια* (4) (*nuclear energy*) and more frequently by *ηλεκτρική ενέργεια* (28) (*electrical power*).

The equivalent of *nuclear power plants* has been as varied as its components. Some given translations were the following: *ηλεκροπαραγωγικούς σταθμούς* (*stations/plants for electricity production*), *πυρηνικούς σταθμούς ηλεκτροπαραγωγής* (*nuclear stations/plants*) and also *σταθμούς παραγωγής ισχύος* (*plants for the production of power*).

The word *energy*, as a synonym of the word *power*, which was used by authors, and apparently also translators, interchangeably, appeared in the following clusters: *renewable energy* (12 times), *energy sources* (14 times) and *energy companies* (10 times). The reason these collocations are stated here –albeit not significant in number- is because their equivalents have also been found in the Greek sub-corpus: *ανανεώσιμης ενέργειας* (7), *πηγές/πηγών ενέργειας* (19), *εταιρείες/επιχειρίσεις ενέργειας* (10).

Finally, the word *electricity* although it did not give statistically significant collocations in English, in Greek can be matched to both single and multi-word terms. In particular, we see that the most frequent equivalent was *ηλεκτρική ενέργεια* (28 instances) but the rest of the cases have been complemented by *ηλεκτροπαραγωγή*, *ηλεκτρισμός, ενέργεια*, *ηλεκτρικό ρεύμα* and *ηλεκτρική ισχύς*.

4.2.4 Dictionary Verification

The looking up of our findings in the dictionary sometimes verified our findings and sometimes not. The encouraging thing is that it also confirmed the existence of some multi-word terms we came across in our corpus, giving us confidence in the method of the extraction of terms we follow in this study.

4.2.5. Summary-Remarks

The analysis of this sub-corpus revealed many interesting facts, among them the issue of multiple equivalents of a term, which is dependent on translator's fluency as well as on external and internal factors of the translation process.

In this sub-corpus, we came across multi-word terms that we did not have the chance to examine extensively in Biology–Anthropology sub-corpus. The existence of these terms in corpus-based technical dictionaries offered us a kind of evidence that our study, albeit in a small scale, is moving to the right direction.

Table 9 Single-word candidate terms

1. carbon	1. άνθρακα
2. fuel	2. ενέργειας
3. emissions	3. ηλεκτροπαραγή
4. power	4. καυσίμου
5. ethanol	5. εκπομπές
6. plants	6. εκπομπών
7. energy	7. αιθανόλη
8. hydrogen	8. θερμοκηπίου
9. plant	9. κυψέλες
10. methane	10. μεθανίου
11. coal	11. πετρελαίου
12. oil	12. μεγαβάτ
13. electricity	13. αιθανόλης
14. gasoline	14. υδρογόνου
15. greenhouse	15. γαιάνθρακα
16. vehicles	16. χώρες
17. cost	17. παραγωγή
18. countries	18. κόστος
19. wind	19. κάυσιμα
20. fuels	20. καυσίμων
21. dioxide	21. υδρογόνο
22. efficiency	22. οχήματα
23. production	23. παραγωγής
24. sea	24. οχημάτων
25. climate	25. καύσιμο
26. waste	26. παγκόσμια
27. warming	27. βενζίνη
28. cars	28. φυτά
29. oxygen	29. σταθμούς
30. MW	30. ατμόσφαιρα
31. corn	31. αερίων
32. year	
33. gas	
34. atmosphere	

Table 10 Multi-word candidate terms

1. carbon dioxide	**1. διοξείδιο του άνθρακα**
2. carbon emissions	**2. εκπομπές άνθρακα**
3. fuel-cells	**3. κυψέλες καυσίμου**
4. gas emissions	**4. κάυσιμο υδρογόνο**
5. nuclear power	**5. εκπομπές αερίων**
6. power plants	**6. ηλεκτρική ενέργεια**
7. climate change	**7. πηγές ενέργειας**
8. refueling stations	**8. παραγωγή υδρογόνου**
9. hydrogen stations	**9. εκπομπές μεθανίου**
10. cellulose ethanol	**10. καύση γαιάνθρακα**
11. mass extinctions	**11. αέρια θερμοκηπίου**
12. coal-fired electricity	**12. εκπομπή αερίων θερμοκηπίου**
13. coal-fired power	**13. οχήματα με κυψέλες καυσίμου**
14. coal powered plant	**14. παραγωγή ενέργειας**
15. coal plants	**15. ορυκτά καύσιμα**
16. greenhouse gas emissions	**16. παραγωγή αιθανόλης**
17. greenhouse gas	**17. παγκόσμια θέρμανση**
18. fuel-cell vehicles	
19. solar photovoltaics	
20. efficiency measures	
21. waste management	
22. nuclear waste	
23. fossil fuels	
24. energy efficiency	
25. global warming	

4.3 Medicine Sub-corpus

4.3.1 Overview

The Medicine sub-corpus is the smallest of our parallel sub-corpora, 13.658 words the English version and 15.110 words the Greek version. Like the Physics sub-corpus, it contains only four articles and their translations; and two of them treat the issue of cancer. Hence it is expected that they will be thematically related with each other, but this is something we will examine thoroughly below.

4.3.2 Quantitative Analysis

The English keyword list that Wordsmith 3.0 gives us contains 44 words of which one is single letter, whereas the Greek one contains 40, of which four are single letters and are not going to be part of the final terms list, at least not as single letters. Let us now examine how close are the two language lists with regards to the candidate terms that appear at the top of them.

Table 11 Medicine Sub-corpus English Keyword List

N	Word	Freq.	Medicen.Lst %	Freq.	Sciamen.Lst %	Keyness	P
1	Cancer	179	1,32	180	0,13	400,7	0,000000
2	Cells	142	1,04	234	0,17	230,9	0,000000
3	T	99	0,73	116	0,08	202,8	0,000000
4	Immune	63	0,46	65	0,05	138,7	0,000000
5	Dogs	57	0,42	59	0,04	125,2	0,000000
6	Disease	56	0,41	62	0,04	118,5	0,000000
7	Regs	51	0,37	51	0,04	114,1	0,000000
8	Autoantibodies	48	0,35	48	0,03	107,4	0,000000
9	Autoimmune	34	0,25	34	0,02	76,1	0,000000
10	Pet	34	0,25	35	0,03	74,9	0,000000
11	Blood	36	0,26	41	0,03	74,9	0,000000
12	Diabetes	31	0,23	31	0,02	69,3	0,000000
13	Cancers	27	0,20	27	0,02	60,4	0,000000
14	Patients	25	0,18	28	0,02	52,5	0,000000
15	Mice	23	0,17	27	0,02	47,0	0,000000
16	Tissues	21	0,15	22	0,02	45,8	0,000000
17	People	44	0,32	117	0,08	45,3	0,000000
18	Proteins	26	0,19	39	0,03	45,2	0,000000
19	Predictive	19	0,14	19	0,01	42,5	0,000000
20	Selection	22	0,16	29	0,02	41,8	0,000000
21	Against	27	0,20	48	0,03	41,2	0,000000
22	Tumor	18	0,13	18	0,01	40,3	0,000000
23	Trials	19	0,14	21	0,02	40,2	0,000000

24	Genes	36	0,26	95	0,07	37,4	0,000000
25	Drugs	17	0,12	18	0,01	36,8	0,000000
26	Evolutionary	19	0,14	29	0,02	32,6	0,000000
27	Tumors	15	0,11	16	0,01	32,4	0,000000
28	Comparative	14	0,10	14	0,01	31,3	0,000000
29	Risk	18	0,13	27	0,02	31,3	0,000000
30	Type	25	0,18	61	0,04	28,3	0,000000
31	Insulin	13	0,10	14	0,01	27,9	0,000000
32	Investigators	19	0,14	36	0,03	27,5	0,000000
33	Oncologists	12	0,09	12	26,8		0,000000
34	Cell	30	0,22	91	0,07	26,2	0,000000
35	Diseases	14	0,10	19	0,01	26,1	0,000000
36	Human	31	0,23	97	0,07	26,0	0,000000
37	Doctors	13	0,10	16	0,01	25,8	0,000000
38	Humans	22	0,16	54	0,04	24,8	0,000001
39	Colon	11	0,08	11	24,6		0,000001
40	Defenses	11	0,08	11	24,6		0,000001
41	Autoantibody	11	0,08	11	24,6		0,000001
42	Disorders	11	0,08	11	24,6		0,000001
43	Therapy	12	0,09	14	0,01	24,6	0,000001
44	Bone	14	0,10	21	0,02	24,3	0,000001
45	The	682	5,01	9.037	6,56	52,5	0,000000
46	Energy	5	0,04	455	0,33	54,7	0,000000

Table 12 Medicine Sub-corpus Greek Keyword List

N	Word	Freq.	Medicgr.Lst %	Freq.	Sciamgr.Lst %	Keyness	P
1	Κύτταρα	104	0,69	139	0,09	198,5	0,000000
2	Τ	90	0,60	107	0,07	184,8	0,000000
3	Καρκίνου	79	0,52	79	0,05	178,8	0,000000
4	Καρκίνο	74	0,49	74	0,05	167,4	0,000000
5	Reg	53	0,35	53	0,03	119,9	0,000000
6	Κυττάρων	52	0,34	72	0,05	96,8	0,000000
7	Σκύλους	35	0,23	37	0,02	76,8	0,000000
8	Αυτοαντισωμάτων	30	0,20	30	0,02	67,8	0,000000
9	Αυτοαντισώματα	26	0,17	26	0,02	58,8	0,000000
10	Δ	24	0,16	24	0,02	54,2	0,000000
11	Σ	24	0,16	24	0,02	54,2	0,000000
12	Ανοσοποιητικού	23	0,15	23	0,01	52,0	0,000000
13	Ανθρώπους	33	0,22	63	0,04	48,1	0,000000
14	Καρκινικά	19	0,13	20	0,01	41,8	0,000000
15	Ερευνητές	46	0,30	144	0,09	39,4	0,000000
16	Σκύλοι	17	0,11	17	0,01	38,4	0,000000
17	Αίμα	19	0,13	24	0,02	37,5	0,000000
18	Ανοσοποιητικό	17	0,11	18	0,01	37,2	0,000000
19	Καρκίνος	16	0,11	16	0,01	36,2	0,000000
20	Εντέρου	16	0,11	16	0,01	36,2	0,000000

21	Στους	58	0,38	228	0,15	35,0	0,000000
22	Γονίδια	27	0,18	60	0,04	34,2	0,000000
23	Άνθρωπο	16	0,11	19	0,01	32,8	0,000000
24	Εξετάσεις	15	0,10	16	0,01	32,7	0,000000
25	Θεραπεία	16	0,11	20	0,01	31,8	0,000000
26	Φάρμακα	16	0,11	20	0,01	31,8	0,000000
27	Εναντίον	14	0,09	14		31,6	0,000000
28	Νοσοπροβλεπτικοί	14	0,09	14		31,6	0,000000
29	Κατοικίδιους	14	0,09	14		31,6	0,000000
30	Πρωτεινες	17	0,11	26	0,02	29,5	0,000000
31	Επιλογή	18	0,12	31	0,02	28,6	0,000000
32	Τύπου	27	0,18	73	0,05	27,8	0,000000
33	Σκύλων	12	0,08	12		27,1	0,000000
34	Ιστούς	12	0,08	12		27,1	0,000000
35	Κατοικίδιοι	12	0,08	12		27,1	0,000000
36	Δοκιμές	15	0,10	22	0,01	26,8	0,000000
37	Ποντικούς	12	0,08	13		26,0	0,000000
38	Εμφάνιση	16	0,11	27	0,02	25,8	0,000000
39	Φυσική	17	0,11	32	0,02	25,1	0,000001
40	P	12	0,08	14		24,9	0,000001
41	Ενέργειας	3	0,02	235	0,15	26,0	0,000000
42	Το	210	1,39	3.518	2,26	55,8	0,000000

The first word in the English list, which was sorted according to keyness, is *cancer*, while in the Greek list the first word to appear at the top of the list is *κύτταρα*, the Greek equivalent for *cells*, which is on the 2nd place of the English list. That is because the Greek equivalent of *cancer* appears in many forms: in different cases but also –as the parallel concordances reveal– in different parts of speech (as noun and as adjective).

The single letter *T* is on the third position of the English list. The high frequency of *T* in both lists, along with the fact that we are examining a medicine sub-corpus is sufficient reason for us to treat it as a candidate term and include it in our analysis.

Almost the same case we come across with the word that appears on the 4th position of the English list. For *immune* and its translation *ανοσοποιητικού* we assume that their presence there implies the existence of *system* –or *συστήματος* for Greek– as a very frequent collocate. Yet, contrary to what one would expect neither *system* nor *συστήματος* are there in the lists, but they do form clusters with *immune* and *ανοσοποιητικού*, as we will find out in the qualitative analysis.

An interesting case that is worth looking at although it is not among the first 10 words of the keyword list is that of the word *tumor* and its plural form *tumors*. Actually, no translation of this term exists in the Greek list. In fact, as was suspected the expected equivalent was not used enough times to appear in the list. Checking the parallel concordances, we noticed that there was a translation equivalent for the term, but in some cases *tumor* and *tumors* have been translated by the Greek equivalent for *cancer* both as noun (*καρκίνος*) and as adjective (*καρκινικά*). This will be further examined during the qualitative analysis.

Table 13 Multiconc Parallel Concordances

H:\multconc\med.en P35 S5 Some findings suggest, for example, that cancer patients have abnormally high numbers of active **T-regs** both in their blood and in the tumors themselves.
H:\multconc\med.gr P35 Μερικά ευρήματα δείχνουν, για παράδειγμα, πως οι καρκινοπαθείς έχουν αφύσικα υψηλούς αριθμούς ενεργών **T-reg** τόσο στο αίμα τους όσο και στον ίδιο τον όγκο.
H:\multconc\med.en P5 S5 they also influence the **immune system's** responses to infectious agents, cancer, organ transplants and pregnancy.
H:\multconc\med.gr P5 Επηρεάζουν επίσης τις αποκρίσεις του **ανοσοποιητικού συστήματος** στους λοιμογόνους παράγοντες, στον καρκίνο, στη μεταμόσχευση οργάνων και στην εγκυμοσύνη
H:\multconc\med.en P19 S2 The **cells** appear capable of suppressing a wide variety of **immune system cells**, impeding the cells' multiplication and also their other activities, such as secretion of cell-to-cell chemical signals (cytokines). H:\multconc\med.gr P19 Τα **κύτταρα** φαίνονται ικανά να καταστείλουν ένα ευρύ φάσμα **ανοσοκυττάρων**, παρεμποδίζοντας τον πολλαπλασιασμό τους, καθώς και άλλες δραστηριότητες τους, όπως η έκκριση χημικών σημάτων κατά την επαφή των κυττάρων μεταξύ τους (κυτοκίνες).
H:\multconc\med.en P87 S1 Imagine a 60-year-old man recuperating at home after **prostate cancer** surgery, drawing comfort from the aged golden retriever beside him.
H:\multconc\med.gr P87 Φανταστείτε έναν εξηντάχρονο άνδρα που αναρρώνει στο σπίτι έπειτα από εγχείρηση **καρκίνου του προστάτη**, βρίσκοντας παρηγοριά στο ηλικιωμένο γκόλντεν ρετρίβερ που κάθεται δίπλα του.
H:\multconc\med.en P88 S2 Despite an unprecedented surge in researchers' understanding of what **cancer cells** can do, the translation of this knowledge into saving lives has been unacceptably slow.
H:\multconc\med.gr P88 Παρά την πρωτοφανή συσσώρευση γνώσεων από τους ερευνητές σχετικά με το τι μπορούν να κάνουν τα **καρκινικά κύτταρα**, η αξιοποίηση αυτής της γνώσης για τη διάσωση ζωών είναι απαράδεκτα αργή.

4.3.3 Qualitative Analysis

Let us start with the most frequent term *cancer*, which more often appears as a single-word term, but when it appears as a cluster, it collocates with the word *cells* forming the collocation *cancer cells* (16 out of 179) and behaving as an adjective. In its other collocations, we could say that it forms multi-word terms because as a cluster it signifies types of *cancer*, such as *bone cancer* (7)*, colon cancer* (7)*, breast cancer* (5)

and *lung cancer* (4). Only the equivalent for *colon cancer* occurs in the Greek keyword list; hence only this equivalent will be included in the final list of terms.

Additionally, the word *cells* appears also in interesting clusters, creating fixed multi-word terms, like *T cells* (28), *beta cells* (8), *white blood cells* (3) and also *cancer cells* (17), *tumor cells* (5), *immune system cells* (5) and others. In Greek, the word *κύτταρα* appears in equivalent clusters: *καρκινικά κύτταρα* (17), *κύτταρα του ανοσοποιητικού συστήματος* (4), *κύτταρα Τ* (18), *κύτταρα β* (6), *καρκινικά κύτταρα* (17).

The letter *T* never appears in the text, but in clusters, which is quite normal; thus we find it either in collocations such as *T-regs* (61/99), in *T-cells* (36/99) or in *T lymphocytes* (2/99). To a great extent, the same occurs with the letter *T* in the Greek keyword list: *T-reg* (52/90) (non-translated in Greek because it is a standardized term), *κύτταρα/κυττάρων Τ* (33/90), *λεμφοκύτταρα Τ* (2/90).

In the case of the adjective *immune* (*ανοσοποιητικού*) which we discussed in the quantitative analysis, we see that our initial assumption is actually verified by the concordances. In 38 cases out of 63 the *immune* collocates with the *system*, and in the 6 of them *immune* collocates both with *system* and with *cells* forming the name of a specific type of *cells*, the *immune system cells*. In the Greek sub-corpus, the translation equivalent for *immune, ανοσοποιητικού* behaves in the same way: *ανοσοποιητικού* (genitive case) / *ανοσοποιητικό* (nominative case) collocate 100% (40/40) with the noun *συστήματος/σύστημα*, and more specifically in 4 of these cases *ανοσοποιητικού* and *συστήματος* collocate with the Greek equivalent of *cells*, forming the cluster, *κύτταρα του ανοσοποιητικού συστήματος*.

Finally, we will check the problematic case of *tumor/tumors,* which caused us problems in the quantitative analysis.

First of all, we assumed that there is a difference between the use of singular and the use of plural form and we were right about that because most of plural forms were nouns in the texts, whereas the majority of singular forms were adjectives. In particular, we saw that the plural form *tumors* is mostly used to denote “the abnormal mass of new tissue growing in or on part of the body” (Oxford Advanced Learner’s Dictionary 1995:1283) and it is translated in Greek as such with the noun *όγκοι* (10/15). The singular *tumor*, on the other side, appears in the following clusters: *tumor cells* (5/18), *tumor suppressor proteins* (2/19) which are translated in Greek

also by similar clusters: *καρκινικά κύτταρα* and *ογκοκατασταλτικές πρωτείνες*. In three of the cases –two plural and one singular- this word has been translated with the Greek equivalent of *cancer*, *καρκίνος/καρκίνοι*.

4.3.4 Dictionary Verification

The stage of dictionary verification turned out to be a challenging process in this sub-corpus. The reason for that may be attributed to its special nature as well as to the high level of technicality of the extracted terms.

Hence, two technical dictionaries have been used: the Chambers Dictionary of Science and Technology (1999) and the bilingual Dorland's Medical English-Greek, Greek-English Dictionary (1989); but they have not covered the whole range of our terms. Therefore, some terms, like *cancer cells, tumor cells, colon cancer* and *T-reg cells*, have been looked up on the internet, in reliable and trustworthy sources (see website).

4.3.5 Summary-Remarks

Overall, in this sub-corpus we observed a considerable lack of terms in the Greek list compared to the English one. That is linked once again to the translator's choices regarding the translation process, as well as its final product. Although the examination of translation strategies would be interesting in this study; nevertheless it is beyond its scope, thus we will not examine them any further.

Another problem that emerged from the analysis of this sub-corpus, during the stage of dictionary verification, is the difficulty that technical dictionaries face in setting the limits of their width and in providing adequate explanations to the experts who are their most frequent users. Fortunately, we have the Internet, which, despite its drawbacks, can be a valuable source of information, when it is wisely used.

Table 14 Single-word candidate terms

1. cancer	**1. κύτταρα**
2. cells	**2. καρκίνου**
3. disease	**3. καρκίνο**
4. autoantibodies	**4. κυττάρων**
5. blood	**5. αυτοαντισωμάτων**
6. diabetes	**6. αυτοαντισώματα**
7. cancers	**7. αίμα**
8. patients	**8. καρκίνος**
9. tissues	**9. εντέρου**
10. proteins	**10. γονίδια**
11. selection	**11. εξετάσεις**
12. tumor	**12. θεραπεία**
13. trials	**13. φάρμακα**
14. genes	**14. πρωτείνες**
15. drugs	**15. επιλογή**
16. tumors	**16. τύπου**
17. risk	**17. ιστούς**
18. type	**18. δοκιμές**
19. insulin	**19. εμφάνιση**
20. cell	
21. diseases	
22. colon	
23. defenses	
24. autoantibody	
25. disorders	
26. therapy	
27. bone	

Table 15 Multiword candidate terms

1. human cancers	**1. καρκίνος του παχέος εντέρου**
2. cancer cells	**2. ανοσοποιητικό σύστημα**
3. T cells	**3. φυσική επιλογή**
4. T-reg	**4. καρκινικά κύτταρα**
5. autoimmune disease	**5. κλινικές δοκιμές**
6. predictive autoantibodies	**6. δοκιμές στον άνθρωπο/στους ανθρώπους**
7. blood sample	**7. κύτταρα Τ**
8. blood vessels	
9. type 1 diabetes	
10. type 2 diabetes	
11. natural selection	
12. tumor cells	
13. tumor growth	
14. tumor suppressor	
15. clinical trials	
16. human trials	
17. prevention trials	
18. cancer genes	
19. immune system cells	
20. immune system	
21. colon cancer	

4.4 Physics Sub-corpus

4.4.1 Overview

The Physics sub-corpus is the second smaller sub-corpus of this study after Medicine sub-corpus. The size of the English sub-corpus is approximately 14,404 words and that of the Greek one 15652 words. It is composed of four articles and their translations. Two of them appear in the magazine section entitled Physics and the two others in the applied physics section. Their topics vary a lot ("THE ULTIMATE WHITE LIGHT", by Alfano; "SEEING WITH SUPERCONDUCTORS", by Irwin; "MAKING SILICON LASE", by Jalali; "WEIGHTY MATTERS" by Robinson), but we are going to see how this dissimilarity will be reflected on the lists with the candidate terms.

4.4.2 Quantitative Analysis

The English and Greek keyword lists retrieved from Wordsmith tool contain 41 and 35 terms respectively. The remarkable thing here is that unlike the corpora we examined till now, these keyword lists are very similar to each other and the matching between the terms is if not obvious, at least easy.

Consequently, the list has the form illustrated in the table below:

Table 16 Physics Sub-corpus English Keyword List

N	Word	Freq.	Physen.Lst %	Freq.	Sciamen.Lst %	Keyness	P
1	Light	141	0,98	224	0,16	223,8	0,000000
2	Silicon	72	0,50	82	0,06	143,5	0,000000
3	Frequency	57	0,40	64	0,05	114,5	0,000000
4	Laser	46	0,32	48	0,03	96,4	0,000000
5	Electrons	48	0,33	67	0,05	83,7	0,000000
6	SC	37	0,26	37	0,03	79,4	0,000000
7	Band	36	0,25	36	0,03	77,3	0,000000
8	Photon	40	0,28	49	0,04	76,1	0,000000
9	Superconducting	38	0,26	43	0,03	76,0	0,000000
10	Detectors	36	0,25	38	0,03	75,0	0,000000
11	Optical	37	0,26	41	0,03	74,9	0,000000
12	Crystal	30	0,21	33	0,02	61,0	0,000000
13	Photons	33	0,23	43	0,03	60,3	0,000000

14	Mass	54	0,38	127	0,09	59,5	0,000000
15	Electron	29	0,20	36	0,03	54,7	0,000000
16	Pulse	26	0,18	28	0,02	53,5	0,000000
17	Quantum	23	0,16	28	0,02	43,9	0,000000
18	The	1.154	8,03	9.037	6,56	42,8	0,000000
19	Detector	22	0,15	27	0,02	41,8	0,000000
20	Measurements	21	0,15	27	0,02	38,7	0,000000
21	Upper	22	0,15	31	0,02	38,1	0,000000
22	Semiconductor	18	0,13	19	0,01	37,5	0,000000
23	Constant	21	0,15	29	0,02	36,9	0,000000
24	Pulses	21	0,15	30	0,02	36,0	0,000000
25	Lasers	17	0,12	18	0,01	35,3	0,000000
26	Kilogram	22	0,15	35	0,03	34,8	0,000000
27	Medium	17	0,12	19	0,01	34,2	0,000000
28	Frequencies	18	0,13	23	0,02	33,3	0,000000
29	Energy	92	0,64	455	0,33	29,0	0,000000
30	Momentum	15	0,10	18	0,01	28,9	0,000000
31	TES	13	0,09	13		27,9	0,000000
32	Voltage	13	0,09	13		27,9	0,000000
33	Sphere	16	0,11	23	0,02	27,3	0,000000
34	Index	14	0,10	17	0,01	26,8	0,000000
35	Atoms	21	0,15	43	0,03	26,7	0,000000
36	Fiber	17	0,12	28	0,02	26,1	0,000000
37	Lasing	12	0,08	12		25,7	0,000000
38	Bands	13	0,09	15	0,01	25,7	0,000000
39	Gamma	13	0,09	16	0,01	24,6	0,000000
40	Emission	17	0,12	31	0,02	24,0	0,000000
41	Atomic	14	0,10	20	0,01	24,0	0,000000
42	May	4	0,03	236	0,17	25,0	0,000000
43	We	12	0,08	433	0,31	32,2	0,000000

Table 17 Physics Sub-corpus Greek Keyword List

N	Word	Freq.	Physgr.Lst %	Freq.	Sciamgr.Lst %	Keyness	P
1	Λέιζερ	68	0,43	70	0,05	147,4	0,000000
2	Φως	83	0,53	127	0,08	139,5	0,000000
3	Φωτός	83	0,53	130	0,08	137,1	0,000000

4	Πυριτίου	43	0,27	52	0,03	84,7	0,000000
5	ΥΣ	38	0,24	38	0,02	83,6	0,000000
6	Πυρίτιο	31	0,20	33	0,02	65,9	0,000000
7	Ηλεκτρόνια	35	0,22	47	0,03	64,4	0,000000
8	Ανιχνευτές	30	0,19	32	0,02	63,7	0,000000
9	Συχνοτήτων	27	0,17	28	0,02	58,2	0,000000
10	Φωτονίων	25	0,16	29	0,02	50,5	0,000000
11	Συχνότητας	24	0,15	27	0,02	49,4	0,000000
12	Κύματος	31	0,20	51	0,03	49,2	0,000000
13	Ζώνη	30	0,19	52	0,03	45,7	0,000000
14	Φωτόνια	21	0,13	25	0,02	41,8	0,000000
15	Συχνότητα	21	0,13	28	0,02	38,8	0,000000
16	Ακρίβεια	25	0,16	44	0,03	37,6	0,000000
17	Παλμού	16	0,10	16	0,01	35,2	0,000000
18	Φωτόνιο	20	0,13	29	0,02	34,9	0,000000
19	Μάζα	28	0,18	62	0,04	34,2	0,000000
20	Ανώτερη	17	0,11	20	0,01	34,1	0,000000
21	Ηλεκτρονίων	18	0,11	24	0,02	33,3	0,000000
22	Οποία	93	0,59	455	0,29	32,9	0,000000
23	Υλικό	20	0,13	32	0,02	32,5	0,000000
24	Διάθλασης	14	0,09	14		30,8	0,000000
25	Φωτονίου	14	0,09	14		30,8	0,000000
26	Φαινόμενο	28	0,18	71	0,05	29,6	0,000000
27	Μέτρηση	15	0,10	19	0,01	28,7	0,000000
28	Χιλιόγραμμου	13	0,08	13		28,6	0,000000
29	Ανιχνευτών	13	0,08	13		28,6	0,000000
30	Σταθεράς	13	0,08	13		28,6	0,000000
31	Μετρήσεις	16	0,10	24	0,02	27,3	0,000000
32	Δευτερόλεπτο	17	0,11	28	0,02	27,0	0,000000
33	Κρύσταλλο	12	0,08	12		26,4	0,000000
34	Ένα	137	0,87	830	0,53	25,4	0,000000
35	Ηλεκτρόνιο	13	0,08	17	0,01	24,4	0,000001
36	Ότι	39	0,25	915	0,59	36,7	0,000000

High on the list, there are words like *light, silicon, frequency, laser* and others. The most frequent terms in the Greek list include *λέιζερ* (*laser*), *φως* (*light* in nominative/accusative case), *φωτός* (*light* in genitive case), *πυριτίου* (*silicon*).

The difference in numbers can be justified by true evidence. For instance, the difference between *light* and its translation *φως/φωτός* can be settled by the sum of the two frequencies of the Greek equivalents. Their total (166) outnumbers its English original and this could be explained by Baker's universal feature of explicitation, according to which "addition of extra information, insertion of explanations, repetition of previously mentioned details are done for the purpose of clarity" (Baker 1998: 289). The same could also be claimed for all translation equivalents which occur more times than their originals; but this can be done only at this stage of analysis, since in the qualitative analysis we are interested in looking at real examples extracted from the corpus itself.

The point where the lists differentiate is after the 28th term of the English list. Equivalents for words like: *superconducting*, *optical*, *quantum*, *semiconductor, medium, momentum*, the acronym *TES*, *voltage, sphere, index, atoms, fiber, lasing, bands*, the type *gamma* and *emission,* are absent in the Greek keyword list. Yet, in the Greek list, some terms remain "unmatched" and that is something worrying. However, as we realized while checking Multiconc parallel concordances, this presumable gap in language matching can be filled. To put it more simply, we know that in language there are clusters or fixed collocations, that when we come across one of their components in a sentence, we suppose that next to it or fairly close, there will be another term with which it forms an entity. In the same way, some words –and more specifically some adjectives- such as *superconducting, optical* and the noun *index*, we are used to seeing them in fixed collocations like *superconducting material*, *optical fiber* or *refractive index* and we discovered through Multiconc parallel concordances that these collocations not only exist in Greek but also constitute the "lost piece in the puzzle" of the unmatched terms. That is to say, words like *υλικό* (*material*) and *διάθλασης* (refractive) for which no equivalent term appears in the English keyword list, were a part of the broken collocations *υπεραγώγιμο υλικό* (*superconducting material*) and *δείκτης διάθλασης* (*refractive index)*.

Table 18 Multiconc Parallel Concordances

H:\multconc\phy.en P14 S1 The amount by which the **refractive index** increases depends on the light's intensity, so as the pulse passes by a given location in the medium the refractive index there varies continuously, and so do the induced phase changes.

H:\multconc\phy.gr P14 Το ποσό κατά το οποίο αυξάνεται ο **δείκτης διάθλασης** εξαρτάται από την ένταση του φωτός- έτσι, καθώς ο παλμός διέρχεται από ένα σημείο του μέσου, εκεί ο δείκτης διάθλασης μεταβάλλεται συνεχώς, και ομοίως μεταβάλλονται οι επαγόμενες αλλαγές φάσης του παλμού.

H:\multconc\phy.en P44 S1 Tiny devices made of **superconducting material** that act as superb sensors of photons and other particles are revolutionizing a wide range of research and technology fields

H:\multconc\phy.gr P44 Μικροσκοπικές συσκευές κατασκευασμένες από **υπεραγώγιμο υλικό**, οι οποίες λειτουργούν ως εξαιρετικοί αισθητήρες φωτονίων και άλλων σωματιδίων, φέρνουν επανάσταση σε ένα ευρύ φάσμα ερευνητικών και τεχνολογικών πεδίων

4.4.3 Qualitative Analysis

Let us now look more closely and test our findings from the quantitative analysis.

The candidate term *light* that occurs totally 148 times in the Physics sub-corpus, collocates 8 times with the candidate term *laser*; 6 times with the word *visible;* and 5 times with the adjective *white* forming the following clusters: *laser light, visible light* and *white light*. As a determiner, *light* appears 4 times with *emission, pulses, source* and 3 times with the word *beam*, forming the: *light emission, light pulses, light source* and *light beam*. In the same way in the Greek sub-corpus, the equivalent for light *φως/φωτός* forms the following clusters: *φως/φωτός λέιζερ (laser light)* (15/166)*, ορατού φωτός (visible light)* (8/166)*, λευκό φως/λευκού φωτός (white light)* (6/166), *φως υψηλής/χαμηλής συχνότητας (high/low frequency light)* (3/83), *εκπομπή/εκπομπής φωτός* (*light emission)* (14/83). As we notice the times that *light* co-occurs with another term are too few to be considered.

The word *silicon* collocates with *laser(s)* 11 times out of 72 that *silicon* appears in the sub-corpus, while their Greek equivalents *πυριτίου* collocates with *λέιζερ* 10 times out of 71, showing that the collocation *silicon laser* is equal to its equivalent *λέιζερ πυριτίου* (where the equivalent for *silicon*, *πυρίτιο* is a genitive).

Another interesting case is that of the acronym *SC – ΥΣ* in Greek- which stands for Supercontinuum – Υπερσυνεχές. The concordances show that SC has a tendency to appear alone in the English sub-corpus, collocating with *light* only 7 out of 121 times. On the contrary, *ΥΣ* in the Greek sub-corpus collocates almost exclusively (34 out 38 times) with the equivalent of light forming the clusters: *ΥΣ φως* or *ΥΣ του φωτός*.

Then the word *band*, which stands for electron's energy band, co-occurs -as expected- with *upper (upper band)* (13/36), *energy (energy band)* (8/36), *lower (lower band)* (3/36). In the Greek sub-corpus, important collocations are *ανώτερη ζώνη* (13/36) and *ενεργειακή ζώνη* (3/36).

The word *detectors* almost never occurs alone in the sub-corpus but with some determiner: *superconducting detectors* (8/36) and *TES detectors* (4/40) are the most frequent. However it also co-occurs with other words forming collocations –but not significant in number- like *x-ray/gamma-ray/photon detectors*. We observe that the same thing happens also in the Greek sub-corpus with the equivalent of detectors, *ανιχνευτές* forming the following collocations: *υπεραγώγιμοι ανιχνευτές* (*superconducting detectors*) (6/30)*, ανιχνευτές TES* (*TES detectors*) (4/30), and the Greek collocations: *ανιχνευτές ακτίνων Χ, ανιχνευτές ακτίνων/ακτινοβολίας γ, ανιχνευτές φωτονίων*.

Another interesting case is that of the term *constant*. As we noticed in the concordances, the word *constant* has been used in the sub-corpus both as a noun denoting "a quantity (or parameter), which remains the same while the variables change" (Chambers Dictionary 1999:255) and as an adjective, implying something firm and stable (Oxford Dictionary 1995:246). When it appears as a noun it is almost always preceded by the name of the constant, like here *Planck's* or *Avogadro constant*. In our Physics sub-corpus is more frequently appeared as a noun: *Planck's constant* (7/21) – *σταθερά του Planck* (6/13); *Avogadro constant* (6/21) - *σταθερά του Avogadro* (6/21).

Last, we will comment on one of the cases we came across during the quantitative analysis with the unmatched terms. The term we will look at is the noun *index*, which as appears from the concordances constitutes a multi-word term with the adjective *refractive* (10/14). A look in the Greek concordances can definitely convince us that it is about a multi-word term, since the 14/14 of the two terms co-occurrence (*δείκτης διάθλασης*) does not let us any doubt about it.

4.4.4 Dictionary Verification

Once again, the dictionary verified a large number of single but not so many multi-word terms. The existence of multi-word terms like *white light, visible light, energy*

band, band gap, Planck's and Avogadro's constant and *refractive index* have been also confirmed by the dictionary.

4.4.5 Summary-Remarks

The Physics sub-corpus has been an interesting case as it contains single and multi-word terms of a high level of technicality.

A problem that arises also in this sub-corpus is that many word forms of the same lemma are spread all over the lists and thus the rules of the analysis require the checking of every single term in every single list.

Another thing we observed here is that the collocation of terms which appear in the same keyword list, and more especially in close positions is another index that these terms may form together a multi-word term.

Finally, the dictionary verification stage confirmed once again our view about the lack of organization and the scarcity of multi-word terms which could offer a wider spectrum of scientific knowledge to the specialists.

Table 19 Single-word candidate terms

1. light	1. λέιζερ
2. silicon	2. φως
3. frequency	3. φωτός
4. laser	4. πυριτίου
5. electrons	5. ΥΣ
6. SC	6. πυρίτιο
7. band	7. ηλεκρόνια
8. photon	8. ανιχνευτές
9. detectors	9. συχνοτήτων
10. crystal	10. φωτονίων
11. photons	11. συχνότητας
12. mass	12. κύματος
13. electron	13. ζώνη
14. pulse	14. φωτόνια
15. quantum	15. συχνότητα
16. detector	16. ακρίβεια
17. measurements	17. παλμού
18. semiconductor	18. φωτόνιο
19. constant	19. μάζα
20. pulses	20. ανώτερη
21. lasers	21. ηλεκτρονίων
22. kilogram	22. όποια
23. medium	23. υλικό
24. frequencies	24. διάθλασης
25. energy	25. φωτονίου
26. momentum	26. φαινόμενο
27. TES	27. μέτρηση
28. voltage	28. χιλιογράμμου
29. sphere	29. ανιχνευτών
30. index	30. σταθεράς
31. atoms	31. μετρήσεις
32. fiber	32. δευτερόλεπτο
33. lasing	33. κρύσταλλο
34. bands	34. ένα
35. gamma	35. ηλεκτρόνιο
36. emission	
37. atomic	

Table 20 Multi-word candidate terms

1. silicon laser	**1. φως λέιζερ**
2. frequency measurements	**2. εκπομπή φωτός**
3. high-frequency	**3. λέιζερ πυριτίου**
4. laser light	**4. ηλεκτρονίων στην ανώτερη ζώνη**
5. SC light	**5. ΥΣ φως**
6. upper band	**6. παραγωγή του ΥΣ φωτός**
7. energy band	**7. ανώτερη ζώνη**
8. photon detectors	**8. ενέργεια των φωτονίων**
9. photon's energy	**9. υπεραγώγιμοι ανιχνευτές**
10. x/gamma-ray detectors	**10. ενέργεια των φωτονίων**
11. superconducting detectors	**11. μετρήσεις συχνότητας**
12. TES detectors	**12. μετρήσεις μάζας**
13. laser pulse	**13. σταθερά του Planck**
14. quantum confinement	**14. σταθερά του Avogadro**
15. quantum-mechnical	**15. ζώνη συχνοτήτων**
16. pair-breaking detectors	**16. εύρος της ζώνης συχνοτήτων**
17. stimulated emission	**17. δείκτης διάθλασης**
18. light emission	
19. emission efficiency	
20. optical fiber	
21. refractive index	

4.5 Planetology-Cosmology Corpus

4.5.1 Overview

Planetology-Cosmology corpus is the 6th of the seven sub-corpora that compose our popular science English-Greek parallel corpus. It consists of 8 articles originally written in English and 8 translations of the articles in Greek. In this study, it is the second biggest corpus after Energy–Environment–Geology corpus and is approximately 24.767 words (the English version) and 27.211 words (the Greek version).

The reason for the double title of the corpus is explained by the headings of the magazine sections the articles appear in. The topic is related to cosmological issues and celestial bodies; thus we expect to see a high level of consistency among the articles and within the corpus.

4.5.2 Quantitative Analysis

At first sight, the keyword lists appear to be very similar. This is also an indication that the translations are close enough to the originals and our results are reliable. Therefore, for the planetology/cosmology sub-corpus, we get keyword lists –when sorting the results by keyness-, that look like this:

Table 21 Planetology-Cosmology English Keyword List

N	Word	Freq.	Planen.Lst %	Freq.	Sciamen.Lst %	Keyness	
1	Galaxies	108	0,44	108	0,08	143,3	0,000000
2	Stars	77	0,31	79	0,06	100,0	0,000000
3	Galaxy	70	0,28	71	0,05	91,8	0,000000
4	Planets	68	0,28	68	0,05	90,2	0,000000
5	Planet	70	0,28	81	0,06	82,0	0,000000
6	Star	60	0,24	68	0,05	71,6	0,000000
7	Universe	59	0,24	66	0,05	71,2	0,000000
8	Dark	64	0,26	79	0,06	70,6	0,000000
9	Black	65	0,26	86	0,06	66,9	0,000000
10	Moons	50	0,20	50	0,04	66,3	0,000000
11	Mars	49	0,20	49	0,04	64,9	0,000000
12	Astronomers	49	0,20	53	0,04	60,9	0,000000
13	Gas	104	0,42	227	0,16	54,7	0,000000

14	Cluster	41	0,17	44	0,03	51,3	0,000000
15	Constellations	38	0,15	38	0,03	50,4	0,000000
16	The	1.930	7,81	9.037	6,56	50,3	0,000000
17	Orbits	35	0,14	35	0,03	46,4	0,000000
18	Clusters	36	0,15	39	0,03	44,7	0,000000
19	Hole	36	0,15	39	0,03	44,7	0,000000
20	Formed	36	0,15	42	0,03	41,8	0,000000
21	Holes	35	0,14	43	0,03	38,7	0,000000
22	Cosmic	35	0,14	43	0,03	38,7	0,000000
23	Bodies	42	0,17	62	0,04	38,4	0,000000
24	Massive	39	0,16	54	0,04	38,3	0,000000
25	Formation	37	0,15	51	0,04	36,5	0,000000
26	Irregular	29	0,12	32	0,02	35,4	0,000000
27	Earth	42	0,17	70	0,05	33,1	0,000000
28	Bubbles	29	0,12	35	0,03	32,6	0,000000
29	B	34	0,14	49	0,04	31,9	0,000000
30	Sun	30	0,12	39	0,03	31,4	0,000000
31	Gravitational	24	0,10	26	0,02	29,8	0,000000
32	Asteroids	24	0,10	26	0,02	29,8	0,000000
33	Orbital	23	0,09	25	0,02	28,4	0,000000
34	Moon	22	0,09	23	0,02	28,1	0,000000
35	C	35	0,14	59	0,04	27,2	0,000000
36	Matter	40	0,16	75	0,05	26,8	0,000000
37	Shock	21	0,08	22	0,02	26,8	0,000000
38	Space	41	0,17	80	0,06	25,9	0,000000
39	Supernovae	19	0,08	19	0,01	25,2	0,000001
40	Jupiter	19	0,08	19	0,01	25,2	0,000001
41	Pluto	19	0,08	19	0,01	25,2	0,000001
42	Martian	19	0,08	19	0,01	25,2	0,000001
43	Use	4	0,02	138	0,10	24,2	0,000001
44	For	128	0,52	1.166	0,85	31,9	0,000000
45	Power	5	0,02	189	0,14	34,8	0,000000
46	Carbon	4	0,02	180	0,13	35,9	0,000000
47	To	453	1,83	3.462	2,51	44,1	0,000000
48	Fuel	4	0,02	222	0,16	48,1	0,000000

Table 22 Planetology-Cosmology Greek Keyword List

N	Word	Freq.	Plangr.lst %	Freq.	Sciamgr.lst %	Keyness	P
1	Γαλαξιών	66	0,24	67	0,04	88,5	0,000000
2	Γαλαξίες	62	0,23	62	0,04	84,1	0,000000
3	Άστρα	55	0,20	56	0,04	73,6	0,000000
4	Πλανήτες	53	0,19	53	0,03	71,9	0,000000
5	Άρη	47	0,17	47	0,03	63,8	0,000000
6	Άστρων	46	0,17	51	0,03	57,3	0,000000
7	Αστρονόμοι	42	0,15	45	0,03	53,9	0,000000
8	Χ	59	0,22	93	0,06	51,6	0,000000
9	Σύμπαν	38	0,14	40	0,03	49,5	0,000000
10	Σκοτεινή	37	0,14	40	0,03	47,1	0,000000
11	Πλανήτη	55	0,20	95	0,06	42,9	0,000000
12	Δορυφόρων	31	0,11	32	0,02	41,0	0,000000
13	Σμήνους	29	0,11	29	0,02	39,3	0,000000
14	Σώματα	35	0,13	44	0,03	38,9	0,000000
15	Μάζας	39	0,14	55	0,04	38,6	0,000000
16	Αέριο	57	0,21	109	0,07	38,6	0,000000
17	Τρύπα	29	0,11	30	0,02	38,3	0,000000
18	Πλανητών	29	0,11	30	0,02	38,3	0,000000
19	Δορυφόροι	28	0,10	29	0,02	36,9	0,000000
20	Π	32	0,12	39	0,03	36,6	0,000000
21	Μαύρες	27	0,10	28	0,02	35,6	0,000000
22	Γαλαξία	26	0,10	26	0,02	35,3	0,000000
23	Τρύπες	28	0,10	31	0,02	34,9	0,000000
24	Σουπερνόβα	25	0,09	25	0,02	33,9	0,000000
25	Τροχιές	25	0,09	25	0,02	33,9	0,000000
26	Σμήνη	25	0,09	25	0,02	33,9	0,000000
27	Ύλη	34	0,12	50	0,03	32,2	0,000000
28	Οι	499	1,83	2.133	1,37	32,1	0,000000
29	Γύρω	41	0,15	74	0,05	30,1	0,000000
30	Μαύρη	25	0,09	29	0,02	29,9	0,000000
31	Σπιν	22	0,08	22	0,01	29,8	0,000000
32	Πλανήτης	23	0,08	26	0,02	28,2	0,000000
33	Αστερισμούς	20	0,07	20	0,01	27,1	0,000000

34	Τους	291	1,07	1.172	0,75	26,5	0,000000
35	Αστεροειδείς	21	0,08	23	0,01	26,4	0,000000
36	Σύμπαντος	24	0,09	31	0,02	26,0	0,000000
37	Δία	19	0,07	19	0,01	25,8	0,000000
38	Γη	30	0,11	49	0,03	25,1	0,000001
39	Σωμάτων	20	0,07	22	0,01	25,1	0,000001
40	Άστρο	18	0,07	18	0,01	24,4	0,000001
41	Σμήνος	18	0,07	18	0,01	24,4	0,000001
42	Έκρηξη	22	0,08	28	0,02	24,2	0,000001
43	Να	505	1,85	3.656	2,35	27,4	0,000000
44	Άνθρακα	4	0,01	182	0,12	35,4	0,000000
45	Για	181	0,66	1.654	1,06	41,5	0,000000

Another remarkable point as a result of the lack of lemmatisation in the corpus is the appearance of many word forms of the same lemma. In other words, in the English list, we have both singular and plural forms of the same word, like *galaxies-galaxy, stars-star, planets-planet, moons-moon, cluster-clusters, hole-holes*; whereas in the more highly-inflected Greek, words appear in both numbers and in different cases, e.g. *γαλαξιών* (plural; genitive) – *γαλαξίες* (plural; nominative/accusative) – *γαλαξία* (singular; genitive/accusative), *άστρα* (plural; nominative/accusative) – *άστρων* (plural; genitive) – *άστρο* (singular; nominative/accusative), *πλανήτες* (plural; nominative/accusative) – *πλανήτη* (singular; genitive/accusative) – *πλανητών* (plural; genitive) – *πλανήτης* (singular; nominative), *σύμπαν* (singular; nominative/accusative) – *σύμπαντος* (singular; genitive), *δορυφόρων* (plural; genitive) – *δορυφόροι* (plural; nominative), *σμήνους* (singular; genitive) – *σμήνη* (plural; nominative/accusative) – *σμήνος* (singular; nominative/accusative), *σώματα* (plural; nominative/accusative) – *σωμάτων* (plural; genitive), *τρύπα* (singular; nominative/accusative) – *τρύπες* (plural; nominative/accusative).

But let us now see the results of the keyword lists and check what the figures tell us about them. The word *galaxies* is expected to match with the plural forms of its Greek equivalent: *γαλαξιών* and *γαλαξίες*. Their sum (128), however, outnumbers by 20 the number of occurrences of *galaxies*. One reason for this might be that, for example, English uses a pronoun ("they/them", for example) where the Greek translator uses a

noun –this if it happens would be in line with ideas about translations being more "explicit" than originals.

Furthermore, the singular form *galaxy*, this time, outnumbers its unique Greek equivalent that appears in the list, *γαλαξία*. Their arithmetic difference is significant; their parallel concordances, however, demonstrate that *galaxy* is also translated in Greek as an adjective (*γαλαξιακών*), as a noun in a different case (*γαλαξίας*) and even as a noun in a different number (*γαλαξίες*).

More or less the case is the same for most of the words appearing in more than one word form in the list enumerated above, and the difference in figures is more clearly illustrated in the table of Multiconc parallel concordances. Consequently, rather than continue exhaustively with this type of counting and listing it would be more useful to comment on the few cases they have been left in the lists.

Consequently, the words that interest us are: *universe* which shows the same number of occurrences (22) as its two equivalents, *σύμπαν* and *σύμπαντος*; *constellations* that differs in number from its Greek equivalent *αστερισμούς*, its equivalent noun that occurs in the keyword list; *orbits* which also differs from the Greek *τροχιές* by 10 instances, the looking at the concordances showed that in some cases, the translator decided to translate *orbits* either by another word or periphrastically; *asteroids* which differs from the Greek *αστεροειδείς* only by 1.25%; *matter* which corresponds in 85% of the citations to the Greek *ύλη*; and supernovae which has a significant difference of 24% from the Greek *σουπερνόβα*. Here we have to say that *σουπερνόβα* is the equivalent of both singular and plural form, and the form *supernova* does not appear in the English list so as to be added to the plural form. All in all, we assume that the number of instances missing correspond to the singular form *supernova*.

Table 23 Multiconc Parallel Concordances

H:\multconc\plan.en P35 S2 One realization has already sunk in: although **dark energy** betrayed its existence through its effect on the universe as a whole, it may also shape the evolution of the universe's inhabitants--**stars**, **galaxies**, **galaxy clusters**.
H:\multconc\plan.gr P35 Ήδη έχουν αρχίσει να σκιαγραφούνται κάποια χαρακτηριστικά της: Αν και η **σκοτεινή ενέργεια** πρόδωσε την ύπαρξη της μέσω της επίδρασης που ασκεί στο Σύμπαν ως όλον, ενδέχεται ίσως να διαμορφώνει και την εξέλιξη όσων το ενοικούν —των **άστρων**, των **γαλαξιών** και των **γαλαξιακών σμηνών**.
H:\multconc\plan.en P118 S5 In the 1970s theorists proposed three possible mechanisms, all functioning during or soon after the epoch of **planet formation**.
H:\multconc\plan.gr P118 Στη δεκαετία του 1970, διάφοροι θεωρητικοί πρότειναν τρεις σχετικούς μηχανισμούς, όλοι τους δε θεωρείται πως επενήργησαν κατά τη διάρκεια ή λίγο μετά την εποχή **σχηματισμού των πλανητών**.
H:\multconc\plan.en P228 S2 This argument overlooks the fact that astronomers classify all objects that **orbit planets** as "**moons**," although two of them are larger than the planet Mercury and many are captured **asteroids** and comets.
H:\multconc\plan.gr P228 Αυτό το επιχείρημα παραβλέπει το γεγονός ότι οι αστρονόμοι κατατάσσουν όλα τα **αντικείμενα που κινούνται σε τροχιά** γύρω από πλανήτες ως «**δορυφόρους**», αν και δύο από αυτούς ξεπερνούν σε μέγεθος τον πλανήτη Ερμή και πολλοί από αυτούς είναι «αιχμαλωτισμένοι» **αστεροειδείς** και κομήτες.
H:\multconc\plan.en P52 S3 Such **holes** power quasars and other types of **active galaxies**, which are rare in the modern **universe**; <s>the **black holes** in our **galaxy** and others are quiescent.

H:\multconc\plan.gr P52 Τέτοιες **μαύρες τρύπες** τροφοδοτούν τους κβάζαρ και άλλους τύπους **ενεργών γαλαξιών**, οι οποίοι απαντούν σπάνια στο σύγχρονο **Σύμπαν** οι **μαύρες τρύπες** του δικού μας **Γαλαξία**, καθώς και άλλων γαλαξιών φυσικά, είναι ανενεργές.
H:\multconc\plan.en P265 S1 The jets blast through the **galaxy** and out into the **cluster gas**, where their energy converts to heat. H:\multconc\plan.gr P265 Οι πίδακες εκτινάσσουν ύλη και ενέργεια τόσο μέσα στο **γαλαξία** όσο και έξω από αυτόν, στο χώρο του **σμήνους που καλύπτεται από το αέριο•**η ενέργεια εκεί μετατρέπεται σε θερμότητα.
H:\multconc\plan.en P265 S5 Millions of years later the **hot gas** in the central region of the **cluster** finally cools sufficiently to initiate a new season of growth for the **galaxy** and its **supermassive black hole**, and thus the cycle continues.
H:\multconc\plan.gr P265 Εκατομμύρια χρόνια αργότερα, το **θερμό αέριο** στην κεντρική περιοχή του **σμήνους** ψύχεται επαρκώς, δίνοντας το έναυσμα για μια νέα εποχή ανάπτυξης του **γαλαξία** και της **υπέρμαζης μαύρης τρύπας** του, και με αυτό τον τρόπο ο κύκλος συνεχίζεται.
H:\multconc\plan.en P269 S1 THE SCENARIO IS ENRICHED by **galaxy collisions**, an ever present hazard in the central regions of **galaxy clusters**.
H:\multconc\plan.gr P269 ΤΟ ΠΑΡΑΠΑΝΩ ΣΕΝΑΡΙΟ ενισχύεται από τις **γαλαξιακές συγκρούσεις** —ένας πανταχού παρών κίνδυνος στις κεντρικές περιοχές των **γαλαξιακών σμηνών**.

4.5.3 Qualitative Analysis

From the quantitative results, we conclude that these words are representative in the planetology/cosmology sub-corpus and some of them may be terms as well. In order to verify this, we check the concordances of the words which we think may be candidate terms.

Here, we have to point out that adjectives which appear in the English keyword list will also constitute part of the study, but they will not appear as single-word terms in the final term list.

Starting from the word with the highest keyness and the highest frequency, *galaxies*, we observe that it appears in clusters such as: *massive galaxies* (11/108) and *dwarf galaxies* (4/108). Its singular form, *galaxy* is also used as a noun: *central galaxy* (5/70), but it is frequently used as an adjective as well: *galaxy cluster(s)* (17/70), *galaxy formation* (5/70), *galaxy merger(s)* (5/70).

For the highly inflectional Greek, we have to face a very complicated situation as we described in the quantitative analysis. The fact that in every case we talk about different word forms of the same lemma increases the chances to come across the same clusters in all of its forms. If these are not the same, we still have something important to talk about; since we would have discovered an important collocation which is only representative of the number or the case in which it appears.

Hence, for the plural, genitive form *γαλαξιών*, we have the following collocations: *σχηματισμού/οί/ό γαλαξιών* (*galaxy formation*) (7/66), *σμήνη/ους γαλαξιών* (*galaxy cluster(s)*) (6/66), *συγχωνεύσεις γαλαξιών* (*galaxy merger(s)*) (3/66). For the plural and nominative form *γαλαξίες*, we have the collocations: *γαλαξίες μεγάλης μάζας* (*hypermassive galaxies*) (8/66), *νάνοι/ους γαλαξίες* (*dwarf galaxies*) (4/66), *σπειροειδείς γαλαξίες* (*spiral galaxies*) (3/66). For the singular form *γαλαξίας,* we

have the following clusters: *κεντρικό γαλαξία* (central galaxy) (5/26), *γιγαντιαίο/ου γαλαξία* (*giant galaxy*) (2/26). The above is a representative example of translator's decisions with regards to the translation product. We may not know the reasons which led him to make this or that decision in a particular time and space frame, but we see how all these are reflected in his translation and what impact these may have on the target audience.

The same thing we see it happening in the cases of *star* and *planet*. The plural form of *star* is found in clusters like *new stars* (6/77), *neutron stars* (2/77), *massive stars* (2/77) and the plural form of *planet* in *giant planets* (9/68)*, host planets* (2/68) *and terrestrial planets* (2/68). Their singular forms are used:

- for *star* in: *star formation* (15/60), *star groups* (5/60) and *star pictures* (4/60) but also in the same clusters as in plural: *neutron star* (4/60) and *new star* (2/60)
- for *planet* in: *planet formation* (5/70) and *Red Planet* (9/70)

The various inflectional forms of the above candidate terms in Greek are demonstrated in the following translator's choices:

- for *πλανήτης* in: *γίγαντες πλανήτες* (*giant planets* in accusative, plural) (7/52), *γιγάντων πλανητών* (*giant planets* in genitive plural) (3/29), *Κόκκινου Πλανήτη* (*Red Planet* in genitive singular) (5/55), *Κόκκινος Πλανήτης* (*Red Planet* in nominative singular) (2/23), and
- for *άστρο* in: *σχηματισμούς άστρων* (*star formations* in genitive plural) (13/46), *ομάδες/ων άστρων* (*star groups* in genitive plural) (4/46)

Another interesting case is *moons* –as for being or not a multi-word term– because of the high frequency in which it appears next to the adjective *irregular* (21/50). To the same way, in Greek, the translation of the above cluster by the same pattern *ανώμαλων δορυφόρων* (13/31) (adjective+noun in plural, genitive) and *ανώμαλοι δορυφόροι* (9/28) (adjective+noun in plural, nominative) is also very frequent, thus we decided to include it in the list –along with its translation- as a multi-word term.

The next candidate term to be investigated is *cluster*. In the concordances, *cluster* is found alone, but most of the times, it appears in collocations, where the words that occur next to it, define what kind of cluster it is, e.g. *cluster gas* (5/41), *Perseus*

cluster (3/41), *Virgo cluster* (2/41), etc. In plural, *clusters*, collocates more frequently with *galaxy*, forming the *galaxy clusters* (14/36). In Greek, the collocations are of the same kind: *σμήνους γαλαξιών* (2/29) (*galaxy clusters* in genitive, singular), *σμήνη γαλαξιών* (4/25) (*galaxy clusters* in nominative/accusative plural), *γαλαξιακά σμήνη* (5/25) (*galaxy clusters* but in pattern adjective+noun in plural), *σμήνος της Παρθένου* (4/18) (*Virgo cluster*), *σμήνος του Περσέα* (3/18) (*Perseus cluster*), *γαλαξιακό σμήνος* (2/18) (*galaxy clusters* but in pattern adjective+noun in singular).

Last, we will comment on the fixed collocation *black holes* which is a multi-word term because none of its components can give alone the meaning that they both form in cluster. They show 100% co-occurrence (36/36) in both the English corpus and the Greek one (*μαύρες τρύπες*) (27/27).

4.5.4 Dictionary Verification

Dictionary checking confirmed the existence of many single-word and some multi-word terms. The appearance in technical dictionaries of terms like *galaxy, star, planet, moons and universe* is expected. Nonetheless, collocations like *active galaxies, giant planets, black holes, dark matter* and others which exist as well in Chambers Dictionary of Science and Technology are there as a result of their frequent use and their standardization in language. The rest that have not been found in the dictionary have been crosschecked in other available and reliable resources (see website).

4.5.5 Summary-Remarks

Again during the analysis of this corpus, we attempted to remain faithful to the targets we set at the beginning of this study and try to understand the relation between the quantitative results we get from the keyword lists and the actual clues we get from the concordance lines.

Unfortunately, the lack of lemmatizers and taggers hampers the analysis and restricts to a great extent our potentials. However we were fully aware of that since the beginning and thus the scope of our study had to come to terms with it. The methodology that is being suggested here intends to show a primary connection between keyword lists, concordances and dictionaries for Greek and shed light to what constitutes or not representativeness within a genre. A more detailed study would definitely demand deeper research in all levels: grammatical, syntactical, stylistic, psychological and others.

Table 24 Single-word candidate terms

1. galaxies	**1. γαλαξιών**
2. stars	**2. γαλαξίες**
3. galaxy	**3. άστρα**
4. planets	**4. πλανήτες**
5. planet	**5. άστρων**
6. star	**6. σύμπαν**
7. universe	**7. πλανήτη**
8. moons	**8. δορυφόρων**
9. gas	**9. σμήνους**
10. cluster	**10. σώματα**
11. constellations	**11. μάζας**
12. orbits	**12. αέριο**
13. clusters	**13. τρύπα**
14. hole	**14. πλανητών**
15. holes	**15. δορυφόροι**
16. bodies	**16. μαύρες**
17. formation	**17. γαλαξία**
18. earth	**18. τρύπες**
19. bubbles	**19. σούπερνόβα**
20. sun	**20. τροχιές**
21. asteroids	**21. σμήνη**
22. moon	**22. ύλη**
23. matter	**23. σπιν**
24. shock	**24. πλανήτης**
25. space	**25. αστερισμούς**
26. supernovae	**26. αστεροειδείς**
	27. σύμπαντος
	28. σωμάτων
	29. άστρο
	30. σμήνος
	31. έκρηξη

Table 25 Multi-word candidate terms

1. **massive galaxies**	1. **σχηματισμός γαλαξιών**
2. **galaxy cluster**	2. **γαλαξίες μεγάλης μάζας**
3. **star formation**	3. **γίγαντες πλανήτες**
4. **giant planets**	4. **σχηματισμός άστρων**
5. **Red planet**	5. **πρώιμο σύμπαν**
6. **irregular moons**	6. **ανώμαλοι δορυφόροι**
7. **cluster gas**	7. **σκοτεινή ενέργεια**
8. **Greek constellations**	8. **σκοτεινή ύλη**
9. **black holes**	9. **μαύρες τρύπες μεγάλης μάζας**
10. **dark energy**	10. **σώμα μεγάλης μάζας**
11. **dark matter**	11. **ομαλοί δορυφόροι**
12. **cosmic expansion**	12. **μαύρες τρύπες**
13. **shock wave**	13. **υπέρμαζη μαύρη τρύπα**
14. **shock front**	14. **περιστρεφόμενη μαύρη τρύπα**
15. **strong shock waves**	15. **κεντρικό γαλαξία**
16. **small bodies**	16. **έκρηξης σουπερνόβα**
17. **dark ages**	17. **υπέρμαζες μάυρες τρύπες**
18. **cosmic history**	18. **σμήνη γαλαξιών**
	19. **γαλαξιακά σμήνη**
	20. **σκοτεινή ύλη**
	21. **θερμοκρασία σπιν**
	22. **σπιν του ηλεκτρονίου**
	23. **αντιστροφή του σπιν**
	24. **ουράνιων σωμάτων**
	25. **σμήνος της Παρθένου**
	26. **σμήνος του Περσέα**
	27. **γαλαξιακό σμήνος**
	28. **Μεγάλη Έκρηξη**

4.6 Psychology Corpus

4.6.1 Overview

The Psychology corpus consists of 5 articles and 5 translations of these articles and its size is estimated at 17.131 words the English version and 19.222 words the translated Greek version. The special characteristic of this corpus is the thematic relevance of the component articles. In the keyword lists we extracted for every article, there are many common terms that appear in more than two separate articles keyword lists. More specifically, the word *neuron(s)* appears in three articles keyword lists; the word *mirror* –which as we will see below constitutes a multi-word term together with the word *neuron(s)*- appears in two articles keyword lists; and the word *brain* in two articles keyword lists as well.

4.6.2 Quantitative Analysis

The keyword lists we are about to examine in Psychology corpus contain 36 (the English) and 24 words (the Greek) respectively.

The keyword lists has the structure shown below:

Table 26 Psychology Corpus English Keyword List

N	Word	Freq.	Psychen.Lst %	Freq.	Sciamen.Lst %	Keyness	P
1	Neurons	119	0,70	121	0,09	221,4	0,000000
2	Mirror	87	0,51	93	0,07	156,7	0,000000
3	Brain	75	0,44	87	0,06	127,9	0,000000
4	Autism	64	0,38	64	0,05	120,2	0,000000
5	Neuron	53	0,31	53	0,04	99,5	0,000000
6	Color	51	0,30	63	0,05	83,2	0,000000
7	Chess	44	0,26	44	0,03	82,6	0,000000
8	Neural	35	0,21	35	0,03	65,7	0,000000
9	Cortex	31	0,18	31	0,02	58,2	0,000000
10	Motor	34	0,20	49	0,04	49,2	0,000000
11	He	42	0,25	88	0,06	42,4	0,000000
12	Visual	24	0,14	27	0,02	41,7	0,000000
13	Action	27	0,16	37	0,03	40,7	0,000000
14	Monkey	22	0,13	23	0,02	40,1	0,000000
15	Whiskers	20	0,12	20	0,01	37,5	0,000000

16	Rats	21	0,12	23	0,02	37,2	0,000000
17	Memory	23	0,14	29	0,02	36,9	0,000000
18	Activity	35	0,21	74	0,05	35,0	0,000000
19	Children	25	0,15	39	0,03	33,8	0,000000
20	Subjects	20	0,12	25	0,02	32,3	0,000000
21	Cortical	17	0,10	17	0,01	31,9	0,000000
22	We	103	0,60	433	0,31	31,2	0,000000
23	Grasping	17	0,10	18	0,01	30,8	0,000000
24	His	33	0,19	75	0,05	30,2	0,000000
25	Tactile	16	0,09	16	0,01	30,0	0,000000
26	Sensory	16	0,09	17	0,01	28,9	0,000000
27	Monkeys	16	0,09	17	0,01	28,9	0,000000
28	Responses	19	0,11	26	0,02	28,7	0,000000
29	When	80	0,47	318	0,23	27,9	0,000000
30	Actions	19	0,11	27	0,02	27,8	0,000000
31	Master	16	0,09	19	0,01	26,8	0,000000
32	Information	27	0,16	59	0,04	25,9	0,000000
33	Movements	15	0,09	18	0,01	24,9	0,000001
34	Grandmaster	13	0,08	13		24,4	
35	VPM	13	0,08	13		24,4	
36	Players	15	0,09	19	0,01	24,0	0,000001
37	Years	8	0,05	261	0,19	24,2	0,000001
38	Are	52	0,31	799	0,58	24,4	0,000001

Table 27 Psychology Corpus Greek Keyword List

N	Word	Freq.	Psychgr.Lst %	Freq.	Sciamgr.Lst %	Keyness	P
1	Νευρώνων	84	0,42	84	0,05	152,9	0,000000
2	Νευρώνες	67	0,34	69	0,04	119,6	0,000000
3	Μετρ	43	0,22	43	0,03	78,2	0,000000
4	Κατόπτρων	38	0,19	41	0,03	65,8	0,000000
5	Εγκεφάλου	36	0,18	42	0,03	59,0	0,000000
6	Κινήσεις	32	0,16	35	0,02	54,9	0,000000
7	Κάτοπτρα	31	0,16	34	0,02	53,1	0,000000
8	Χρώμα	29	0,15	34	0,02	47,4	0,000000
9	Εγκέφαλο	26	0,13	27	0,02	46,1	0,000000
10	Αυτισμό	24	0,12	24	0,02	43,6	0,000000

11	Παιδιά	29	0,15	39	0,03	42,7	0,000000
12	Φλοιού	24	0,12	26	0,02	41,4	0,000000
13	Δραστηριότητα	30	0,15	48	0,03	38,1	0,000000
14	Σκακιού	20	0,10	20	0,01	36,4	0,000000
15	Ερεθισμάτων	18	0,09	18	0,01	32,7	0,000000
16	Φλοιό	19	0,10	21	0,01	32,3	0,000000
17	Αυτισμού	17	0,09	17	0,01	30,9	0,000000
18	Όταν	76	0,38	276	0,18	29,9	0,000000
19	Μουστάκια	16	0,08	16	0,01	29,1	0,000000
20	Σκάκι	14	0,07	14		25,4	0,000000
21	Μας	96	0,48	410	0,26	24,8	0,000001
22	Χρώματος	15	0,08	18	0,01	24,1	0,000001
23	Μνήμη	15	0,08	18	0,01	24,1	0,000001
24	Αποκρίσεις	15	0,08	18	0,01	24,1	0,000001
25	Θα	104	0,52	1.317	0,85	26,3	0,000000
26	Ενέργεια	4	0,02	248	0,16	36,2	0,000000
27	Ενέργειας	3	0,02	235	0,15	37,6	0,000000

At the top of both lists, we see the term *neurons*, which, as we mentioned in the overview, is representative of the corpus, since it exists in the three of the six articles that consist the sub-corpus. More analytically, in the English list we have two word forms (one singular and one plural) of *neuron*. In the Greek list, although we have also two word forms of *νευρώνας* (*neuron*), they are both plural but in different case (*νευρώνες*: nominative/accusative; *νευρώνων*: genitive). This is because the singular *neuron* is also used as a determiner preceding other nouns; whereas in a similar case in Greek, the genitive case is more frequently used instead of an adjective.

The second term of our list is *mirror* in singular. Its Greek equivalent appears in two word forms which appear in the 4th (*κατόπτρων*: plural; genitive) and 7th place (*κάτοπτρα*: plural; nominative/accusative) respectively. Their accordance in number and case with the word forms *νευρώνων* and *νευρώνες* lead us assume that they may be multi-word terms; but the quantitative analysis is at too early stage to decide on that.

The third candidate term of the English list is *brain*. In the Greek list its equivalents come fifth (*εγκεφάλου*: singular; genitive) and ninth (*εγκέφαλο*: singular; accusative). Their sum differs from the English original by 13 instances, but as we checked in Multiconc parallel concordances, the remaining instances are translated in Greek by an adjective, which, due to the fact that its times of occurrence are few, does not appear in the keyword list.

The fourth term of the list is the word *autism*. Its Greek equivalent comes 10th (*αυτισμό*: nominative/accusative) and 17th (*αυτισμού*: genitive) (the two word forms appear again in different cases) in the list. There are 23 more occurrences of these forms in Greek than in English and the parallel concordances show that some of these extra occurrences are due to the term being translated by an adjective, or the expression "people with autism" being translated as αυτιστικοί (literally, "autistics") Here, we would like to comment on the fact that while in the English list there are some adjectives, such as *neural, visual, cortical, tactile* and *sensory*; in the Greek list there are no adjectives. Moreover, in the Greek list we have an abundance of genitives, which reinforce our initial claim that the English adjectives can frequently be translated by Greek genitives.

Another case that confirms this argument is that of the candidate term *cortex*. Its Greek equivalents are *φλοιού* (genitive) and *φλοιό* (accusative) but their sum (13) is not equal to the 31 instances of *cortex*. The existence, however, of the adjective *cortical* on the 21st position of the English list seems to compensate for this imbalance. We cannot however be absolutely sure that all occurrences of *cortical* are translated by a genitive and not by an adjective; because the adjective does not appear in the Greek list.

Table 28 Multiconc Parallel Concordances

H:\multconc\psy.en P52 S2 Many **people with autism** have problems understanding metaphors, sometimes interpreting them literally.
H:\multconc\psy.gr P52 Πολλοί **αυτιστικοί** εμφανίζουν δυσκολίες στην κατανόηση μεταφορικών γλωσσικών σχημάτων, τα οποία συχνά ερμηνεύουν κατά κυριολεξία.
H:\multconc\psy.en P116 S5 And because lack of emotional mirroring ability appears to be a hallmark of **autism**, we are also working with young **autistic** children to learn whether they have detectable motor deficits that could signal a general dysfunction of the mirror neuron system.
H:\multconc\psy.gr P116 Και επειδή μάλιστα μία από τις «σφραγίδες» του **αυτισμού** είναι η αδυναμία για συναισθηματικό καθρέφτισμα, τελευταία εργαζόμαστε και με νεαρά **αυτιστικά** παιδιά για να εξετάσουμε αν έχουν ανιχνεύσιμα κινητικά ελλείμματα που θα σηματοδοτούσαν μια γενική δυσλειτουργία του συστήματος των νευρώνων-κατόπτρων τους.

H:\multconc\psy.en P53 S3) Eric Courchesne of U.... and other anatomists have shown elegantly that **children with autism** have characteristic abnormalities in the cerebellum, the **brain** structure responsible for coordinating complex voluntary muscle movements.

H:\multconc\psy.gr P53) Ο Eric Courchesne, του Πανεπιστημίου της Καλιφόρνιας στο Σαν Ντιέγκο, και άλλοι ανατόμοι έχουν δείξει ότι τα **αυτιστικά** παιδιά εμφανίζουν χαρακτηριστικές ανωμαλίες της παρεγκεφαλίδας, της **εγκεφαλικής** δομής που είναι υπεύθυνη για το συντονισμό περίπλοκων, εκούσιων συσπάσεων των μυών.

H:\multconc\psy.en P56 S3 **Brain**-imaging techniques subsequently showed that these so-called **mirror neurons** also exist in the corresponding regions of the human **cortex**.

H:\multconc\psy.gr P56 Μελέτες απεικόνισης του **εγκεφάλου** έδειξαν αργότερα ότι τούτοι οι αποκαλούμενοι **νευρώνες-κάτοπτρα** υπάρχουν επίσης και στις αντίστοιχες περιοχές του ανθρώπινου εγκεφαλικού φλοιού.

4.6.3 Qualitative Analysis

First of all, let us begin with the term with the highest keyness in both lists: the word *neurons* and its translation, *νευρώνες/νευρώνων*. In English corpus, we find it in the following clusters, starting from the most frequent: *mirror-neurons* (42/119), *VPM neurons* (10/119), *individual neurons* (7/119), *single neurons* (5/119), *cortical neurons* (4/119) and some others with lower frequency, to which for the sake of economy we will not refer here. In the Greek corpus, the Greek equivalents of *neurons*, *νευρώνες/νευρώνων* appear in the same clusters: *νευρώνων-κατόπτρων* (38/84) / *νευρώνες- κάτοπτρα* (30/67), *νευρώνων του VPM* (4/84), *μεμονωμένων νευρώνων* (11/84) / *μεμονωμένους νευρώνες* (2/67), *νευρώνες του φλοιού* (5/64).

As illustrated above, the most frequent is the collocation mirror-neuron(s) – *νευρώνων-κατόπτρων*/*νευρώνες-κάτοπτρα*. We assume that this cluster is a multi-word term. To this claim it is added the percentage of 90% (78/87) of the times that *mirror* collocates with *neuron(s)* and the percentage of 100% (38/38)/(30/30) of the times that *νευρώνων* collocates with *κατόπτρων* and *νευρώνες* collocates with *κάτοπτρα* in the Greek corpus. Together they form another strong collocation with a third term, that of *system* (22/87) – *συστήματος/σύστημα* (15/38) and appear as a three-word term in *mirror-neuron system – συστήματος/σύστημα νευρώνων-κατόπτρων*.

Furthermore, the term *brain* is found in the following clusters: *human brain* (5/75)*, brain areas* (3/75)*, brain stem* (3/75) *and brain structures* (3/75). In Greek corpus, the Greek equivalent *εγκεφάλου* is found in the next collocations -which are relatively few in number: *ανθρώπινου εγκεφάλου (human brain) (*3/36), *περιοχές του εγκεφάλου* (*brain areas)* (6/36), *δομές του εγκεφάλου (brain structures)* (2/36). In the case of *brain*, we will go the other way round. That is to say, for the word *brain*, in Greek there are two equivalents: *μυαλό* and *εγκέφαλος*, which according to Dorland's

Medical Dictionary (1997:1014) *μυαλό* is more colloquial than *εγκέφαλος*, which is more technical. Consequently, because the term *εγκέφαλος* appears in the Greek corpus as equivalent for *brain*, we conclude through inductive thinking that *brain* constitutes also a technical term.

The next term to be examined is *cortex–φλοιού/φλοιό*. Although the clusters in which it appears are not so numerous, they are all highly technical terms existing in technical dictionaries, e.g. *visual cortex* (6/31) (1999:1241); *motor cortex* (5/31) (1999:759); *cingulate cortex* (1999:213), etc. Equivalent translations in the Greek corpus which correspond to the above English collocations are the following: *οπτικού φλοιού* (2/24); *κινητικού φλοιού* (2/24). Other collocations in Greek are: *νευρώνες/νευρώνων του φλοιού* (*cortical neurons*) (6/24) and *περιοχές του φλοιού* (*neuron areas*) (5/24).

Finally, words like *action(s), activity* and *responses*, have been already mentioned above as collocating with some of the above-examined technical terms and sometimes even forming multi-word terms. Some examples are: *mirror neuron activity – δραστηριότητα των νευρώνων-κατόπτρων* and *neuron responses – αποκρίσεις των νευρώνων*.

4.6.4 Dictionary Verification

The above terms have been checked and the results showed that single words like *neuron(s), brain, autism,* and *cortex* occur in technical dictionaries; but words like *color, chess, monkey, rats,* and *children* do not.

As for the multi-word terms, some of them have been verified in technical dictionaries and some of them on the Internet.

4.6.5 Summary-Remarks

The analysis of this corpus brought up issues like: what is and what is not a technical term; whether there are, and if so, how many levels of technicality in the ranking of a technical term? And how can we distinguish between a multi-word term and a collocation? (for the same issues see also Chung, Nation: 2004)

All these are hard to answer, since there will always be a small percentage of doubt, because we are talking about language. Let us now analyze in greater depth some of our findings from the analysis of Psychology Corpus.

In the keyword list we have terms like *color, chess, monkey, whiskers, rats, children* and others which at first sight do not seem to be technical enough to be comprised in the term list. Nevertheless, these words have been checked in the concordances as well as in the dictionary, but even then, they did not demonstrate any signs of technicality and thus they have been left out of the final term list.

Although the scope of this study is not to evaluate the translations, we cannot skip a translator's slip we noticed during the comparative analysis of a term. The term is the adjective *tactile* which according to Oxford's Dictionary (1995:1214) is "something of or using the sense of touch". In the Greek translation, however, and during the parallel concordances examination, we discovered that *tactile* has been translated in Greek as *οπτικός* (*optical*) and that is another one (but rare) reason for numbers not to correspond to each other across the two languages.

Table 29 Single-word candidate terms

1. neurons	**1. νευρώνων**
2. mirror	**2. νευρώνες**
3. brain	**3. μετρ**
4. autism	**4. κατόπτρων**
5. neuron	**5. εγκεφάλου**
6. color	**6. κινήσεις**
7. chess	**7. κάτοπτρα**
8. neural	**8. χρώμα**
9. cortex	**9. εγκέφαλο**
10. motor	**10. αυτισμό**
11. action	**11. φλοιού**
12. memory	**12. δραστηριότητα**
13. activity	**13. σκακιού**
14. subjects	**14. ερεθισμάτων**
15. responses	**15. φλοιό**
16. actions	**16. αυτισμού**
17. master	**17. σκάκι**
18. information	**18. χρώματος**
19. movements	**19. μνήμη**
20. grandmaster	**20. αποκρίσεις**
21. VPM	

Table 30 Multi-word candidate terms

1. **mirror neurons**	1. **νευρώνων-κατόπτρων**
2. **mirror neuron system**	2. **μεμονωμένων νευρώνων**
3. **neuron activity**	3. **σύστημα των νευρώνων-κατόπτρων**
4. **mirror neuron**	4. **νευρώνες-κάτοπτρα**
5. **motor cortex**	5. **μεγάλος μετρ του σκακιού**
6. **visual cortex**	6. **μετρ του σκακιού**
7. **chess master**	7. **διεθνείς μετρ σκακιού**
8. **chess players**	8. **δραστηριότητα των νευρώνων-κατόπτρων**
9. **chess position**	9. **δραστηριότητα των νευρώνων**
10. **motor acts**	10. **δραστηριότητα των μεμονωμένων νευρώνων**
11. **motor command neuron**	11. **περιοχές του εγκεφάλου**
12. **action potentials**	12. **νευρώνες/ων φλοιού**
13. **cortical neurons**	13. **περιοχές του φλοιού**
14. **cortical layer**	14. **δραστηριότητα του εγκεφάλου**
15. **long-term memory**	15. **εγκεφαλική δραστηριότητα**
16. **working memory**	16. **τοπίο προεξερχόντων ερεθισμάτων**
17. **VPM neurons**	17. **οπτικών ερεθισμάτων**
18. **sensory information**	18. **οπτικό φλοιό**
19. **tactile information**	19. **κινητικό φλοιό**
20. **neuron responses**	20. **μακρόχρονη μνήμη**
21. **emotional responses**	21. **μνήμη εργασίας**
22. **autonomic responses**	22. **αποκρίσεις των νευρώνων**
	23. **αποκρίσεις του αυτόνομου νευρικού συστήματος**

4.7 Technology Corpus

4.7.1. Overview

Technology corpus is the last corpus we are going to investigate in this study. It is composed of 6 articles (15.694 words) and 6 translations (17.782 words) and it is a bit bigger than Biology-Anthropology corpus which contains the same number of articles.

Although the texts included cover a big range of topics: Aeronautics (1 article), Robotics (2 articles) and Information Technology (3 articles), it has been decided to choose a title that would be wide enough to comprise all topics and at the same time to be representative of all of them; thus the title Technology was considered as the most appropriate one for this purpose.

4.7.2 Quantitative Analysis

The keyword lists we retrieved from Wordsmith 3.0 contain 33 words (the English list) and 22 words (the Greek list) respectively, and apart from a couple of cases, the Greek terms are all matched with their English equivalents –although this is not the case for all the terms that appear in the English list. However, there is a significant difference in numbers on which we are going to comment both in the quantitative and the qualitative analysis; but let us now see how the two keyword lists look like:

Table 31 Technology Corpus English Keyword List

N	Word	Freq.	Techen.Lst %	Freq.	Sciamen.Lst %	Keyness	P
1	Scramjet	41	0,26	41	0,03	81,4	0,000000
2	Robots	37	0,23	39	0,03	71,1	0,000000
3	Mobile	35	0,22	37	0,03	67,2	0,000000
4	Computer	42	0,26	62	0,04	63,9	0,000000
5	Malware	32	0,20	32	0,02	63,5	0,000000
6	Mach	31	0,19	31	0,02	61,5	0,000000
7	Robot	31	0,19	32	0,02	60,4	0,000000
8	Engine	39	0,25	57	0,04	59,8	0,000000
9	Analog	30	0,19	30	0,02	59,5	0,000000
10	Software	30	0,19	32	0,02	57,3	0,000000
11	Digital	32	0,20	39	0,03	56,0	0,000000
12	Air	41	0,26	78	0,06	49,8	0,000000
13	Cable	26	0,16	28	0,02	49,3	0,000000

14	Robotics	24	0,15	24	0,02	47,6	0,000000
15	Flight	24	0,15	24	0,02	47,6	0,000000
16	TV	24	0,15	24	0,02	47,6	0,000000
17	Phones	24	0,15	25	0,02	46,5	0,000000
18	Channels	25	0,16	32	0,02	42,3	0,000000
19	PC	20	0,13	20	0,01	39,7	0,000000
20	Tracing	19	0,12	19	0,01	37,7	0,000000
21	Ball	20	0,13	23	0,02	36,4	0,000000
22	Smartphones	18	0,11	18	0,01	35,7	0,000000
23	DTV	18	0,11	18	0,01	35,7	0,000000
24	Ballbot	17	0,11	17	0,01	33,7	0,000000
25	Program	22	0,14	35	0,03	31,5	0,000000
26	To	516	3,25	3.462	2,51	28,4	0,000000
27	Phone	15	0,09	17	0,01	27,5	0,000000
28	Ray	26	0,16	57	0,04	27,2	0,000000
29	Hytech	13	0,08	13		5,8	0,000000
30	HDTV	13	0,08	13		25,8	0,000000
31	Devices	23	0,14	48	0,03	25,4	0,000000
32	Viruses	13	0,08	14	0,01	24,7	0,000001
33	Computers	15	0,09	20	0,01	24,6	0,000001
34	Of	436	2,74	5.090	3,69	40,1	0,000000
35	Energy	10	0,06	455	0,33	48,3	0,000000

Table 32 Technology Corpus Greek Keyword List

N	Word	Freq.	Techgr.lst %	Freq.	Sciamgr.lst %	Keyness	P
1	Ρομπότ	70	0,39	72	0,05	137,5	0,000000
2	ΑΥΚ	41	0,23	41	0,03	81,8	0,000000
3	Υπολογιστών	35	0,20	38	0,02	66,5	0,000000
4	Ευκρίνειας	28	0,16	33	0,02	50,4	0,000000
5	Σφαιρομπότ	25	0,14	25	0,02	49,9	0,000000
6	Μαχ	24	0,13	24	0,02	47,9	0,000000
7	Κινητήρα	30	0,17	42	0,03	47,8	0,000000
8	Τηλέφωνα	24	0,13	25	0,02	46,7	0,000000
9	Κινητών	21	0,12	21	0,01	41,9	0,000000
10	Λογισμικό	22	0,12	24	0,02	41,6	0,000000
11	Καύσης	24	0,13	32	0,02	39,7	0,000000

12	Τηλεόρασης	20	0,11	21	0,01	38,8	0,000000
13	Συσκευές	29	0,16	51	0,03	38,3	0,000000
14	Ψηφιακή	19	0,11	22	0,01	34,6	0,000000
15	Υψηλής	35	0,20	82	0,05	34,3	0,000000
16	Τηλεόραση	16	0,09	16	0,01	31,9	0,000000
17	Έξυπνα	17	0,10	19	0,01	31,7	0,000000
18	Ρομποτικής	16	0,09	17	0,01	30,8	0,000000
19	Αέρα	27	0,15	57	0,04	29,7	0,000000
20	Κανάλια	23	0,13	44	0,03	28,0	0,000000
21	Ψηφιακής	14	0,08	14		27,9	0,000000
22	Ιχνηλάτηση	14	0,08	14		27,9	0,000000
23	Κινητά	14	0,08	15		26,8	0,000000
24	Κακοβουλισμικό	13	0,07	13		25,9	0,000000
25	Υπολογιστές	15	0,08	20	0,01	24,8	0,000001
26	Αποκωδικοποιητής	12	0,07	12		23,9	0,000001
27	Λογισμικού	12	0,07	12		23,9	0,000001
28	Κακοβουλισμικού	12	0,07	12		23,9	0,000001

Scramjet comes 1st in the English list and *robots* comes 2nd; whereas *ρομπότ* (*robot(s))* and *ΑΥΚ* (the Greek acronym that stands for *αυλωθητής υπερηχητικής καύσης*; *scramjet*) occur in the 1st and 2nd position of the Greek list, respectively. Here, we have to mention that the Greek word *ρομπότ* stands for both the singular (*robot*) and the plural (*robots*) form, since it ends in –t (or to be extremely precise, a consonant which is not otherwise a regular final consonant in Greek nouns) and it has only one form for all cases in both numbers. Furthermore, both terms: *scramjet* and the plural and singular form of *robot* are perfectly matched in number to their Greek equivalents.

Third on the list is the word *mobile* which does not match in the list position with its Greek equivalent. This is because *mobile* is actually translated by two word forms (*κινητών* and *κινητά*) which correspond to two different cases of the same word.

The 4th English term is the word *computer*, which most probably corresponds to the Greek: *υπολογιστών* (plural; genitive) and *υπολογιστές* (plural; nominative/accusative). Nevertheless, we would like first to point out that *computer* is a singular form –which could probably be used as a determiner to a noun- and its two equivalents

υπολογιστών and *υπολογιστές* are both of them in plural. Second, in the 19th position of the English list, there is the well-known acronym *PC* which stands for *personal computer* (Chambers Dictionary 1999:843) and which has been standardized and nowadays used as such in many languages, including Greek; however, as we saw in the parallel concordances, the term *υπολογιστής* is also used to translate this acronym. Thus, we could say that we have two English words that correspond to one translation. This is partly right because while checking the parallel concordances, we observed that there is a tendency for *computer* to be translated as *υπολογιστής* and *PC* to be kept as such in the Greek text. Nevertheless, after a thorough observation of all occurrences of the form *PC* in the Greek texts, we concluded that this is a trait of a specific article and a certain technique of its translator. In the 5th position of the keyword list we have the word *malware*, a compound of MALicious softWARE, which has been smartly translated in Greek by the term *κακοβουλισμικό*. The latter appears in the list in two forms *κακοβουλισμικό* and *κακοβουλισμικού* (nominative/accusative and genitive respectively). Nonetheless, the total number of their occurrences is less than the number of occurrences of the English term *malware*. The reason for that will be revealed in the qualitative analysis.

In the 8th position we have the word *engine* which corresponds to the Greek *κινητήρα*. The arithmetic difference between them is explained by a range of compounds, such as *στροβιλοαντιδραστήρα* (equivalent of *jet engine*) that have been used in Greek, instead of the word κινητήρα (usual equivalent of *engine*). This has not been done arbitrarily but it can be attributed to the various collocations the word *engine* makes with its contextual neighbors.

Ninth in the list occurs the word *software* which corresponds to the Greek *λογισμικό* (nominative/accusative) and *λογισμικού* (genitive). The number of instances of both originals and translations does not differ dramatically; therefore an unproblematic matching is expected.

Finally the two acronyms, *DTV* (Digital TV) and *HDTV* (High Definition TV) which appear at the bottom of the English list bring about a problem concerning their matching to an equivalent on the opposite list. After a careful checking of the parallel concordances in Multiconc, it has been found that *DTV* has been mostly translated as *ψηφιακή τηλεόραση* (*digital TV*) in Greek, but in cases where *DTV* was collocating with words like *tuners* or *reception*, the word *τηλεόραση* from the Greek synecdoche

ψηφιακή τηλεόραση tended to disappear and the adjective preceding it, *ψηφιακή* was according in number and in case with the following noun.

The *HDTV*, on the other hand, is mostly translated by the collocation *τηλεόραση υψηλής ευκρίνειας* which however outnumbers its English equivalent. The reason for that is, as we observed in the parallel concordances, that *υψηλής ευκρίνειας* (equivalent only for high definition) can also stand for other notions besides *TV*, like *programs*, for instance. Nevertheless that is something that cannot be clearly seen at the stage of the keyword lists analysis.

Table 33 Multiconc Parallel Concordances

H:\multconc\techn.en P155 S1 THE **SCRAMJET** is not a new propulsion concept.

H:\multconc\techn.gr P155 Ο **ΑΥΛΩΘΗΤΗΣ ΥΠΕΡΗΧΗΤΙΚΗΣ ΚΑΥΣΗΣ** δεν αποτελεί καινούργια ιδέα στον τομέα της προώθησης.

H:\multconc\techn.en P50 S3 The goal was to see if it was possible to provide the same kind of common, low-level foundation for integrating hardware and software into **robot** designs that Microsoft BASIC provided for computer programmers.H:\multconc\techn.gr P50 Στόχος μας ήταν να δούμε αν μπορούσαμε να προσφέρουμε το ίδιο είδος κοινού, χαμηλού επιπέδου υπόβαθρο για την ενσωμάτωση υλισμικου και λογισμικού σε **ρομποτικές** σχεδιάσεις με αυτό που παρείχε η Microsoft BASIC στους προγραμματιστές των υπολογιστών.

H:\multconc\techn.en P119 S3 The target population for malicious **mobile software** is enormous and growing by leaps.
H:\multconc\techn.gr P119 Ο στοχευμένος πληθυσμός από κακόβουλο **λογισμικό κινητών** είναι τεράστιος και αυξάνει αλματωδώς.

H:\multconc\techn.en P43 S2 One trend that has helped them is the increasing availability of tremendous amounts of **computer** power.

H:\multconc\techn.gr P43 Μια τάση που τους έχει βοηθήσει είναι η συνεχώς αυξανόμενη διαθεσιμότητα τεράστιων ποσοτήτων **υπολογιστικής** ισχύος.

H:\multconc\techn.en P49 S5 Although a great many individuals made essential contributions to the development of the **personal computer**, Microsoft BASIC was one of the key catalysts for the software and hardware innovations that made the **PC** revolution possible.

H:\multconc\techn.gr P49 Αν και στην ανάπτυξη του **προσωπικού υπολογιστή** έχουν συμβάλει με ουσιαστικό τρόπο πάρα πολλοί άνθρωποι, η Microsoft BASIC αποτέλεσε έναν από τους βασικούς καταλύτες για τις καινοτομίες στο λογισμικό και το υλισμικό οι οποίες κατέστησαν δυνατή την επανάσταση των **προσωπικών υπολογιστών**.

H:\multconc\techn.en P115 S1 Despite Herculean efforts to rein it in, **PC malware** continues at a gallop: more than 200,000 forms have been identified so far, and today an unprotected **PC** is often infected within minutes of connecting to the Internet.
H:\multconc\techn.gr P115 Παρ' όλες τις ηράκλειες προσπάθειες να το χαλιναγωγήσουμε, το **κακοβουλισμικό για PC** συνεχίζει καλπάζοντας: περισσότερες από 200. μορφές έχουν ταυτοποιηθεί μέχρι στιγμής, και σήμερα ένα απροστάτευτο **PC** συχνά μολύνεται εντός ολίγων λεπτών αφότου συνδεθεί στο Διαδίκτυο.

H:\multconc\techn.en P148 S1 Creating a revolutionary **jet engine** that could propel a space plane to orbit affordably and routinely is a tough but seemingly achievable task

H:\multconc\techn.gr P148 Η δημιουργία ενός επαναστατικού **στροβιλοαντιδραστήρα** που θα μπορούσε να θέτει σε τροχιά ένα διαστημοπλάνο άνετα και οικονομικά είναι δύσκολο αλλά, καθώς φαίνεται, εφικτό σχέδιο

H:\multconc\techn.en P110 S2 I and other researchers who study malicious forms of **software** knew that it was only a matter of time until such malware appeared on mobile phones as well.

H:\multconc\techn.gr P110 Εγώ και άλλοι ερευνητές που μελετάμε κακόβουλες μορφές **λογισμικού** ξέραμε ότι ήταν θέμα χρόνου μέχρι να εμφανιστεί τέτοιο κακοβουλισμικό (malware) και στα κινητά τηλέφωνα.

H:\multconc\techn.en P22 S2 Or another option is that the companies may simply wait well into the next decade, when sufficient numbers of viewers will have finally replaced their long-lasting analog sets with ones containing **DTV tuners** as well as other so-called conditional-access systems, such as credit-card-size CableCARDs or their **software**-only counterparts.

H:\multconc\techn.gr P22 Μια άλλη επιλογή είναι οι εταιρείες να περιμένουν μέχρι την επόμενη δεκαετία, όταν ένας επαρκής αριθμός τηλεθεατών θα έχει πλέον αντικαταστήσει τις παλιές του αναλογικές συσκευές με καινούργιες οι οποίες θα περιλαμβάνουν **ψηφιακούς δέκτες** καθώς και άλλα συστήματα τύπου «ελεγχόμενης πρόσβασης στο περιεχόμενο», όπως αναγνώστες καρτών τύπου CableCARD ή αντίστοιχες λειτουργίες που υλοποιούνται μόνο με **λογισμικό**.

H:\multconc\techn.en P25 S3 But DBS faces its own bandwidth constraints as channels overall have proliferated, the number of network **HDTV** affiliates has swelled, and subscribers have increasingly had their local channels beamed to them by satellite.

H:\multconc\techn.gr P25 Ωστόσο, η DBS αντιμετωπίζει τους δικούς της περιορισμούς στο εύρος ζώνης, καθώς το σύνολο των καναλιών έχει αυξηθεί, ο αριθμός των θυγατρικών δικτύων που εκπέμπουν σε **υψηλή ευκρίνεια** έχει πολλαπλασιαστεί, ενώ όλο και περισσότεροι συνδρομητές λαμβάνουν το τοπικά τους κανάλια μέσω δορυφόρου.

4.7.3 Qualitative Analysis

Following the order of the keyword list and that of the quantitative analysis, we start with the candidate term *scramjet* and its Greek equivalent *ΑΥΚ*. Their appearance in clusters is not that important in terms of numbers (*scramjet engine* 5/41; *scramjet operation* 2/41; *scramjet performance* 2/41) (*λειτουργία του ΑΥΚ* 3/41; *επιδόσεων του ΑΥΚ* 2/41). Nonetheless, *scramjet* and its Greek three-word equivalent *αυλωθητής υπερηχητικής καύσης (ΑΥΚ)* are going to be included in the final candidate term list after they have been also verified by the dictionary.

The term *mobile* is an adjective and it usually appears in cluster with the noun *phone* to denote a gadget that came in our life approximately 17 years ago; however, for the sake of speech economy the word *phone* started to fade away. As a result, now *mobile* is used most of the times alone to refer to the multi-word term *mobile phone*. The same thing has happened in Greek. *Κινητό τηλέφωνο* is a multi-word term which has been also standardized in Greek as *κινητό*. As a matter of fact, in this corpus, *mobile* has been found in the following clusters: *mobile malware* (12/35), *mobile robots* (4/35) and *mobile virus(es)* (3/35). In Greek the equivalent clusters are: *κακοβουλισμικό κινητών* (7/21), *κινητών ρομπότ* (2/21) and *ιοί κινητών* (2/21). The fact that in the concordance lines of *malware* the word *malware* collocates 12 times out of 32 with *mobile*, reinforces our assumption that *mobile malware* constitutes a multi-word term.

In Greek corpus however, the equivalent cluster for *mobile malware* does not occur so many times (only 3 out of 13). A reason for that might be that the translator likes to "play around" with his options.

Finally, an interesting case are the two words *tracing* and *ray* which both appear in the English list, unlike their Greek equivalents, for which only *ιχνηλάτηση* is in the list as an equivalent of *tracing*. The fact is that *ray tracing* constitutes a multi-word term as their times of co-occurrence (20/26) leave no doubt. The question is whether these will appear on the list because of the absence of an equivalent of *ray* in the Greek list. The answer is that it will be included since at least one of the two terms appears in the keyword list.

4.7.4 Dictionary Verification

The dictionary includes mostly single-word terms like *robot* (Chambers Dictionary of Science and Technology 1999:995), *computer*, *engine, software* (ibid.) and *ρομπότ* (English-Greek, Greek-English Dictionary of Technology and Science 2001:1721), *υπολογιστής*, *κινητήρας* and *λογισμικό* (ibid.). However, multi-word terms, like *scramjet engine* or *coaxial cable* and *κακοβουλισμικό κινητών* (equivalent to *mobile malware*) or *αυλωθητής υπερηχητικής κάυσης* (*AYK*) (equivalent to scramjet) are hard to be found in a dictionary.

In addition, there is also the issue of the standardized and non-standardized terms, such as *scramjet* and *ballbots*, which do not appear in a technical dictionary (at least in none of the ones we have at our disposal).

4.7.5 Summary–Remarks

At this point, we would like to express our conviction that small corpora (Maia 1997, Zanettin 1998) provide us with greater flexibility, in that we can easily become familiar with details. Numbers are only part of the understanding of the language, but their role is important in showing what is typical in it. However, numbers alone are unable to shape language's complete picture. Human interpretation of data explains what causes the numbers to be the expected or unexpected way they are.

Table 34 Single-word candidate terms

1. **scramjet**	1. **ρομπότ**
2. **robots**	2. **ΑΥΚ**
3. **mobile**	3. **υπολογιστών**
4. **computer**	4. **ευκρίνειας**
5. **malware**	5. **σφαιρομπότ**
6. **mach**	6. **MAX**
7. **robot**	7. **κινητήρα**
8. **engine**	8. **τηλέφωνα**
9. **software**	9. **κινητών**
10. **air**	10. **λογισμικό**
11. **cable**	11. **καύσης**
12. **robotics**	12. **τηλεόρασης**
13. **flight**	13. **συσκευές**
14. **tv**	14. **τηλεόραση**
15. **phones**	15. **ρομποτικής**
16. **channels**	16. **αέρα**
17. **PC**	17. **κανάλια**
18. **tracing**	18. **ψηφιακής**
19. **ball**	19. **ιχνηλάτηση**
20. **smartphones**	20. **κινητά**
21. **DTV**	21. **κακοβουλισμικό**
22. **ballbot**	22. **υπολογιστές**
23. **program**	23. **αποκωδικοποιητές**
24. **phone**	24. **λογισμικού**
25. **ray**	25. **κακοβουλισμικού**
26. **HYTECH**	
27. **HDTV**	
28. **devices**	
29. **viruses**	
30. **computers**	

Table 35 Multi-word candidate terms

1. scramjet engine	**1. προσωπικών υπολογιστών**
2. mobile robots	**2. υψηλής ευκρίνειας**
3. mobile malware	**3. τηλεόραση υψηλής ευκρίνειας**
4. mobile phones	**4. κανάλια υψηλής ευκρίνειας**
5. cable system	**5. προγράμματα υψηλής ευκρίνειας**
6. cable operators	**6. εκπομπή υψηλής ευκρίνειας**
7. coaxial cable	**7. υπολογιστής του σφαιρορομπότ**
8. robotics industry	**8. έξυπνα τηλέφωνα**
9. analog TV	**9. κινητά τηλέφωνα**
10. digital TV	**10. κακοβουλισμικό κινητών**
11. TV channels	**11. κακόβουλο λογισμικό**
12. TV sets	**12. λογισμικό υποκλοπής**
13. cell phones	**13. Αυλωθητής Υπερηχητικής Κάυσης (ΑΥΚ)**
14. analog channels	**14. Θάλαμος Υπερηχητικής Κάυσης**
15. high-definition channels	**15. ψηφιακή τηλεόραση**
16. DTV tuners	**16. καλωδιακή τηλεόραση**
17. HDTV programs	**17. δορυφορική τηλεόραση**
18. ray tracing	**18. αναλογική τηλεόραση**
19. computer viruses	**19. βιομηχανία της Ρομποτικής**
20. PC viruses	**20. ροή του αέρα**
21. mobile viruses	
22. ball rotation	
23. PC malware	
24. Hytech program	

5. CONCLUSION

5.1 Commenting on the final term lists

Our last section will start with comments on the lists of single-word and multi-word terms, we retrieved from the analysis of every sub-corpus. This will be done here because the reader may feel the need to understand why we ended up with these lists and why we chose to include in the lists these terms and not others. The main reason why we did not provide an explanation for every single list earlier is because our interest was mainly focused on the method and not on its results. Consequently, we preferred to draw general conclusions on the final term lists.

For every sub-corpus we provided a detailed analysis of the procedure we followed to extract and finally comprise in a list the candidate terms. Furthermore, the summary–remarks section, in the end of the analysis of every sub-corpus constituted a kind of conclusion which summarized all noteworthy points of every sub-corpus.

There are two kinds of lists: the single-word term lists and the multi-word term lists for every sub-corpus in both English and Greek. No matching between languages was attempted for the emerging terms, because such an act would be out of the scope of this study which is to provide translation teachers and students with a method ready to be applied and the issue of the choice of terms which are going to be used in a translation course is completely independent and up to the people involved in the teaching procedure.

The extraction of the single-word terms are almost exclusively based on the initial keyword lists we retrieved from Wordsmith 3.0. However, the decision on which terms to include in the final lists was shaped according to the criteria we set at the beginning of our study.

Table 36 Indicative examples of equivalent multi-word terms

Biol/Anthr	mRNA transcript	μετάγραφο mRNA
En/Env/Geol	greenhouse gas	αέρια θερμοκηπίου
Medicine	cancer cells	καρκινικά κύτταρα
Physics	laser light	φως λέιζερ
Plan/Cosm	irregular moons	ανώμαλοι δορυφόροι
Psychology	mirror neuron	νευρώνων–κατόπτρων
Technology	mobile malware	κακοβουλισμικό κινητών

The compilation of multi-word term lists was a more complex procedure. For the extraction of multi-word terms we took as starting point the single-word terms and we checked in the concordance lines for any fixed and repeated collocations of them which could lead us to the assumption that they constitute multi-word terms. The appearance of these terms in the final list was strongly linked firstly to the 10%, and above, frequency of co-occurrence of the components of a multi-word term and then to their existence or not in technical dictionaries or online glossaries.

What would be useful for the term lists that we did not include in this study, but we regard as something important is the rating of the degrees of technicality of the terms. We envisage it in the future as an important technique which could be based either on the level of technicality of the contextual elements of a single-word term or on the level of technicality of the single-word components of a multi-word term.

Table 37 Indicative examples of different levels of terms' technicality

Biol/Anthr	*water* VS *pseudogenes*	*νερού* VS *ψευδογονιδίων*
En/Env/Geol	*countries* VS *hydrogen*	*χώρες* VS *υδρογόνο*
Medicine	*drugs* VS *autoantibodies*	*φάρμακα* VS *αυτοαντισώματα*
Physics	not obvious in this sub-corpus	not obvious in this sub-corpus
Plan/Cosm	*stars formation* VS *dark matter*	*σχηματισμός άστρων* VS *σκοτεινή ύλη*
Psychology	*chess players* VS *motor cortex*	*μετρ του σκακιού VS κινητικό φλοιό*

Technology	not obvious in this sub-corpus	not obvious in this sub-corpus

5.2 Results and research problems

In this paper, we attempted the development of a methodology for the extraction of terms from parallel corpora. More precisely, we used a corpus-based approach to fish candidate single and multi-word terms out of a specially created parallel English–Greek popular science corpus. The aim of this study was to show a way to translation teachers and trainees of how to make good use of facilities to which they can easily get access, such as word processors, texts of wide circulation (e.g. from popular science magazines like Scientific American) and their translations, or maybe even students' own translations of these texts, as well as simple corpus tools (facilities for bilingual concordancing, like Multiconc).

Overall, we could characterize our analysis as horizontal but multi-leveled, because even though it examined seven sub-corpora, it can be regarded as detailed in that it used three stages of analysis: the quantitative, the qualitative and the dictionary verification.

More precisely, in our study, we came across the following cases:

- the issue of unequal keyword lists: this is due to the fact that the initial lists are not lemmatized. This means that the more inflected a language is, the less likely the rare forms of a word can appear in the keyword list, and therefore the number of occurrences of a term in –let us say– "top ten" or "top twenty" of the list is likely to be lower in Greek than in English
- the issue of inflectional languages in translation: Greek is more inflectional than English. That is to say, in the English corpus we had to cope only with two different forms of singular and plural, whereas in Greek we had to cope

with the different numbers and the different cases of the noun terms. This had as a consequence a problematic matching of equivalents between source and target text which has been reflected also in the term lists.

- the issue of syntactic patterns matching: in our criteria we committed ourselves to extracting mainly nouns. However, most of the times, the matching of noun equivalents between languages proved to be a trivial matter; especially when nouns were substituted by adjectives, determiners or paraphrases.
- the issue of acronyms and single letters: For the former we made the decision to include them in the term lists, whereas for the latter we decided not to include them in the single-word lists, but only in the multi-word lists when these happened to function as components of a multi-word term.
- the issue of validity of technical dictionaries: In this study we decided to use technical dictionaries and not term banks because the access to the former was easier than the access to the latter. From our research, we remarked, on the one hand, the fairly poor ability of technical dictionaries to verify all nouns and noun phrases we suggested as terms. On the other hand, we realized the growing importance of these tools to the extraction and standardization of terminology.

5.3 Applications and future work

As we previously mentioned the present study introduces a method which can be potentially used in the translation classroom. Its function is mainly assistive to the teaching procedure. Supposing students have access to the software and to a personal computer, they can create easily small parallel corpora by gathering the teaching material (the source texts with which their teachers supply them) and their own

translations. Then, by using Multiconc, they can align their parallel corpus and with Wordsmith they can get frequency lists. Keyword lists are a simple way to retrieve terms because they show what is representative in a corpus; hence these lists provide interesting data, when the articles that consist a corpus are technical. Wordsmith is also important because it gives the most significant collocates of a node word, only by clicking on the word the students are interested in examining. Wordsmith, with its Viewer & Aligner Tool can show the parallel texts in a sentence-to-sentence form, but Multiconc is recommended for sentence or paragraph matching because it has the alignment tool. The advantage of such parallel corpora is that they can be used as repertories of natural language and be enriched any time by both translation teachers and students.

The students can also face some problems during the corpus compilation process. One issue is the quality of the translations, if they are going to use as a translated text material their piece of work. Another issue is the accessibility to language material. For instance, for our research, we got access to the source language material electronically through our university subscription. For the translations, we could not have access to the electronic issues of Greek Scientific American, thus, we photocopied and scanned the most recent six-month issues that existed in the University Library. Of course, we realize that no student would like to undertake such a laborious task; hence we assume an ideal –but, hopefully not far from the truth– situation, where translation teachers provide students with the source language material and the student's translations are accurate.

Additionally, translation teacher's contribution to the compilation and the computerized analysis of a parallel corpus could be seen as a helpful necessity and not as an extra burden to their already busy schedule. After all, our approach is aiming to

teachers who are motivated, interested in new ideas and willing to spend some of their time in developing a promising teaching method. As we said above, such corpora can be used as repertories and their analysis can also be conducted –after the alignment has been completed– by one software program (e.g. Wordsmith). An important issue upon which we must draw teachers' attention is the maintenance of the corpus consistency while this is enriched with new material.

All in all, we also presuppose the understanding by the students of notions like keyness, frequency lists, which will help them to undertake such a task. We also understand that the applicability of such a method to big and ambitious projects is restricted, but we regarded it as sufficient and easy to use by students.

For Greek language, however, we are restricted to working with limited means. In the future, if our technique is to be used for Greek in a broader scale, we envisage a range of techniques that can be applied to English-Greek parallel corpora, such as greater availability of accurate and efficient taggers for Greek, syntactic and structural analysis of the text at the level of chunks and phrases, tokenization, development of techniques for the alignment of exact translation equivalents of multi-word terms, statistical alignment techniques which prioritize the one-to-one correspondence and calculation of scores and filters for the alignment and the matching of equivalents. But above all, we wish to see a greater interest on the part of translation departments working from and into Greek.

BIBLIOGRAPHY

Aarts, J. (1999). '*The description of Language Use'* In Hasselgård, Hilde and Signe Oksefjell (Eds.) *Out of corpora. Studies in honour of Stig Johansson*, Amsterdam/Atlanta, GA: Rodopi. 1999

Biber, D., **Conrad**, S. & **Reppen**, R. (1998). '*Corpus Linguistics: Investigating Language Structure and Use*', Cambridge: Cambridge University Press.

Baker, M. (1993). '*Corpus Linguistics and translation studies: implications and applications'* In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In honour of John Sinclair*. Amsterdam: John Benjamins 17-45.

Baker, M. (1995). '*Corpora in Translation Studies: An Overview and Some Suggestions for Future Research*', Target*,* 7 (2), 223-243.

Baker, M. (1996b). '*Corpus-based Translation Studies: The Challenges that Lie Ahead*', in Harold Somers (ed) *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, Amsterdam and Philadelphia: John Benjamins: 175-186

Baker, M (1998). '*Norms*'. In Mona Baker '*Encyclopedia of Translation Studies*', London and New York: Routledge, 163-165.

Baker, M (1998). '*Translation Studies*'. In Mona Baker '*Encyclopedia of Translation Studies*', London and New York: Routledge, 277-280

Baker, M (1998). '*Réexplorer la langue de la traduction: une approche par corpus*', Meta, XLIII, 4, 1998

Baker, M. (1999). '*The Role of Corpora in Investigating the Linguistic Behaviour of Professional Translators*', International Journal of Corpus Linguistics 4 (2): 281-298

Baker, M. (2004). '*The Treatment of Variation in Corpus-based Translation Studies*', in Karin Aijmer and Hilde Hasselgard (eds) *Translation and Corpora* (Gotenborg Studies in English), Gothenburg University, Sweden, 7-17.

Baker, M. (2004). '*A Corpus-based View of Similarity and Difference in Translation*', International Journal of Corpus Linguistics 9(2): 167:193.

Bononno, R (2000). '*Terminology for Translators— an Implementation of ISO 12620*', META XLV (4): 646-669.

Cabré, M. T. (1999). '*Terminology: Theory, methods and applications*' Amsterdam and Philadelphia: John Benjamins B.V.

Cabré Castellvi, M. T.; **Bagot**, R. E.; **Palatresi**, J. V. (2001). '*Automatic Term Detection: A review of current systems*', In Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme, *Recent Advances in Computational Terminology*, John Benjamins Publishing Company, 53-87.

Chung, T.M. and **Nation P**. (2004) '*Identifying Technical Vocabulary*', System 32 (2), 251-263

Church, K. W.; Gale, W. (1991). '*Identifying Word Correspondences in Parallel Texts*', In the 4th DARPA Workshop on Speech and Natural Language, 152-157

Cmerjek, M.; Curin, J. (2001). '*Automatic Extraction of Terminological Translation Lexicon from Czech-English Parallel Texts*', International Journal of Corpus Linguistics Vol. 6 (special issue), John Benjamin Publishing Co., 1-12

van der Eijk, P. (1993). '*Automating the Acquisition of Bilingual Terminology*', In the proceedings of EACL'93. 113-119

Erjavec, T. (2002). '*The IJS-ELAN Slovene-English parallel corpus*', International Journal of Corpus Linguistics Vol.7 (1), John Benjamin Publishing Co., 1-20

Gale, W. A.; Church, K. W. (1993). '*A program for aligning sentences in bilingual corpora*.' In the proceedings of the 29th annual meeting on Association for Computational Linguistics, June 18-21, 1991, Berkeley, California, 177-184

Hatim, B. (2001) '*Teaching and Researching Translation*', Essex: Pearson Education Limited

Heid, U. (1999). '*Extracting Terminologically Relevant Collocations from German Technical Texts*', In: Sandrini, Peter (ed.): *Terminology and Knowledge Engineering* (TKE '99), Innsbruck. Wien: TermNet, 241-255

Holmes, J. (1988b/2000). '*The Name and Nature of Translation Studies*' in Lawrence Venuti *The Translation Studies Reader*, New York and London: Routledge, 180-192

Hunston, S. (2002). '*Corpora in Applied Linguistics*', Cambridge: Cambridge University Press.

Johansson, S. (2003). '*Reflections on Corpora and their Uses in Cross-linguistic Research*', in Zanettin, Bernardini and Stewart '*Corpora in Translation Education*' Manchester: StJerome: 135-143

Kennedy, G. (1998). '*An Introduction to Corpus Linguistics*', London and New York: Longman.

Kenny, D. (1998). '*Corpora in Translation Studies*'. In Mona Baker '*Encyclopedia of Translation Studies*', London and New York: Routledge, 50-53

Kenny, D. (2000) 'Translators at play: exploitations of collocational norms in German-English translation.' in B. Dodd (ed.) (2000) *Working with German Corpora.* Birmingham: Univeristy of Birmingham Press. 143-160

Kenny, D. (2001). '*Lexis and Creativity in Translation: A Corpus-based Study*', Manchester and Northampton: St. Jerome Publishing

Kubler, N. (2003). '*Corpora and LSP Translation*', in Zanettin, Bernardini and Stewart '*Corpora in Translation Education*' Manchester: St Jerome: 25-41

Laviosa, S. (1998). '*Universals of translation*', In Mona Baker '*Encyclopedia of Translation Studies*', London and New York: Routledge, 288-291.

Laviosa, S. (1998). '*The Corpus-based Approach: a new paradigm in Translation Studies*', Meta XLIII, 4, electronic version

Lemay C.; **L'Homme**, M. C.; **Drouin** P., (2005) '*Two methods for extracting "specific" single-word terms from specialized corpora: Experimentation and Evaluation*', *International Journal of Corpus Linguistics*, 10 (2), 227-253.

Maia, B. (1997) '*Do-it-yourself corpora... with a little bit of help from your friends*' In the Proceedings of PALC '97, 403-409

Maia, B**.** 2002. '*Corpora for terminology extraction – the differing perspectives and objectives of researchers, teachers and language services providers*', In the proceedings of the Workshop Language Resources for Translation Work and Research - held in conjunction with LREC 2002, Universidad de Las Canárias, Spain.. 25-28.

Maia, B. (2003). '*Some Languages are more Equal than Others: Training Translators in Terminology and Information Retrieval using Comparable and Parallel Corpora*' in Zanettin, Bernardini and Stewart '*Corpora in Translation Education*' Manchester: StJerome: 43-53

McEnery, T. & **Wilson** A. (1996). '*Corpus Linguistics'*, Edinburgh: Edinburgh University Press.

Munday, J. (2001). '*Introducing Translation Studies: Theories and Applications*', London and New York: Routledge.

Oakes, M. P. (1998). '*Statistics for Corpus Linguistics*', Edinburgh: Edinburgh University Press

Olohan, M. (2004). '*Introducing Corpora in Translation Studies*', London: Routledge.

Olohan, M. and **Baker**, M. (2000) '*Reporting that in translated English: evidence for subconscious processes of explicitation?*" Across Languages and Cultures 1:141-172

Patry, A.; **Langlais** P. (2005). '*Corpus-Based Terminology Extraction*', In *Proceeding of the 7th International Conference on Terminology and Language Engineering*, Copenhagen, Danemark, August 17-18, 313-321.

Pearson, (2003). J., '*Using parallel texts in the translator training environment*' in Zanettin, Bernardini and Stewart '*Corpora in Translation Education*' Manchester: StJerome:15-24.

Penas, A**.; Verdejo** F.; **Gonzalo** J. (2001) '*Corpus-based Terminology Extraction applied to Information Systems*' In Proceedings of Corpus Linguistics 2001

Sager, J. C. (1990/1996) '*A practical course in Terminology Processing*', Amsterdam, Philadelphia: John Benjamins B. V.

Sager, J. C. (1994) '*Language Engineering and Translation Consequences of Automation*', Amsterdam, Philadelphia: John Benjamins B. V.

Sinclair, J. (1991). '*Corpus, Concordance, Collocation*', Oxford: Oxford University Press.

Storjohann, P. (2005) '*Corpus-driven vs. corpus-based approach to the study of relational patterns*', Proceedings from The Corpus Linguistics Conference Series, Vol. 1, no. 1 Corpus Linguistics 2005, http://www.corpus.bham.ac.uk/PCLC/

Stubbs, M. (1996). '*Text and Corpus Analysis*', Oxford and Massachusetts: Blackwell Publishers Ltd.

Stubbs, M. **(**2001). '*Words and Phrases. Corpus Studies of Lexical Semantics*', Oxford and Massachusetts: Blackwell Publishers Ltd.

Teubert, W. (2002) '*The role of parallel corpora in translation and multilingual lexicography*'. In B. Altenberg, S. Granger (eds.) *Lexis in Contrast.* Amsterdam/Philadelphia: Benjamins, 189-214

Tiedemann, J. (1998). '*Extraction of translation equivalents from parallel corpora*', In the proceedings of the 11th Nordic Conference on Computational Linguistics, Center for Sprogteknologi, Copenhagen, 1998, 1-7

Tognini-Bonelli, E. (2001). '*Corpus Linguistics at Work*', Amsterdam/Philadelphia: John Benjamins Publishing Company

Toury, G. (1978). '*The Nature and Role of Norms in Literary Translation*' in Gideon Toury '*Literature and Translation*', Leuven: ACCO

Toury, G. (1995). '*Descriptive Translation Studies and Beyond*', Tel Aviv University: Benjamin Translation Library 4

Utka, A. (2004) '*Phases of Translation: compilation and analysis*.' *International Journal of Corpus Linguistics* 9/2:195-224

Vanderauwera, R. (1985) '*Dutch Novels translated into English: The Transformation of a 'Minority' Literature*', In M. Baker G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: In honour of John Sinclair*. Amsterdam: John Benjamins 17-45.

Varantola, K. (2003). '*Translators and Disposable Corpora*', in Zanettin, Bernardini and Stewart '*Corpora in Translation Education*' Manchester: StJerome: 55-70.

Venuti, L. (2000, 2004). '*Translation, Community, Utopia*' in Lawrence Venuti *The Translation Studies Reader*, New York and London: Routledge, 482-502.

Vintar, S. (2001). '*Using Corpora for Translation-Oriented Term Extraction*' *Babel* Vol. 47 (2), 121-132

Veronis, J. (2000). '*Chapter 1 From the Rosetta stone to the information society A survey of parallel text processing*', In the *Parallel text processing: alignment and use of translation corpora*, Kluwer Academic Publishers. Universite de Provence, France, 2000, 1-24

Varadi, T. (2000). '*Fishing for Translation Equivalents Using Grammatical Anchors*', *International Journal of Corpus Linguistics* Vol. 5(1), 2000, 1-16

Zanettin, F. (1998) '*Bilingual Comparable Corpora and the training of translators*' Meta, XLIII (4) http://www.erudit.org/revue/meta/1998/v43/n4/004638ar.html

Dictionaries used in this study

Chambers (1999) '*Dictionary of Science and Technology*', Edinburgh: Chambers Harrap Publishers Ltd

Dorland's (1997) '*Medical Dictionary: English-Greek Greek-English*', Athens: Medical Publications M.Ch. Paschalidis

Hornby, A.S. (1995) '*Oxford Advanced Learner's Dictionary of Current English'*, Oxford: Oxford University Press.

Pantazis, N. (2002) '*New English-Greek Dictionary of Technical Terms*', Athens: Kastaniotis Publishing

Siemens, '*Dictionary of Technology and Sciences:English-Greek Greek-English*' Athens: Stafylidis Publishing

Software web resources

Multiconc: http://artsweb.bham.ac.uk/pking/multiconc/lingua.htm

Wordsmith 3.0: http://www.lexically.net/wordsmith/version3/manual.pdf

Terms verification websites

For the Biology/Anthropology sub-corpus:

riboswitches: http://www.yale.edu/breaker/riboswitch.htm

cancer cells : http://www.cancerhelp.org.uk/help/default.asp?page=96

beta cells: http://www.diabetesnet.com/betacl.php

t-reg: http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/T/Treg.html

colon cancer: http://www.medicinenet.com/colon_cancer/article.htm

immune system cells: http://www.aidsmap.com/en/docs/F7193E5D-540C-4843-BE2A-1B21D42FC5EA.asp

tumor suppressor protein: http://cti.itc.virginia.edu/~cmg/Demo/pdb/p53/p53.html

tumor cells: http://www.iscid.org/encyclopedia/Tumor_Cell

visible light:

http://www.windows.ucar.edu/tour/link=/physical_science/magnetism/em_visible_light.html

For the Physics Sub-corpus:

supercontinuum : http://www.rp-photonics.com/supercontinuum_generation.html

superconducting detectors: http://www.rp-photonics.com/photon_counting.html

silicon laser: http://www.rp-photonics.com/silicon_photonics.html

malware: http://www.wetstonetech.com/page/page/1972572.htm