ROYAL INSTITUTE
OF TECHNOLOGY

Master's Thesis

Using parallel corpora to create Greek-English dictionary for web site searching

November 2006

Department of Computer and Systems Sciences
Royal Institute of Technology / Stockholm University

Author: Konstantinos Charitakis

Supervisor: Associate professor Hercules Dalianis

* This thesis corresponds to 20 weeks of full-time work

Abstract

The importance of parallel corpora has been appreciated for many years. After the emergence of statistical translation methods there were many projects that have been working on automated dictionary extraction using parallel corpora. Many corpora processing systems and tools have been implemented and have been applied to parallel corpora of most of the popular natural languages. However there are not many projects on automated creation of a dictionary between the Greek and English language pair.

This thesis project focuses on the creation of a machine readable bilingual dictionary from Greek-English parallel corpora that were created manually by collected documents retrieved from the Internet. The English corpora contained 196.048 words in total, with 10.450 unique words identified, while the Greek corpora contained 204.043 words in total, with 18.117 unique words identified respectively. The parallel corpora processing was performed by the Uplug system without the use of language specific information. A sample was extracted from the population of suggested translations included in the resulted dictionary, and was included in questionnaires that were sent out to Greek-English speakers who evaluated the sample based on the quality of the translation pairs. For the suggested translation pairs of the sample belonging to the stratum with the higher frequency of occurrence, 67.11% of correct translations have been achieved. With an overall of 50,63% correct translations of the sample, the results were promising considering the minimal optimisation of the corpus and the many differences between the two languages.

The resulted dictionary could be used as input to special software tools that in their turn could be used by search engines for web site searching, or it can be utilised by Multilingual Information Retrieval applications in order to facilitate web retrieval and act as a bridge between different languages. The dictionary can also be used as a translation tool between Greek and other small languages with English acting as a pivot language.

Acknowledgements

I would like to thank my friends and especially my family for their confidence and moral support throughout my studies.

I would also like to thank my supervisor associate professor Hercules Dalianis for his support, guidance and enthusiastic supervision throughout the project.

Special thanks also to Martin Rimka for his help in the practical part of the project and for the enlightening conversations in the beginning of the project.

Finally I would like to thank all the people who put the time and effort in filling the rather long questionnaire and helped with the evaluation of the extracted dictionary.

Table of contents

List of figures

List of tables

1. Introduction

Greece is a small country and the native Greek speakers worldwide are roughly 15 millions [The General Secretariat for Greeks Abroad, 2006] of which approximately 2.7 millions constitute the Greek speaking online population [Global Reach, 2004].

In order for Greek resources to be available but above all understandable and therefore appreciated by the global online community, there must be found a way for all these resources to be available to non Greek speaking users. For instance, to researchers seeking to determine who has conducted research on a particular topic, intelligence agencies seeking global intelligence, companies seeking international business communications and opportunities and so on.

In many cases, organisations or individuals have included another language version to their web sites in order to allow Greek resources to be visible into other languages and facilitate non Greek speaking users, but this seems to be a transitory solution. Apparently this also applies the other way around with Greek speaking users seeking resources in other languages.

The World Wide Web has become a major channel for information service. However the majority of the web content is in English. Nearly 70% of the web content is in English while paradoxically only 35.8% of the online population are native English speakers [Global Reach, 2004]. Therefore the following question arises. How is that affecting the effectiveness and the power or even the availability of the resources over the web? The question comes right in time if someone considers the importance and the rapid evolvement of the web in all fields of human life, including not only research and education but also marketing and trade as well as entertainment. The web has become undoubtedly the biggest resource of information. This implies that it becomes more and more important to know how to use Internet services and, as a part of this, to read and write in English.

Nowadays research is turned towards incorporating Multilingual Information Retrieval (MLIR) applications in web retrieval and web site searching taking advantage of the vast amount of information over the web [Zhou, Qin, Chen and Nunamaker, 2005]. In the field of Multilingual Information Retrieval the most common approach for searching across languages makes use of translation of the search query to all target languages. This is performed by means of bilingual dictionaries in the source and target languages. Therefore in order to facilitate multilingual web search and improve its efficiency and performance, the existence of bilingual dictionaries between the source language and the target languages are valuable tools of great importance.

Due to the diversity of the known languages and the vast amount of resources required to produce a bilingual dictionary, people turned their efforts towards the automation of the task. Within the field of machine translation the emergence of statistical methods have shown promising results and combined with machine learning techniques, they have given results accurate enough, with less effort and resources required that could be used in order to facilitate the task of automated dictionary extraction [Brown et al., 1990][Forsberg, 2005]. Parallel corpora, which are texts aligned together with their translation in one or more languages, are

extensively used in statistical translation methods as they contain a vast amount of bilingual lexical information.

This thesis is focused on the extraction and evaluation of a Greek-English dictionary created from bilingual parallel corpora using the Uplug system.

Uplug origins from a project in Uppsala University and provides a collection of tools for linguistic corpus processing, word alignment and term extraction from parallel corpora using only statistical methods, without use of linguistic information [Tiedemann, 2004b].

1.1 Problem

As it is going to be described later in the thesis machine readable bilingual dictionaries are used as translation means in search query translation, in order to facilitate web retrieval and web site searching. However the creation of bilingual dictionaries is a tedious task which requires a lot of resources. The automation of bilingual dictionary extraction without the use of linguistic information is also a difficult task especially in the case of the Greek language which has many differences from other languages.

A similar work of extraction of Greek English dictionary was attempted by Piperidis, Boutsis and Demiros [1997] although the approach was slightly different as it employed statistical techniques coupled with linguistic processing for better results and it was applied on a small corpus in software domain.

If there is another attempt which didn't come to our attention it would be very interesting but also very useful, to compare the results and derive more accurate conclusions.

1.2 Goal of the project

The goal of this thesis project is to use parallel corpora in order to create a bilingual Greek-English dictionary using the Uplug system without the use of linguistic information.

1.3 Purpose of the project

The implementation of a Greek-English dictionary will contribute towards the efforts of many other contributors, for an easier and more accurate query search within resources over the Internet from Greek to many other target languages and vice versa.

The resulted dictionary could be used as input to special software tools that in their turn could be used by search engines for web site searching, or it can be utilised by Multilingual Information Retrieval applications in order to facilitate web retrieval and act as a bridge between different languages.

The dictionary can also be used as a translation tool between Greek and other small languages with English acting as a pivot language. For example the Greek-English

dictionary produced could be combined with an English-Swedish dictionary for Greek-Swedish translations.

1.4 Method

The thesis work involves the extraction and evaluation of a bilingual Greek-English dictionary using parallel corpora.

The extraction of the dictionary involves the laboration with the Uplug system using parallel corpora of a specific domain.

The methods used for the evaluation of the resulted dictionary encompass the use of the stratified[1] sampling method for the extraction of sample data and the use of questionnaires for the evaluation of the sample data by Greek-English speaking persons who evaluated the results based on the quality of the translations.

1.5 Limitations

A corpus could never be considered big enough to fully represent a language. Due to time constraints the laboration involves only a respectful size of parallel corpora from a certain domain that was considered efficient enough to create a bilingual dictionary using Uplug's full potential in corpus processing.

There are many approaches followed for the extraction of word equivalences that comprise bilingual dictionaries. The report will focus on the most important ones, the ones that are most commonly used and are more relative to the system that is going to be used.

1.6 Overview of thesis

The thesis report is consisted of the following chapters:

Chapter 1 - Introduction
In this chapter a brief introduction to the subject area, a description of the problem definition, the aims and objectives as well as an overview of the thesis report are included.

Chapter 2 - Background
This chapter gives a thorough introduction to different concepts of automated dictionary extraction. It includes an introduction to linguistic corpora with a focus on parallel corpora and their importance in language engineering. Then follows a description of the differences between the language pair concerned (Greek and English) followed by a description of corpora processing and the approaches used for the creation of a bilingual dictionary from parallel corpora. After that a description of the methods used for the evaluation of the alignment systems is included.

[1] A stratum is a subset of the population that share at least one common characteristic.

The last part of the chapter describes the system that is used for the creation of the bilingual dictionary. An introduction of the Uplug system is included together with a short description of the main approach behind the individual tools that are parts of the system and are used for the parallel corpora processing. After that a description of the trends in web site searching is described.

Chapter 3 – Methodology
This chapter includes a description of all stages involved in the extraction and evaluation of the bilingual dictionary in this particular thesis project.

Chapter 4 - Evaluation
This chapter includes a description of the resulted suggested translations after the corpora processing, followed by an analysis of the results of the evaluation of the extracted sample.

Chapter 5 - Conclusions
In this chapter final conclusions about the project as a whole and the evaluation of the results in particular are presented, followed by suggestions of possible future work that could be done.

Finally this chapter includes a critical appraisal on the completeness of the goals set in chapter one.

2. Background

This chapter gives a thorough introduction to different concepts and techniques of automated dictionary extraction. It includes an introduction to linguistic corpora with special focus on parallel corpora and their applications followed by a brief introduction of the characteristics and differences of the Greek and English languages. After that, there is an introduction to statistical translation approaches followed by the main approaches that are used in automated extraction of translation equivalents and the main evaluation methods of alignment systems. Following that a description of the Uplug system which is used for the creation of the bilingual dictionary is included together with a short description of the main approach behind its individual tools. Finally this chapter includes a brief introduction to web searching across languages and the later trends of research on this subject

2.1 Linguistic corpora

A valuable tool in the fields of machine translation and multilingual web retrieval are linguistic corpora. Apart from their use in extraction of information on natural languages they can be utilised by statistical translation methods for automated bilingual dictionary extraction.

The term “corpora” is the plural of the word “corpus” which comes from the Latin language and means “body”. In modern Linguistics the term is used to refer to large collections of texts, in electronic form, selected to represent as more as possible a language or a variety of languages for the purpose of linguistic research [Sinclair, 2004]. If the collection of texts contains documents in more than one language it is referred to as multilingual corpora. Multilingual corpora in their turn are divided in two main categories: comparable and parallel corpora.

A collection of texts in different languages but not translations of each other, within the same main topic and similar in content, is called comparable corpora. Comparable corpora are used to compare different languages in similar circumstances of communication. An example of comparable corpora would be a collection of news articles in different languages but on the same topic.

Parallel corpora are texts in some source language aligned together with their translation in one or more other target languages. Parallel corpora hold a huge amount of linguistic information and this is the reason why they have many applications in the field of natural language processing. The type of corpora that is going to be used in this thesis project is parallel corpora.

2.1.1 Parallel corpora

Parallel corpora are great tools in the hands of researchers working with machine translation and applied statistical methods. They have become an important resource for building natural language processing tools that become more and more necessary, mainly because of the diversity of available languages in the rapidly evolved information society.

Parallel corpora have been used since ancient times. The most famous example of parallel corpora is the Rosetta stone [The British Museum, 2000]. It is a stone incised with the same text in two Egyptian language scripts and in one classical Greek. With the use of comparative translation[2] applied on these three different versions of the same text, scientists managed to translate its content but also the content of many other previously untranslatable scripts of hieroglyphic writing.

Parallel corpora can be found in two main formats. They can be raw parallel texts, which are useful for simple investigations of different languages or they can be aligned texts, also called *bitexts* by many scientists. The alignment of the translated texts most commonly is done in sentence level but it can also be done in paragraph level or word level, or even in smaller level by a given number of characters.

Despite the fact that processing of parallel corpora was used early in the 1950s as one of the first non-numerical applications of computers mainly for information retrieval for military purpose between USA and Russia, it was not until the 1980s that parallel texts were used systematically in order to process natural languages [Hutchins, 2000].

2.1.2 Applications of parallel corpora

Parallel corpora are turned out to be a powerful tool in the hands of scientists, translators and linguists. For the last two decades researchers in the field of natural language processing and the general applied linguistics have been working with parallel corpora. Nowadays parallel corpora are in electronic form and they have become an important resource in language engineering while they are used widely in multilingual lexicography and terminology, human and Machine Translation (MT), Multilingual Information Retrieval, language learning and so on.

In language learning parallel corpora can be used by extracting basic linguistic information from texts for teaching and learning of the language pairs. They can be used by students in order to find translation pairs and learn translation techniques [Danielsson and Mahlberg, 2003]. It is considered as a challenge for the student to understand the translated sentences and built concepts and structure, based on the original one, supplementing in this way the teaching process.

Parallel corpora can be found useful in multilingual terminology. As the technology evolves, new terms are introduced in new subject areas that are not included in existing dictionaries. Analysis of parallel corpora at a word alignment level is a useful mean in the extraction of multilingual terminology which is used by terminologists and translators [van der Eijk, 1993].

In the field of Multilingual Information Retrieval, the query written in one language must be translated in to the target languages of the documents under demand. The difficulty occurs when multi terms of the query form a phrase, unable to be identified by bilingual dictionaries. Parallel corpora can be used then for a word to word

[2] Comparative translation involves the identification of equivalences between signs or terms and the application of that knowledge for translating unknown texts by making educated guesses about what signs or words stand for.

translation based on translation probability using larger blocks of aligned text [Bannard and Callison-Burch, 2005].

Parallel corpora are turned out to be a powerful tool for automated translation. They are utilised in statistical methods in order to automatically extract word translation equivalents with minimal or without the use of linguistic information.

2.2 Characteristics and differences between the Greek and English languages

Languages can be classified based on two main categories: genetic and typological [Greenberg, 2001]. A genetic classification divides them in to language families according to their development through history. A typological classification divides languages according to their similarities or differences based on their structure, morphology and so on.

Both Greek and English languages belong to the Indo-European language family. English together with other languages belong to the group of Germanic languages forming a branch of the Indo European family while the Greek language forms an independent branch on its own.

The main difference between the Greek language and all the Indo-European languages and therefore including English is its unique alphabet. Greek language has a unique alphabet which differs a lot from the English one. This is considered to be an interesting point if someone considers the fact that there are statistical translation methods that make use of the number of common characters between two words (String matching) to identify translation equivalents [National Virtual Translation Centre].

Based on the typological classification, the two languages have similarities but also many differences. Formally, both languages have the same word order following the Subject-Verb-Object syntax. However, the word order in Greek is relatively flexible and there are many correct ways to form a sentence concerning the word ordering, depending on what is to be emphasised [Foundalis].

Another major difference of the two is that Greek is a highly inflected[3] language in contrast to the English one which appears minimal inflection. That is a big issue when it comes to statistical translation methods because the same word might be appearing in the text with many different forms. In his research for the implementation of a Greek stemmer, Ntais [2005] identified 166 different suffixes for the general forms of the main inflectional types of the Greek language. That means that there might be created translation pairs with the same English word translated to a different form of the same Greek word each time. This is a common issue that makes alignment of words difficult just because languages with minimal inflectional morphology like English do not contain enough information for producing the correct full form in the target language as happens with the morphologically rich languages [Ueffing and Ney, 2003].

[3] Inflection is the variation in the form of a word by means of an affix, to reflect grammatical information.

The two languages also differ in the level of their compound-constructing ability. This property of the Greek language allows the combination of terms in order to construct compound words that could used to express what other languages would express with a whole sentence.

2.3 Statistical translation methods

The idea of building a system that could manage automated translation between languages has been first dated sometime in the seventeenth century although it never turned into practice. The first successful attempt was recorded back in the early 1930s by a French-Armenian called George Artsrouni and a Russian Peter Troyanskii when they applied for patents for their "translating machines" [Hutchins, 1986]. Later during the 1950s after the appearance of computers, researchers proposed translation using computers. This was basically the first non numerical application of computers. The following years there were attempts with promising results. Until the 1980s research on automated translation was based on linguistic representation which involved morphological, syntactical and semantic analysis. By that time there was a respectful computer power available and researchers realised its potential in processing parallel texts. It was at the late 1980s when research was turned towards statistical translation methods utilising parallel corpora [Kay and Röscheisen, 1993][Brown et al. 1990].

Statistical translation methods involve the application of statistic techniques applied on parallel corpora in order to determine correspondences between the source and the target languages. Borin [1998] has proved that translation methods using pure statistical techniques are superior to the ones incorporating linguistic information through experiments on the same corpus using a pure statistical method and a method incorporating linguistic information respectively.

Extraction of bilingual dictionary from parallel corpora using statistical methods usually involves the following interdependent general steps:

- Pre-processing of corpus
- Text alignment
- Word alignment

There are many methods and approaches followed for the extraction of word equivalences that comprise bilingual dictionaries. However this section will not list every possible translation approach but it will focus on the most important ones, the ones that are most commonly used and are more relative to the system that is going to be used.

2.4 Pre-processing of corpus

Once the corpus is created there are important issues that have to be taken care of as long as the properties of the corpus are concerned. The basic units of the corpus have to be identified. This is done with pre-processing of the corpus and it takes place before the actual alignment process. The pre-processing of the corpus may include

many steps including identifying the boundaries of paragraphs and sentences, perform tokenization of the texts and usually annotation of the text.

Identification of paragraph and sentence boundaries
This is basically the process of identification of the start and end of the paragraphs and sentences.

Corpus Tokenization
Tokenization is a very important task in corpus processing. It refers to the isolation of word units (Tokens) from text and further separation of punctuation marks, numbers and so on.

Corpus annotation
Corpus annotation is the process of attaching special codes (Tags) to words in order to indicate their special features. Tagging may incorporate linguistic information and depending on the linguistic information that is employed, different methods can be used. The most popular and common annotation methods used are:

Par-of-Speech (POS) *annotation* - refers to the process of assigning part-of-speech tags to each word in the text. Part-of-speech tags describe grammatical features of words such as noun type and number, adjective type and so on, and it is the most common method in corpus linguistics. Part-of-speech tags basically serve two purposes, to allow disregarding words of certain class (prepositions, conjunctions etc.) and to allow distinction between words in different classes [Ueffing and Ney, 2003].

Lemmatization - involves the reduction of the words in a corpus to their respective lemma (the original form of the word). Strömbäck [2005] however points out that he impact of lemmatization depends very much on the corpus and it has better results when it is applied to corpus with a small lexical variation. Furthermore in their work with Scandinavian languages using the Uplug system Dalianis and Rimka [2007, forthcoming], concluded that there is not substantial difference in the results when the corpus is lemmatised.

Parsing - is the process or result of making a syntactic analysis. In corpus linguistics parsing involves the procedure of analysing a sentence in order to identify its grammatical components such as nouns, verbs and so on, without specifying their internal structure though.

Other less common annotation methods that are rarely used nowadays are Discoursal and Text linguistic annotation, Phonetic transcription, Prosody, Problem-oriented tagging [University of Essex, 1998].

2.5 Text alignment

As it was described in more detail earlier in section 2.1.1 Parallel corpora of this thesis, parallel corpora exist in two formats. They can be raw parallel texts or they can be aligned texts. The alignment of the texts is of great importance because it affects all following stages. In the case where the corpus is not aligned the parallel corpora should be processed in order to be aligned. The level of alignment also plays an important part for the work following. Texts can be aligned in paragraph level,

sentence level or even in phrase level and word level. Corpora alignment usually takes place at a sentence level because it allows better observation and exploration of particular words in a variety of reasonably complete contexts [Romary, Mehl and Woolls, 1995]. However there have been suggestions for the use of clauses[4] as translation units instead of sentences [Piperidis, Papageorgiou and Boutsis, 2000] [Boutsis and Piperidis, 1998].

Possible mistakes at this level will have negative impact at the following stages and the fact that there is a chance that one sentence might be translated in to two sentences in the target language increases the possibilities of mistakes.

Sentence alignment
Sentence alignment is an important task in translation methods that use parallel corpora. Ideally the process of sentence alignment should be performed without any special knowledge about the corpus.

Kay and Röscheisen [1993] developed an iterative relaxation approach, based only on internal evidence, which appeared to converge to the correct sentence alignment after only a few iterations even when applied to relatively free translations. However it was not efficient enough to be applied in large corpora [Moore, 2002]. Another approach introduced by Brown, Lai and Mercer [1991] was based on the length of the sentences. This particular approach was based on the number of words in each sentence. Another similar approach was introduced by Gale and Church [1991]. They introduced a method and a program based on a statistical model that utilises the character length within sentences. Their method was based on the simple observation that longer sentences are translated to long sentences in the other language and shorter sentences are translated in to short sentences in the other language. Chen [1993] came up with an approach that uses word translation probabilities and word identities that have showed better results than the length based approaches described above. However this approach was claimed to be much slower than Brown's and Gale's. Later, because of the possibility of follow-up errors in length based sentence alignment the use of known translated anchor words in the parallel documents were used to avoid the problem. There were attempts that are basically considered as combinations or variations of the basic ideas proposed by Brown or Gale and Church like for example Melamed [1997] and Simard and Plamondon [1998] who suggested a geometric approach using anchor words or stop words as some call them and achieved sentence alignment results slightly better than Gale and Church.

2.6 Word alignment

The process of corresponding words from a text in one language to its translation to another language is called word alignment. This process is also used in the case of a bilingual dictionary creation from parallel corpora as intended in this particular thesis project.

Word alignment is a bit more difficult task than sentence alignment. In fact the lower the level of alignment in a corpus, the more difficult the task is. This has to do with

[4] A clause is a collection of grammatically-related words including a predicate and a subject. A collection of grammatically-related words without a subject or without a predicate is called a phrase.

the differences between the languages under study such as the difference in the level of inflection between languages, the word order of languages, the compound level of languages and so on as it was previously described at section 2.2 Characteristics and differences between the Greek and English languages.

There are generally two approaches to word alignment, the association approach using measure of correspondence of words of some kind, and the estimation approach employing probabilistic translation models. Both approaches make use of statistics or statistic techniques. Association approach is more commonly used [Tiedemann, 2003c]. All word alignment methods described assume sentence aligned corpus.

2.6.1 Association approaches

Methods following this approach employ heuristics that most of the times are based either on the *co-occurrence* measures or on *string similarity* measures of words in the two languages.

Co-occurrence measures

Co-occurrence measures presuppose that the texts are sentence aligned and they are based on the idea of counting the frequency of word pairs that co-occurred in the aligned sentences. This frequency is then used in association measures for the identification of word correspondences.

One statistical association measure of co-occurrence is to test if co-occurrence of a pair of words appears considerably more than it would be expected, based on chances.

Another method of co-occurrence measure is by using the *Dice coefficient* which is used to measure the correlation between discrete events. In this case the occurrence of two words in one text and its translation. The Dice coefficient takes a value between 0 and 1 (0, 1) with 1 representing the highest probability of one word being a translation of the other.

A third statistical association measure is *Mutual information* derived from information theory and is a quantity that measures the mutual dependence of two random variables. In the case of word alignment it measures the amount of common information between two words. The idea behind it is that words that are assumed to have a lot of information in common are likely to be translations of one another [Tiedemann, 2003c].

String Similarity measures

Another method for alignment is using string similarity measures. String similarity algorithms can be used to compare the number of common characters of two words.

One algorithm that employs this idea of character comparison is the *Longest Common Subsequence* algorithm (LCS). By using this algorithm, a longest common subsequence ratio can be calculated and therefore a comparison between a pair of words is possible. In the case of a pair of languages with different alphabets an algorithm that maps the different characters of both languages is employed in parallel [Tiedemann, 1999].

Another method utilised for string similarity measures is the *N-grams* method. The main idea of the N-grams approach is the grouping of words that contain many common substrings of N subsequent characters. In this way the character structure of the word is compared and used to find pairs or words and word variants [Kosinov, 2001].

2.6.2 Estimation approaches

Estimation approach makes use of parallel corpora to estimate probabilistic alignment models. This approach has been influenced by statistical approaches in machine translation [Brown et al., 1990] and it is used to handle words that do not have an equivalent correspondence in the other language. In the estimation approach, alignment is modelled as hidden connections in a statistical translation model [Och and Ney, 2003], where each word in a target language string is connected to not more than one word in the source language [Tiedemann, 2003c].

2.6.3 Combined approaches

Combined approaches combine methods of the approaches described above. Weight is assigned on the result of each method that is decided to be used, and the final evaluation of a candidate translation pair is based on the sum of the results of all the evaluations multiplied with their respective weight [Tiedemann, 2003c].

2.7 Evaluation of word alignment systems

There are different ways to evaluate extracted dictionaries. Some of the most common ways are the use of *gold standards* [Ahrenberg, Merkel, Sågvall and Tiedemann, 2000], methods that incorporate the classification of the translations into categories [Sjöbergh, 2005] or comparison of randomly selected pairs of existing dictionaries to the suggested translations. The gold standard method is based on *recall* and *precision* evaluation metrics.

Gold standards

The evaluation of alignment output can be performed by comparing it to gold standards (also called reference data) which is constructed before the alignment process takes place. Gold standards are consisted of sample text and its equivalent in the target languages that is pre-linked by the reviewers and then it is used to test the alignment results automatically. There are two approaches used with gold standards.

The first approach of performing a complete alignment of the sample, breaks down to segments the sentences in the source and target languages and then the translation equivalences are marked.

The second approach is using the “translation spotting” method. In this method a number of words or phrases are extracted from the source text and then all the sentences of the target text that contain these words or phrases are presented to the reviewer in order to choose the corresponding target word or phrase and compare the equivalences.

Classification to categories
In this evaluation method a number of randomly selected words are classified by experts into categories depending on the translation quality (e.g. good translation, acceptable translation or wrong translation etc.).

Use of bilingual dictionaries
In this method a number of randomly selected translation pairs are selected from existing bilingual dictionaries and then it is counted how many of these are correctly matched in the alignment output. The result of the evaluation though is very much depended on the domain of the corpus and the dictionary used.

Evaluation metrics
Evaluation of the output can be performed by experts who perform the evaluation after the alignment. Metrics for evaluation in this case are recall and precision. Recall is defined as the ratio of the correct translations to the possible correct translations.

$$\text{Recall} = \frac{\text{Number of correctly aligned items}}{\text{Numner of possible correct items}}$$

Precision is defined as the ratio of the correct translations over the sum of all translations.

$$\text{Precision} = \frac{\text{Number of correctly aligned items}}{\text{Number of obtained items}}$$

2.8 Uplug system introduction

Uplug-system origins from a project in Uppsala University and provides a collection of tools for linguistic corpus processing, word alignment and term extraction from parallel corpora. It was developed within the on-going PLUG project which stands for Parallel Corpora in Linköping, Uppsala and Göteborg. The purpose of this software is to provide a modular platform for the integration of text processing tools [Tiedemann, 1999b]. Based on that idea every independent external tool which performs a specific task can be used and combined with existing modules for building of specific task applications.

In particular, Uplug's pre-processing tools include a sentence splitter, tokenizer and external part-of-speech tagger and shallow parsers. The following external tools are used: The *TreeTagger* for English, French, Italian, and German, the *TnT tagger* for English, German and Swedish, the *Grok system* for English (tagging and chunking), and the morphological analyzer *ChaSen* for Japanese. Translated documents can be sentence aligned using the length-based approach by *Gale&Church*. Words and phrases can be aligned using the *Clue alignment* approach and the toolbox for statistical machine translation *GIZA++* [Tiedemann, 2004b]. Corpora are pre-processed with language-specific pre-processing modules if available. Otherwise, Uplug will use the *basic* pre-processing modules that adds simple XML markup and runs the sentence splitter and the general tokenizer.

TreeTagger
The TreeTagger is a tool for annotating text with part-of-speech and lemma information which has been developed within the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Spanish, Bulgarian, Russian, Greek, Portuguese and old French texts and is easily adaptable to other languages if a lexicon and a manually tagged training corpus are available [IMS Textcorpora and Lexicon Group, 2003].

TnT tagger
TnT, the short form of Trigrams'n'Tags, is a statistical part-of-speech tagger that is trainable on different languages and virtually any tag set. TnT is not optimized for a particular language. Instead, it is optimized for training on a large variety of corpora. The component for parameter generation trains on tagged corpora [Brants, 1998].

Grok system
Grok is a library of natural language processing components, including support for parsing with categorial grammars and various pre-processing tasks such as part-of-speech tagging, sentence detection, and tokenization [Baldridge, 2001].

ChaSen
ChaSen is a tokenizer and morphological analyser for Japanese [Matsumoto Laboratory, 2003].

The sentence aligner
The sentence aligner applies the approach proposed by Gale and Church [1991] which is basically based on sentence length comparisons between the source and the target language texts.

The word clue aligner
The word aligner implemented in the Uplug system is the Clue Aligner. The word alignment approach used is based on the combination of word alignment clues. The idea is that features like frequency, part-of-speech, parsing and word form as described above, together with similarity and frequency measures are taken into account and are considered as association clues between words. All these association clues are then combined together in order to find links between words in the source and target languages [Tiedemann, 2003].

GIZA++ toolbox
GIZA++ is based of the existing statistical machine translation toolkit GIZA and is extended with implemented training algorithms for statistical translation models [Och, 2001].

Iterative size reduction
Uplug makes use of all the above tools in order to extract a basic one-to-one (1:1), one-to-many (1:X) and many-to-one (X:1) dictionary. This basic dictionary is then used to analyse the rest of the test and remove known translations. The size of the remaining text is getting smaller and a new one-to-one (1:1) alignment is performed. The new obtained alignments are then added to the basic dictionary. This new improved dictionary is then used to analyse the remaining alignments of the previous

step. This process is repeated iteratively until there are no new one-to-one (1:1) word alignments [Tiedemann, 1998].

2.9 Web searching across languages

Most Internet users perform web site searching using a search engine to locate online information or services. Search engines are devoted to facilitate user searches and they have dedicated most of their resources in order to achieve efficiency in their results [Baeza-Yates and Ribeiro-Neto, 1999]. Efficiency is an important factor for the users in a web search. Although there are services provided by search engines to facilitate web site searching they are not even close to make the most out of the Web's potential for polyglot and non polyglot users successfully.

There are search engines which introduced services that allow users to choose the language in which they would like their results to appear or they even provide machine translation services, where a user can translate a web page in the language of his preference and therefore take advantage of resources in languages unfamiliar to them. However these solutions are considered as monolingual searches in a way that the results are basically in the language that they are requested. Moreover the results are very much depended on the variety of languages that a search engine utilises. Furthermore research has shown that even polyglot users, they do not use search engines as multilingual tools and they do not make as much use of these services as it would be expected. Rieh and Rieh [2005] in their research on the preferences and behaviour of bilingual users in Korea, came to the conclusion that the users still insert queries for web site searching one language at a time, in the languages they are more familiar with and in the language that represent their information need more accurately. For example in relation to the language pair concerned in this thesis, if the users would like to get results in English and in Greek, they will first use a query written in English and then they would insert a query written in Greek.

2.10 Multilingual Web Retrieval

The broad applications of the web and its potential from the perspective of information resources are seen by many as a challenge in the field of Information Retrieval (IR) [Zhou et al., 2005]. "*Information retrieval deals with the representation, storage, organization of, and access to information items*" [Baeza-Yates and Ribeiro-Neto, 1999]. With the emergence of machine translation, Information retrieval has been evolved and research has been turned towards Multilingual Information Retrieval.

"*The term Multilingual Information Retrieval (MLIR) refers to the ability to process a query for information in any language, search a collection of objects, including text, images, sound files, etc., and return the most relevant objects, translated if necessary into the user's language." [Klavans and Hovy,* 1999]

Considering the diversity of the languages used by the non English speaking population, a user might find multilingual web retrieval extremely useful when it comes to web site searching. Web retrieval refers to the ability to process a query for

information over the web in any language and return results relevant to the user's query also in any language.

Although there is scepticism about the integration of Multilingual Information Retrieval applications to web retrieval, there have been attempts that have shown promising results of its successful application to multilingual web retrieval and web site searching [Zhou et al., 2005].

2.10.1 Approaches to Multilingual Web Retrieval

There are two approaches used in Multilingual Information Retrieval: document translation or query translation. The second is closely related to the outcome of this thesis work.

In the document translation approach all documents in the whole database are translated. It is therefore easy to understand that the document translation approach requires a large amount of computational and storage resources, it is very cost intensive and thus it is avoided.

In contrast query translation translates the query into all target document languages and then monolingual retrieval is performed separately for each document language. This approach is most commonly used as it is much easier to implement it and the only requirement is a tool for the translation of the query text, usually a machine readable bilingual dictionary [Kishida and Kando, 2005].

2.10.2 Approaches for query translation

Looking at the approaches available for query translation it is clear that the outcome of this thesis is closely related to searching over the Internet in multiple languages. There are three approaches for query translation adopted: using machine translation, a parallel corpus, or a bilingual dictionary.

Machine translation based approach uses existing machine translation techniques to perform automatic translation of the queries. The application of this approach is simple but the quality of the results is not very satisfying. The reason for that appears to be the fact that queries usually do not contain enough contextual information that is necessary to machine translation in order to achieve word sense disambiguation [Sakai Tetsuya, 2000].

A corpus based approach uses large collection of parallel texts (corpora) to construct a statistical translation model. It does not depend on manual creation of bilingual dictionaries, however this approach is very much depended on the quality of the corpus while sometimes it is difficult to find parallel corpus, especially for languages that are not very popular [Oard, 1997].

The main idea in a dictionary based approach is to replace each term of the query with the equivalent term or set of terms in the desired language. The equivalent terms are looked up into a bilingual dictionary. This is the most popular approach because of the simplicity of its application and the existence of a variety of machine readable bilingual dictionaries. However this approach lacks of consistency in the quality of

the results. The reason behind this is the fact that this approach fails to translate many terms as a phrase, while morphological differences between languages might introduce noise because of many definitions of a word [Ballesteros and Croft, 1996].

However there are not many machine readable bilingual dictionaries for small language pairs. The work of this thesis is a contribution towards the creation of a bilingual Greek-English dictionary that could be used as a tool for Web site searching.

3. Methodology

The method used for the practical part of the project is described in the following five sections.

3.1 Preparation of working environment

For the particular thesis project and the processing of the corpus the Uplug system was used. Uplug is a collection of tools for linguistic corpus processing, word alignment and term extraction from parallel corpora. Uplug exists in two versions: UplugWeb which is a web interface of the system and Uplug as a standalone application.

UplugWeb is the web interface of the corpus tools and its original version is installed at the Department of Linguistics and Philology at Uppsala University[5]. It can be used by registered users with small size corpora. Processing corpora with UplugWeb might take relatively more time and be really slow because UplugWeb processes are queued on the local system and have a lower priority to the university's server.

Uplug as a stand alone application is a free for non commercial use application, under the GNU General Public License[6] (GPL) and is running on UNIX like operating systems.

In order to have a better control of the system's processes and take advantage of the processing power for faster processing of the corpus, Uplug system was installed and configured in order to run on a local server in the Department of Systems and Sciences (DSV) at the IT University.

3.2 Collection of parallel corpora

In order the resulted dictionary to be as more accurate as possible, a big amount of parallel corpora was needed.

There are many available public corpora over the web. The most interesting attempt of publicly available parallel corpora resource though is the OPUS corpus [Tiedemann and Nygaard, 2004]. OPUS is a collection of translated open source documents available on the internet. However the corpus provided is already aligned and encoded using XML format and UTF-8 character encoding conversions. There were concerns about the optimised corpora available in the way that optimised corpora would give optimised results while the intention of this thesis project is to work with as more realistic input elements as possible.

In order to test the full potential of the Uplug system including its sentence alignment process and in combination with the point made above, about optimising the corpus for better results, it was thought necessary the use of raw text parallel corpora. Therefore a manually created corpus was created.

[5] http://www.lingfil.uu.se/
[6] http://www.gnu.org/copyleft/gpl.html

The collection of translated documents was based on criteria such as:

- The domain of the texts, so the parallel texts should have a relatively similar vocabulary
- The type of the text (journals, articles etc.)
- The language consistency (English of UK, USA or Australia etc.)
- The completeness of the texts, in a way that the full document should be included or at least as close to that, and not only part of it as it affects the representativeness of the sample

The above criteria had been considered in order to achieve as higher frequency of word occurrence as possible in the parallel texts. Especially for the extraction of bilingual dictionaries specialised text corpus on certain domain is selected in order to achieve maximum coverage on a specific topic and improve results.

The documents included in the corpus used for this thesis project were mainly collected from the European Union's portal web site [Europa, 2006]. All the information made available on the Web by the institutions and bodies of the European Union, can be found translated in at least the languages which were official at the date of publication, including in most cases a Greek translation. However, there was one document included that was retrieved from NATO's on-line library [North Atlantic Treaty Organisation, 2006] also translated in English and Greek.

Documents translated in Greek and English were found and compared before included in the corpus. All web documents of the respective languages were stripped from their HTML format and were included in plain unformatted text in one single text source file. Documents available in .PDF format had to be included also in unformatted text in the text source file. By doing this though all previous alignment was gone and therefore all text had to be aligned again at a document and paragraph level to their original condition manually.

Moreover because Uplug is a memory-demanding application a certain number of blank lines had to be included in the text source file in order to allow Uplug to run smoothly and do not run out of memory. These blank lines are translated in page brakes during the XML tagging process as it is going to be explained later in the report. These page breaks are used to separate the source text in different parts. The page breaks are inserted at document limits or in places within large documents in order to break them down to reasonable sized parts. Uplug then creates a virtual matrix containing these parts. The matrix is then used to facilitate text alignment and the sentence alignment process. This is done in a way that the equivalent of a sentence of a certain part of text in the source language is looked only in the respective part of the text in the target language and not in the whole text source file of the target language. This way allows a more efficient usage of memory. A description of the matrix is shown in Table 1 below.

Greek Text / English Text	**Text part_1_el**	**Text part_2_el**	**. . . . . .**	**Text part_n_el**
Text part_1_en	**X**			
Text part_2_en		**X**		
. . .			**. . . . . .**	
Text part_n_en				**X**

Table 1: Virtual matrix used by Uplug
Description of the virtual matrix used by Uplug utilising the page breaks by separating the whole text in smaller parts in order to facilitate text alignment. Sentence alignment is performed only between equivalent documents. The equivalent of a sentence in Text part_1_en of the English text will be looked only in Text part_1_el and not in the whole Greek text, a sentence of Text part_2_en will be looked only in Text part_2_el and so on.

Finally the resulted corpora were converted and saved using UTF-8 (8-bit UCS/Unicode Transformation Format) character encoding. After some problems with other character encodings encountered in various computers without a Greek character set installed, the use of UTF-8 format was the simplest working solution as it is able to represent any universal character in the Unicode standard.

The final parallel corpora were proof read, compared and double checked. This part was the most tedious and time consuming process and it took a little more than one week of manual work.

3.3 Processing of parallel corpora

Pre-processing of the parallel corpora was performed using the collection of tools provided by Uplug. Uplug uses language-specific pre-processing modules if available. In other case Uplug uses the *basic* pre-processing modules. The process through the corpus pre-processing and the actual sentence and word alignment is described in Figure 1 bellow.

Source text in English is named after Text_en (en stands for English) while source text in Greek is named after Text_el (el stands for Greek).

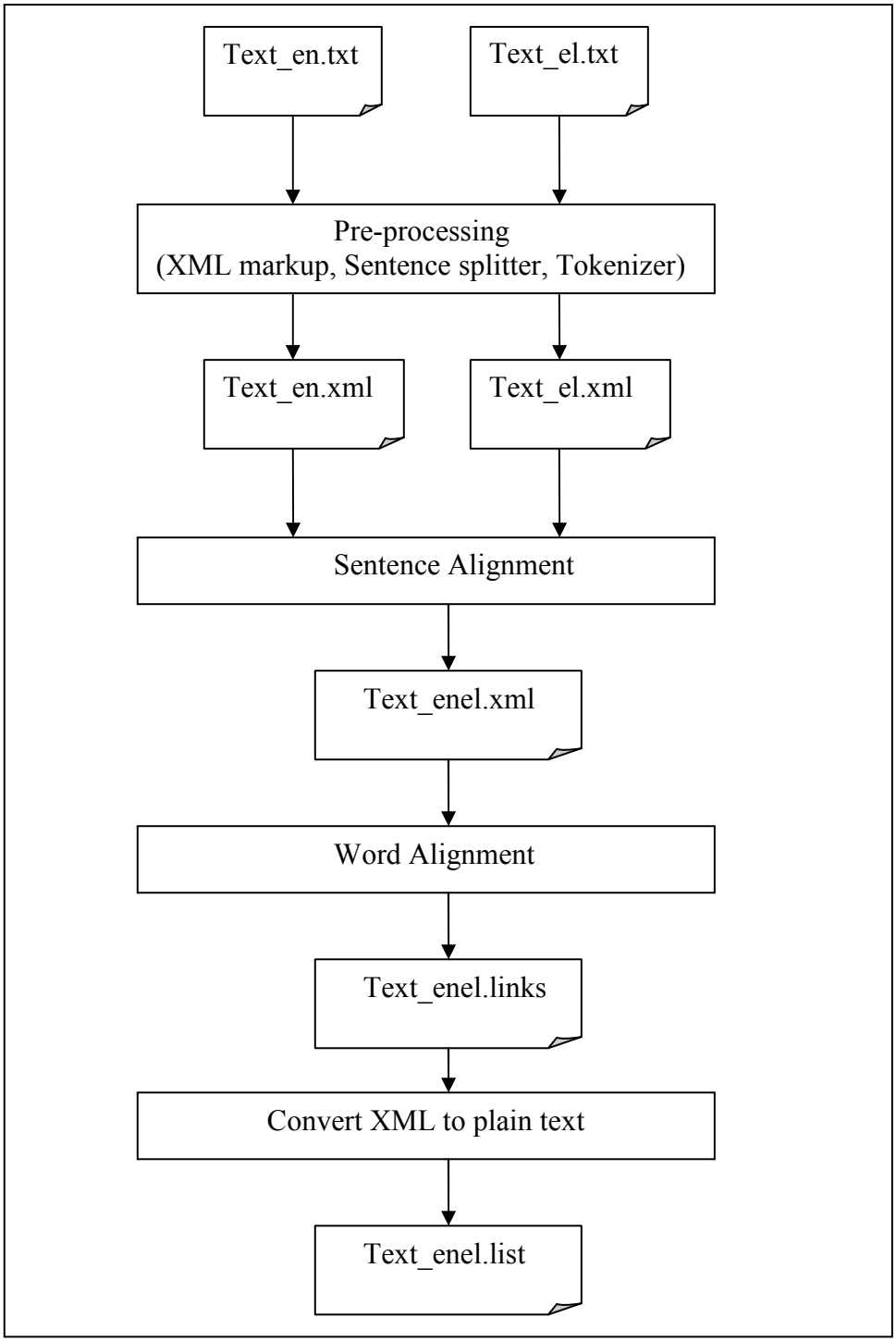

Figure 1: Parallel corpora processing with Uplug
Description of the process of parallel corpora and the outcomes throughout the different stages of pre-processing, sentence and word alignment performed by Uplug system.

As it is depicted in Figure 1 above the two documents in the source and target language are fed as an input to Uplug. Basic pre-processing is performed including addition of simple XML markup and running the sentence splitter and the general tokenizer. XML markup adds some basic markup to the plain text (paragraph breaks and page breaks) basically at empty lines in the file. The sentence splitter adds sentence boundaries to the XML-encoded documents. The tokenizer adds word boundaries to the XML-encoded documents. The results of this process are two respective XML files which are tokenized and marked with XML tags.

```
<p id="6">
<s id="s6.1">
 <w id="w6.1.1">The</w>
 <w id="w6.1.2">Council</w>
 <w id="w6.1.3">of</w>
 <w id="w6.1.4">the</w>
 <w id="w6.1.5">European</w>
 <w id="w6.1.6">Union</w>
 <w id="w6.1.7">passes</w>
 <w id="w6.1.8">laws</w>
 <w id="w6.1.9">,</w>
 <w id="w6.1.10">usually</w>
 <w id="w6.1.11">legislating</w>
 <w id="w6.1.12">jointly</w>
 <w id="w6.1.13">with</w>
 <w id="w6.1.14">the</w>
 <w id="w6.1.15">European</w>
 <w id="w6.1.16">Parliament</w>
 <w id="w6.1.17">.</w>
</s></p>
```

```
<p id="6">
<s id="s6.1">
 <w id="w6.1.1">Το</w>
 <w id="w6.1.2">Συμβούλιο</w>
 <w id="w6.1.3">της</w>
 <w id="w6.1.4">Ευρωπαϊκής</w>
 <w id="w6.1.5">Ένωσης</w>
 <w id="w6.1.6">θεσπίζει</w>
 <w id="w6.1.7">νομοθεσία</w>
 <w id="w6.1.8">και</w>
 <w id="w6.1.9">,</w>
 <w id="w6.1.10">συνήθως</w>
 <w id="w6.1.11">,</w>
 <w id="w6.1.12">νομοθετεί</w>
 <w id="w6.1.13">σε</w>
 <w id="w6.1.14">συνεργασία</w>
 <w id="w6.1.15">με</w>
 <w id="w6.1.16">το</w>
 <w id="w6.1.17">Ευρωπαϊκό</w>
 <w id="w6.1.18">Κοινοβούλιο</w>
 <w id="w6.1.19">.</w>
</s></p>
```

Figure 2: The .xml files after pre-processing
Excerpts of the pre-processed texts in the .xml files, showing tagging in respective paragraphs of English and Greek texts.

The excerpts of the .xml files in Figure 2 above show the annotation used for the sixth paragraph of the English text. The sixth paragraph of the text is annotated using basic XML markup as <p id="6">. The first sentence of the sixth paragraph is then annotated as <s id="s6.1"> while for the annotation of tokens included in a sentence the syntax used is indicating the paragraph, sentence and its place in the sentence (e.g. <w id="w6.1.1">The</w>).

The next stage is sentence alignment which links sentences from the source language document to sentences in the target language document using their sentence ID's obtained in the previous step. The result is one XML file (Text_enel.xml) containing sentence link certainty.

```
<link certainty="404" xtargets="s4.1;s4.1" id="SL0.5" />
<link certainty="400" xtargets="s5.1;s5.1" id="SL0.6" />
<link certainty="715" xtargets="s6.1;s6.1" id="SL0.7" />
<link certainty="301" xtargets="s7.1;s7.1" id="SL0.8" />
<link certainty="1115" xtargets="s8.1;s8.1" id="SL0.9" />
```

Figure 3: The .xml file after sentence alignment
Excerpt of the resulted .xml file after sentence alignment showing link certainty between sentence pairs.

In Figure 3 above is shown how Uplug aligned some sentences. Highlighted is the proposed alignment of the sentences in sixth paragraph from example above. *Link certainty* is a ratio used to indicate the possibility of one sentence to be the equivalent of another. The higher the link certainty the more possible the two sentences to be translations of one another. Then the sentence IDs of the proposed equivalence are listed and finally the link is assigned an ID itself.

The word alignment implemented in the Uplug system is performed by the Clue Aligner. Basic clues like co-occurrence measures and string similarity measures are combined in order to assign links between words from the source and target languages as described above in section 2.8. The result of word alignment will be a .link file (Text_enel.links) which contains the links between words of the source and target language texts together with a *wordLink certainty* ratio (see Figure 4 below).

```
<link certainty="715" xtargets="s6.1;s6.1" id="SL0.7">
 <wordLink certainty="0.0196709634389986" lexPair="laws;θεσπίζει"
xtargets="w6.1.8;w6.1.6" />

<wordLink certainty="0.117290454990787" lexPair="The;To"xtargets="w6.1.1;w6.1.1"/>

<wordLink certainty="0.00543996024343188" lexPair="passes usually legislating jointly;
νομοθεσία νομοθετεί" xtargets="w6.1.7+w6.1.10+w6.1.11+w6.1.12;w6.1.7+w6.1.12" />

<wordLink certainty="0.0857450489923296" lexPair="Parliament;Κοινοβούλιο"
xtargets="w6.1.16;w6.1.18" />

<wordLink certainty="0.0501851920446233" lexPair="European European;Ευρωπαϊκής
Ευρωπαϊκό" xtargets="w6.1.5+w6.1.15;w6.1.4+w6.1.17" />

<wordLink certainty="0.00599642766063619" lexPair="Council of the the;Συμβούλιο της
και συνήθως συνεργασία με το"xtargets="w6.1.2+w6.1.3+w6.1.4+w6.1.14;w6.1.2+
w6.1.3+w6.1.8+w6.1.10+w6.1.14+w6.1.15+w6.1.16" />

<wordLink certainty="0.111732453938216" lexPair=",;, ,"
xtargets="w6.1.9;w6.1.9+w6.1.11" />

<wordLink certainty="0.0306590753224423" lexPair="with .;σε ."
xtargets="w6.1.13+w6.1.17;w6.1.13+w6.1.19" />

 <wordLink certainty="0.0866608210304425" lexPair="Union;Ένωσης"
xtargets="w6.1.6;w6.1.5" />
</link>
```

Figure 4: The .links file after word alignment
Excerpt of the resulted .links file after sentence alignment showing wordLink certainty between words.

The last step is the conversion of the XML file to a text file (Text_enel.list). The final readable text file includes the translated word pairs together with a number indicating the frequency of occurrence of each translation pair as shown in Figure 5 bellow.

```
 90 European      Ευρωπαϊκό
 47 accordance    σύμφωνα
 35 between       μεταξύ
 18 members       μέλη
 14 policy        πολιτική
  9 protection    προστασία
  5 citizens      πολίτες
  3 categories    κατηγορίες
```

Figure 5: The readable .list file
Excerpt of the resulted .list file after conversion of the .links file to a text file.

3.4 Extraction of sample data

For the evaluation of the results a sample of the output data was used in order to make safe conclusions for the characteristics of the whole population under study. Therefore the definition of the sample data should be such, in order to be representative of the whole population of results.

For the extraction of the sampling data the stratified sampling method was used. In this method the population is divided in to non overlapping categories (stratums) where the elements of each category share one common characteristic. Then random sampling is used to select a sufficient number of elements from each stratum.

Following this method the population of translation pairs with frequency of occurrence above three was divided in five categories. The five categories are:

- pairs with frequency of occurrence equal to 3 (freq=3)
- pairs with frequency of occurrence equal to 4 (freq=4)
- pairs with frequency of occurrence equal to 5 (freq=5)
- pairs with frequency of occurrence equal to 6 up to 10 ($6 \leq freq < 11$)
- pairs with frequency of occurrence equal to11 up to maximum ($11 \leq freq < max$)

Then a random sample of 100 suggested translation pairs from each category was drawn and five different tables were created. Each table contained 100 translation pairs that were collected randomly from one of the five categories mentioned earlier. These tables were the ones to be used in questionnaires for evaluation of the dictionary.

3.5 Evaluation of results

The results of the dictionary were evaluated by randomly drawing suggested translation pairs of words and classifying them in to categories depending on their translation quality. This is a quite common way to evaluate automatically created bilingual dictionaries [Sjöbergh, 2005].

The first classification of the suggested translations was based on qualitative data and was used for qualitative analysis of the results. It was performed by at least three fluent Greek - English speaking persons and the sample of the suggested translation pairs was categorised according to the five following categories.

1. Accurate – the suggested translation is accurate
2. Somewhat correct – the suggested translation is correct but not accurate
3. Undecided – person evaluating cannot make a decision about the translation pair
4. Somewhat incorrect – the translation is not correct but similar to the correct
5. Wrong – the suggested translation is just plain wrong

The first category "Accurate" is the desirable case where the suggested translation of a word in the target language is an accurate translation of the source word.

The second category "Somewhat correct" is the case where the suggested translation is correct but not accurate. This is the case where the meaning of the word in the target language is correct but might be in a different form or maybe it is a translated synonym of the word in the target language. That means that using the suggested translation of a word someone will understand the meaning of the original word in a text.

The third category is chosen for the case where the reviewers are undecided about the translation pair. This case is always possible and considered necessary to be included in case reviewers are not familiar with a term.

The fourth category "Somewhat incorrect" is the case where the suggested translation is not correct but similar to the correct. This might be the case where translation is not correct but can still be useful for a reader to understand the general meaning of a word in a text.

The fifth and last category "Wrong" is the case when the suggested translation is just plain wrong and therefore cannot be used.

The above categorization was performed by the means of questionnaires. Questionnaires were created in order to be send to Greek English speakers for evaluating the extracted sample of the resulted dictionary. The categorisation has been performed by Greek persons who were fluent Greek-English speakers with a proficiency in English language. The questionnaires included the tables containing the randomly drawn translation pairs as described in section 3.4 Extraction of sample data, together with five checkbox options for each translation pair (see Table 2 bellow). Each on the five options represents one of the five categories described above. The five options were A, B, C, D and E for "Accurate", "Somewhat correct", "Undecided", "Somewhat incorrect" and "Wrong" respectively.

The tables were included in to the questionnaires in random order regarding the frequency of occurrence so that the evaluation will have a better flow and reviewers would not get tired or frustrated realising a pattern in the quality of the translation pairs.

Finally from the analysis of the answers in the questionnaires some quantitative results would be derived that would allow us to derive conclusions about the resulted dictionary.

	English	Greek	Accurate	Somewhat Correct	Undecided	Somewhat Incorrect	Wrong
1.	.	.	A☐	B☐	C☐	D☐	E☐
2.	and	και	A☐	B☐	C☐	D☐	E☐
3.	ARTICLE	ΑΡΘΡΟ	A☐	B☐	C☐	D☐	E☐
4.	and	και	A☐	B☐	C☐	D☐	E☐
5.	Council	Συμβούλιο	A☐	B☐	C☐	D☐	E☐
6.	and	και	A☐	B☐	C☐	D☐	E☐
7.	Constitution/	Constitution/	A☐	B☐	C☐	D☐	E☐
8.	3	3	A☐	B☐	C☐	D☐	E☐
9.	of	της	A☐	B☐	C☐	D☐	E☐
10.	=]	id	A☐	B☐	C☐	D☐	E☐
11.	by	από	A☐	B☐	C☐	D☐	E☐

Table 2: The format of the questionnaire
Excerpt of the questionnaire sent out for the evaluation of the sample of results.

4. Evaluation

This chapter contains a description of the results and their evaluation process together with their analysis and useful information about resources used for future reference.

4.1 Parallel corpora

The corpora used for the creation of the dictionary were created mainly form a collection of documents taken from the European Union's web portal [Europa, 2006] but also one document retrieved from NATO's on-line library [North Atlantic Treaty Organisation, 2006]. Special attention was given so that the Greek and English translations of the respective documents included referred to the same domain, both texts were translations of each other and not translations of a third document, the documents were of a reasonable size and preferably complete documents were included and not parts of them, even though there were a couple of cases where just a part of the original translated documents were included.

The final bilingual corpus created constituted by the Greek text, which contained 204.043 words, and the English text which contained 196.048 words. The Greek text contained 18117 different words while the English text 10450 different words[7] (see Table 3 below).

Corpus	Size	Words	Different words	Characters (no spaces)
English (en)	1,23 MB	196.048	10.450	1.203.662
Greek (el)	2,46 MB	204.043	18.117	1.066.553

Table 3: Characteristics of the parallel corpora

It is worth noticing the difference in the size of the Greek and English corpora. With size 2,46 MB (Mega Bytes) the Greek text is almost double the size of the English text which is of size 1,23 MB. The reason causing this is the UTF-8 character encoding used for the two texts. It happens that Greek text in UTF-8 format increases the size of the file while it seems that this change in the size does not apply to files containing English text.

The difference between the number of total words and number of different words of these two parallel corpora appears because of the richer morphology of the Greek language compared to that of the English language [Boutsis and Piperidis, 1998].

4.2 Results

The final output after the process of the parallel corpora with the Uplug system resulted in a file including all suggested translation pairs. The extracted pairs included many correct but also incorrect translation pairs. These translation pairs might be of the form one-to-one (1:1), one-to-many (1:X), many-to-one (X:1) or many-to-many (X:X) translations (see examples in Figure 6 bellow). In the cases where many words

[7] The word breakdown of the corpora was performed using the TextSTAT - Simple Text Analysis Tool which can be found at URL: http://www.niederlandistik.fu-berlin.de/textstat/

were included in the many side of a translation pair it was noticed that it could include many different terms, a phrase, or even multi-word terms.

```
90 European            Ευρωπαϊκό
37 European            Ευρωπαϊκού
31 European            Ευρωπαϊκής
20 European            Ευρωπαϊκός
20 European            Ευρωπαϊκή
16 European European   Ευρωπαϊκό
13 institutions        θεσμικά όργανα
10 European            Ευρωπαϊκού
10 Member State        κράτος μέλος
 5 European European European    Ευρωπαϊκή
 5 European European European    Ευρωπαϊκό
 3 European            Ευρωπαϊκό Ευρωπαϊκό
 3 European            Ευρωπαϊκού Κοινοβουλίου
```

Figure 6: Types of translation pairs
Examples of one-to-one, one-to-many, many-to-one or many-to-many translations containing many different terms, phrases, or even multi-word terms as well as duplications of terms within a single translation pair or duplications of different translation pairs.

By looking at the examples of results in Figure 6 above it is easy to understand the difficulty to create a Greek English dictionary from parallel corpora without the use of lexical information. In the example above it is obvious that the word “European” can be translated to a few correct translations of Greek terms that all refer to the word European but containing different lexical information and therefore appear in different forms in the Greek language.

There were thoughts of removing pairs which included multi-term translations to decrease the noise substantially but it seemed that the pairs left would consist of a pretty small number of translations. Furthermore it was thought that in multi-term translations there is also a big chance the correct translation to be included in them. If that assumption is made then it is possible to optimize the extracted translations using different methods to compare the resulted translations [Tiedemann, 1997]. Therefore the extracted dictionary was unaltered and all suggested translations were eligible to be included in the evaluation sample.

4.3 Sample

For the evaluation of the extracted dictionary a sample of the output data was used. Many of the suggested translation pairs had a small frequency of occurrence for the size of the corpora processed. Therefore the results had to be filtered and include in the evaluation only the translation pairs with occurrence above a threshold that was such, in order to avoid evaluation of pairs with occurrence that might be based on chance. That threshold was decided to be a frequency of occurrence above or equal to three. Therefore translation pairs with frequency of occurrence less than three were excluded from the process of extraction of the sample and evaluation. The total number of pairs with frequency above or equal to three ($f \geq 3$) was 1276 pairs and 498 of them comprised the sample included in the questionnaires (see Table 4 bellow).

Sample	11 ≤ f<max	6 ≤ f<11	f=5	f=4	f=3	Total (f ≥ 3)
Pairs included in extracted dictionary	221	251	130	218	456	1276
Pairs included in sample	100	100	100	98	100	498

Table 4: Analysis of the extracted sample

It was also decided that it was necessary to clean up the translation pairs comprising the extracted sample of results that were going to be given out for evaluation in the form of questionnaires. Translation pairs with duplicates in the many side were cleaned so that it would be easier for the reviewers to make their decision on the quality of the suggested translation comparing as less terms as possible for each translation. Therefore duplications in the many side were deleted wherever appeared.

As mentioned above there was not any filtering of the duplicate translation pairs performed. All translation pairs with frequency of occurrence above or equal to three were equally possible to be included in the sample. A list of suggested translations in the aligned output is included in **Appendix A.**

4.4 Evaluation method

There are different ways to evaluate extracted dictionaries. Some of the most common metrics used are precision and recall calculations. However, the use of the above metrics is difficult when the alignments are not just one-to-one [Merkel and Ahrenberg, 1998] like it happens in the extracted dictionary as a result of this thesis. Therefore the evaluation method used was based on the judgment of fluent Greek-English speakers on the quality of extracted translation pairs. This is a quite common way to evaluate automatically created bilingual dictionaries as well [Sjöbergh, 2005].

The sample of the extracted dictionary was sent out in form of questionnaires to fluent Greek-English speakers who classified the suggested translations in to one of five categories. The five categories that were given as options were: A, B, C, D and E for "Accurate", "Somewhat correct", "Undecided", "Somewhat incorrect" and "Wrong" respectively. Completed questionnaires were received from twelve persons and analysis of the responses was performed to all twelve of them. A copy of the questionnaire that was sent out is included in **Appendix B.**

The rules for the evaluation were left open so that the evaluation of the quality of the sample was subjective and based on the judgment of the individuals for the classification of each translation pair to one of the suggested categories. No specific rules of how multi term translations, phrases, or even multi-word terms and words with grammatical differences should be judged were given. The reason behind it was the avoidance of biasing the judgment of the reviewers.

4.5 Analysis of results

As it was expected, by analyzing the results of the evaluation it became clear that some reviewers judge the results in a more strict way than others. There were many cases that reviewers judged differently the same translation pair. In some cases reviewers considered pairs classified in the "Somewhat correct" category when just synonyms occurred rather than others who classified pairs in this category when the two words are translated correctly but they differ grammatically (e.g. gender, case, number, tense etc.).

Another point that is worth mentioning is the fact that the questionnaires are a good way to capture the people's choice but in our case they might have inserted a very small error in the results. The questionnaires were created as protected template files where someone could only check the appropriate checkbox field of their respective choice. However there was not any mechanism to ensure that all translation pairs were evaluated. The completeness of the answers was based on the attention and good will of the reviewer. Therefore some translation pairs were accidentally skipped and a choice for their classification had not been recorded. This might insert a small error in the way that people happen to skip different translation pairs. In this way it might be the case that not all translations have equal chances for evaluation. Some pairs might be evaluated more times than others. However this is an insignificant number comparing it to the total number of evaluations but it was thought necessary to be mentioned.

The results of the analysis of the questionnaires are given in the Tables 5, 6 and 7 below.

Sample	**$11 \leq f < max$**	**$6 \leq f < 11$**	**f=5**	**f=4**	**f=3**
Accurate	42,98 %	43,27 %	30,51 %	23,29 %	20,06 %
Somewhat Correct	24,12 %	19,69 %	18,72 %	16,21 %	14,29 %
Undecided	2,08 %	2,28 %	2,25 %	1,70 %	1,58 %
Somewhat Incorrect	7,84 %	8,79 %	10,70 %	10,92 %	13,04 %
Wrong	22,95 %	25,95 %	37,79 %	47,86 %	51,00 %
Total	99,99 %	99,99 %	99,99 %	99,99 %	99,99 %

Table 5: Analytical distribution of the evaluation results for each stratum of the sample

As expected the lower the frequency of occurrence of translation pairs, the lower the quality measured in terms of accuracy or correctness of translations (Table 5 above).

The sum of the percentages of the categories "Accurate" and "Somewhat correct" for each stratum of the sample is presented in the Table 6 below.

Sample	**$11 \leq f < max$**	**$6 \leq f < 11$**	**f=5**	**f=4**	**f=3**
Accurate	42,98 %	43,27 %	30,51 %	23,29 %	20,06 %
Somewhat Correct	24,12 %	19,69 %	18,72 %	16,21 %	14,29 %
Total	67,11 %	62,97 %	49,24 %	39,50 %	34,36 %

Table 6: Analytical distribution of evaluation of the results for the categories "Accurate" and "Somewhat correct"

Based on the results presented above, the overall distribution of the suggested translations based on their quality is given in Table 7 bellow.

Sample	Accurate	Somewhat Correct	Undecided	Somewhat Incorrect	Wrong
Average	32,02 %	18,61 %	1,98 %	10,26 %	37,11 %

Table 7: Overall distribution of translations of the extracted sample based on their quality

Therefore the correct translations could be summed up to 50,63% of the extracted sample of suggested translations. This is reasonable considering the differences of the two languages and the minimum optimization of the corpora used.

In relation to previous work done the percentage achieved with Uplug is relatively lower but this is controversial because of the different methods and systems used. In their work Piperidis et al. [1997] achieved approximately 94% of correct translations. However they make use of language specific information for better results and they use relatively small corpora, created by technical texts from a software documentation manual which decreases the possibility of errors, because these texts usually include special terminology and strict translations, unlike texts from other domains.

4.6 Error analysis

The resulted output contained a lot of noise. The term noise in the case of word alignment is used to describe every translation of a word with something else other than word like for example punctuation marks, numbers, duplicated terms in translation pairs, duplicated translations and so on. Noise appears because of incorrect or inconsisted translations in the corpora used but it may also occur because of incorrect results from the extraction methods used by the system. Free translations could also insert noise that has as a result to affect substantially the performance.

The system extracted many translation pairs with frequency of occurrence less than three (f=2 and f=1). These translations are not considered worth evaluating as they are not containing any sign of consistency and might be based on chance. The majority of these translations are incorrect although there are exceptions of a few correct ones.

It is also noticed that the lower the frequency of occurrence in the extracted dictionary the more translations of a one-to-many and many-to-many appear. This has to do with the iterative size reduction and alignment used by the Uplug system as described in section 2.8 Uplug system introduction. Uncertain sentence and word alignments are left to be processed at the end and this causes a lot of noise in the form of a big number of many-to-many translations with low frequency of occurrence lying at the bottom of the extracted translations.

4.7 Resources used

The information given bellow concerning resources used is presented for future reference. Uplug was running on a server with two Dual Core processors at 2.0 GHz with 4 GB memory, running Linux operating system and it took Uplug 5 hours and 40 minutes to run the whole process on the particular parallel corpora. No other external resources were used.

5. Conclusions

The objective of the thesis was to use parallel corpora for automated extraction of a bilingual dictionary using the Uplug system without the use of linguistic information. The corpora used contained documents in English and Greek retrieved from the Web. The resulted translations of the dictionary were evaluated by bilingual people in order to assess the quality of the suggested translations.

The creation of the dictionary included two main steps: the creation of the parallel corpora, and the application of statistical techniques.

Parallel corpora creation included the collection of the documents and their filtering to strip any formatting they might contain and create a corpus including only plain unformatted text. The corpus was aligned manually at document level by inserting blank lines at document limits or in places within large documents in order to break them down to reasonable sized parts. A certain number of blank lines were translated in page break during corpus pre-processing so Uplug would run smoothly and do not run out of memory.

The application of statistical techniques presupposes some pre-processing of the corpora and then alignment of text firstly at sentence level and then at word level. All processes after the input of the parallel corpora from pre-processing to dictionary extraction were performed automatically by Uplug. Association measures are applied between words and these association measures are then combined and compared in order to find the most accurate equivalence between words in the parallel corpora.

A sample of the extracted dictionary was then evaluated in order to get some results on the quality of the suggested translations based on human perception.

For the suggested translation pairs of the sample belonging to the stratum with the higher frequency of occurrence, 67.11% of correct translations have been achieved.

It was interesting to notice that, the percentage of accurate translations and the frequency of occurrence of translation pairs are directly proportional in contrast to the percentage of wrong translations and the frequency of occurrence of translation pairs which are indirectly proportional (see table 7). In other words it was noticed a decrease of the percentage of correct translations as the frequency of occurrence of translation pairs decreases and on the other hand it was noticed an increase of the percentage of wrong translations as the frequency of occurrence decreases.

This implies that larger corpora with a bigger collection of documents in the same domain that use the same vocabulary and appear a high frequency of usage of the same words, are more appropriate in order to achieve better word alignment quality..

From the analysis of the evaluation of the extracted dictionary sample, it can be concluded that 50,63% of accurate and correct translations has been achieved. This is a respectful percentage of correct translations if someone considers the minimal optimisation of the corpora used. The parallel corpora were not sentence aligned before they were input in to the Uplug system as it happens with the majority of the corpora used in other similar projects of automated dictionary extraction using other

linguistic corpus processing systems. In our case the parallel corpora contained only raw text, free of any formatting. This is an important point in order to understand the difficulty involved in the corpora processing of the particular parallel corpora.

Moreover, the 50,63% of accurate and correct translations is reasonable, considering the relatively small size of corpora used while the English text contained 196.048 words and the Greek text contained 204.043 words with size of 1,23 MB and 2,46 MB respectively.

The evaluation was performed by twelve different persons while in most other projects evaluation is performed by two or three people. Moreover evaluation rules were left open so the evaluation would be as less biased as possible. Therefore the different perspectives of evaluating the same sample involved, gave a more representative assessment of the sample.

To conclude it is worth mentioning the difficulty of the attempt for an automated extraction of a bilingual dictionary between the Greek and English language pair, because of the many differences between the two languages. The different alphabet and the high level of inflection of the Greek language as well as the fact that it is a language with very rich morphology compared to the English one definitely had a significant affect on the quality of the results.

The final dictionary is certainly of a small size with not so many word translations and with the ones included mainly focused on a certain domain but it could definitely be considered as a small contribution towards the efforts of other researchers. It could also be used as an input for special software tools that are used by search engines for web site searching or even in multilingual information retrieval applications.

Projects like this contribute to minimise or eventually stop the isolation of languages like Greek which is not so spread around the world. Gradually resources of many different languages will be accessible and equally appreciated as it happens with other popular languages like English.

Overall the whole project was challenging and interesting to work with while the results were promising. The results were evaluated and it was appreciated that future work could be done to improve them.

Future work

The suggested translations were obtained with minimal optimisation of the corpora. It would be interesting to see the results from an optimised corpus. For example a sentenced aligned corpus would leave out any doubts of incorrect translations because of inconsistencies in the sentence alignment level that would therefore impact the rest of the alignment process at lower levels.

Furthermore it would be interesting to see the impact on the results of lemmatisation or stemming on the same corpus before it was processed by Uplug. Applying these two techniques to the corpus especially in the case of the Greek language might have significantly improved results as the diversity of the forms a word might appear in Greek would be eliminated and therefore the frequency of occurrence between the English words and the stem or the lemma of the respective Greek words would be

increased. Attempts of using linguistic information would be interesting to evaluate as well like for example by attaching part-of-speech information to the corpora.

Of course the same process could be applied to corpora of bigger size or even with corpora of different domains and acquire a bigger number of suggested translations with a bigger variety in the domain of the suggested translations.

6. References

Ahrenberg Lars., Merkel Magnus, Sågvall A.Hein & Tiedemann Jörg. (2000). Evaluation of word alignment systems. In *Proceedings of the 2nd International Conference on Linguistic Resources and Evaluation (LREC 2000)*, Volume III, (pp 1255-1261) Athens,Greece.

Baeza-Yates Ricardo, Ribeiro-Neto Berthier (1999). Modern Information Retrieval. Addison Wesley

Baldridge Jason. (2001). Grok system, Grok's homepage, Sourceforge.net. [On-line]. Accessed on 11/10/06 at URL: http://grok.sourceforge.net/about.html

Ballesteros Lisa and Croft W. Bruce. (1996). Dictionary Methods for Cross-Lingual Information Retrieval. In *Proceedings of the 7th DEXA Conference on Database and Expert Systems Applications*, (pp. 791-801). Zurich, Switzerland.

Bannard Colin and Callison-Burch Chris. (2005). Paraphrasing with bilingual parallel corpora.. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics* (ACL), (pp 597–604). University of Michigan, Michigan.

Borin Lars. (1998). Linguistics isn't always the answer: Word comparison in computational linguistics. In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALIDA '98*, (pp 140-151). Center for Sprogteknologi and Dept. of General and Applied Linguistics, University of Copenhagen.

Brants Thorsten. (1998). TnT -- Statistical Part-of-Speech Tagging. Department of Computational Linguistics, Universität des Saarlandes, Saarbrücken, Germany. [On-line]. Accessed on 11/10/06 at URL: http://www.coli.uni-saarland.de/~thorsten/tnt/

Brown F. Peter, Cocke John, Della-Pietra A. Stephen, Della-Pietra J. Vincent, Jelinek Frederick, Lafferty D. John, Mercer L. Robert and Roossin S. Paul. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79-85.

Brown F. Peter, Lai Jennifer and Mercer L. Robert. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association of Computational Linguistics (ACL)*, (pp 169-176). 18-21 June, Berkley, California.

Boutsis Sotiris and Piperidis Stelios. (1998). Aligning clauses in parallel texts. In *3rd Conference on Empirical Methods in Natural Language Processing*, (pp 17-26). Granada, Spain.

Chen Stanley. (1993).Aligning Sentences in Bilingual Corpora Using lexical Information. In *Proceedings of the 31st Annual Meeting of the Association for Compntational Linguistics*, (pp 9-13). Columbia, Ohio.

Dalianis Hercules and Rimka Martin. (2007, forthcoming). Cross language information retrieval in the Scandinavian languages using Uplug.

Danielsson Pernilla and Mahlberg Michaela. (2003). There is more to knowing a language than knowing its words: Using parallel corpora in the bilingual classroom. English for Specific Purposes World. Issue 3(6), Volume 2, 2003 Corpus linguistics [On-line]. Accessed on 15/09/06 at URL: http://www.esp-world.info/articles_6/DanielssonMahlberg2003.htm

Europa, European Union hope page, Gateway to the European Union (2006). [On-line]. Accessed on 30/08/06 at URL: http://europa.eu/

Forsberg Markus. (2005). Using statistical methods to improve automatic lexicon extraction. Department of Computing Science, Chalmers University, Sweden.

Foundalis Harry. An Overview of Greek Grammar. Center for Research on Concepts and Cognition, Indiana Univesity. [On-line]. Accessed on 11/10/06 at URL: http://www.cogsci.indiana.edu/farg/harry/lan/grkgram.htm

Gale A. William and Church W. Kenneth. (1991). A Program for aligning sentences in bilingual corpora. In *proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (*ACL)*, (pp 177-184). Berkeley, California.

Global Reach. (2004). Global Internet Statistics [On-line]. Accessed on 23/08/06 at URL: http://global-reach.biz/globstats/index.php3

Greenberg H. Joseph. (2001). The Methods and Purposes of Linguistic Genetic Classification. *Language and Linguistics*, Volume2, Number2, (pp 111-135).

Hóning Matthias. (2005). TextSTAT - Simple Text Analysis Tool. Free University of Berlin. [On-line]. Accessed on 08/11/06 at URL: http://www.niederlandistik.fu-berlin.de/textstat/software-en.html

Hutchins W. John. (1986). Machine Translation: past, present, future. Chichester: Ellis Horwood.

Hutchins W. John. (ed.) (2000). Early years in machine translation: memoirs and biographies of pioneers. Amsterdam/Philadelphia, John Benjamins.

IMS Textcorpora and Lexicon Group. (2003). TreeTagger - a language independent part-of-speech tagger. Institute for Natural Language Processing (IMS), Universität Stuttgart, Germany. [On-line]. Accessed on 11/10/06 at URL: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

Kay Martin and Röscheisen Martin. (1993). Text translation alignment. Computational Linguistics 19(1):121-142.

Kishida Kazuaki and Kando Noriko. (2005). Hydrid approach of Query and Document translation with pivot language for Cross-Language Information Retrieval.

Klavans Judith, Hovy Eduard. (1999) Multilingual (or Cross-lingual) Information Retrieval, [On-line]. Accessed on 11/10/06 at URL: http://www.cs.cmu.edu/~ref/mlim/chapter2.html

Kosinov Serhiy. (2001). Evaluation of n-grams conflation approach in text-based information retrieval. In *Proceedings of International Workshop on Information Retrieval*. Oulu, Finland

Matsumoto Laboratory, Nara Institute of Science and Technology. (2003). ChaSen morphological analyzer. NAIST Computational Linguistics Lab. [On-line]. Accessed on 11/10/06 at URL: http://chasen.naist.jp/hiki/ChaSen/

Melamed I.Dan. (1997) A Portable Algorithm for Mapping Bitext Correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, (pp 305–312). Madrid, Spain.

Merkel Magnus and Ahrenberg Lars. (1998).Evaluating Word Alignment Systems. PLUG report. Department of Computer and Information Science, Linköping university, Sweden.

Moore C. Robert. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association forMachine Translation in the Americas on Machine Translation: From Research toReal Users* (AMTA '02), (pp 135–144). San Francisco, California.

National Virtual Translation Centre. English, Languages of the World. [On-line]. Accessed on 11/10/06 at URL:
http://www.nvtc.gov/lotw/months/december/English.html

National Virtual Translation Centre. Greek, Languages of the World. [On-line]. Accessed on 11/10/06 at URL: http://www.nvtc.gov/lotw/months/february/greek.html

North Atlantic Treaty Organisation-NATO. (2006). On-line library, Introductory publications to NATO, [On-line]. Accessed on 05/10/06 at URL:
http://www.nato.int/docu/home.htm#theme

Ntais Georgios. (2005). Development of a stemmer for the Greek language. MSc thesis, Department of Computer and Systems Sciences, Stockholm University / Royal Institute of Technology, Stockholm Sweden.

Oard W. Douglas. (1997). Adaptive Filtering of Multilingual Document Streams. College of Library and Information Services, University of Maryland, USA.

Och Franz Joseph. (2001). GIZA++: Training of statistical translation models. University of Technology, Aachen, Germany. [On-line]. Accessed on 11/10/06 at URL:
http://www-i6.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html

Och Franz Joseph and Ney Hermann. (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19–51.

Piperidis Stelios, Boutsis Sotiris and Demiros Iason. (1997). Automatic translation lexicon generation from multilingual texts. In *Workshop on Multilinguality in the Software Industry: the AI Contribution (MULSAIC'97), Fifteenth International Joint Conference on Artificial Intelligence (IJCAI'97)*, (pp 57-62). Nagoya, Japan.

Piperidis Stelios, Papageorgiou Haris. and Boutsis Sotiris. (2000). From sentences to words and clauses. In Veronis, J. (Ed) Parallel Text Processing, Alignment and use of translation corpora, Kluwer Academic Publishers, Text Speech and Language Technology Series, (pp. 117-138).

Rieh Hae-Young. and Rieh Soo Young. (2005). Web searching across languages: preference and behavior of bilingual academic users in Korea. Library & Information Science Research. Vol. 27 no. 2, (pp 249-263). RR.

Romary Laurent, Mehl Nathalie and Woolls David. (1995). The Lingua Parallel Concordancing Project: managing multilingual texts for educational purposes. Text Technology 5 (3).

Sakai Tetsuya. (2000). MT-based Japanese-English Cross-language IR Experiments Using the TREC Test Collections. In *Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, (pp. 181-188). Hong Kong, China.

Simard Michel and Plamondon Pierre. (1998). Bilingual Sentence Alignment: Balancing Robustness and Accuracy. Machine Translation 13(1) (pp59–80).

Sinclair John. Developing Linguistic Corpora: a Guide to Good Practice Corpus and Text - Basic Principles. (2004). Tuscan Word Centre. *AHDS Literature, Languages and Linguistics (2006)*. [On-line]. Accessed on 29/08/06 at URL: *http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm*

Sjöbergh Jonas. (2005). Creating a free digital Japanese-Swedish lexicon. In *proceedings of PACLING '05*, (pp 296-300). Meisei University, Japan.

Strömbäck, Peter (2005). The impact of lemmatization in word alignment. Master's thesis, Uppsala University, Department of Linguistics and Philology Unversity, Uppsala, Sweden.

Sågvall A. Hein. (1999). The PLUG-project: Parallel Corpora in Linköping, Uppsala, Göteborg. Uppsala University.

The British Museum. (2000).The Rosetta Stone. [On-line]. Accessed on 13/11/06 at URL: http://www.thebritishmuseum.ac.uk/compass/ixbin/goto?id=OBJ67

The General Secretariat for Greeks Abroad (GGAE). Diaspora General Facts (2006). [On-line]. Accessed on 23/08/06 at URL: http://www.ggae.gr/gabroad/organosi.en.asp

Tiedemann Jörg. (1997). Automatical lexicon extraction from aligned bilingual corpora., Diploma thesis, Department of Computer Science, University of Magdeburg, Germany.

Tiedemann Jörg. (1998). Extraction of translation equivalents from parallel corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics NODALIDA '98*, (pp 120-128). Center for Sprogteknologi and Dept. of General and Applied Linguistics, University of Copenhagen.

Tiedemann Jörg. (1999). Automatic Construction of Weighted String Similarity Measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, University of Maryland, MD, USA, 1999.

Tiedemann Jörg. (1999b). Uplug- a modular corpus tool for parallel corpora. In the *Parallel Corpus Symposium (PKS99)*. Uppsala University, Sweden.

Tiedemann Jörg. (1999c). Word alignment - Step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics NODALIDA '99*. University of Trondheim.

Tiedemann Jörg. (2003). Combining clues for word alignment. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL03)* Budapest, Hungary.

Tiedemann Jörg. (2003c). Recycling translations. Extraction of lexical data from parallel corpora and their application in natural language processing. Acta Universitatis Upsaliensis. Studia Linguistica Upsaliensia 1.

Tiedemann Jörg and Nygaard Lars. (2004). The OPUS corpus - parallel & free. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal.

Tiedemann Jörg. (2004b). The Uplug homepage, Source forge.net. [On-line]. Accessed on 23/08/06 at URL: http://uplug.sourceforge.net/

Ueffing Nicola & Ney Hermann, (2003). Using POS information for statistical machine translation into morphologically rich languages. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, (pp347–354). Budapest, Hungary.

University of Essex. Corpus Anotation. (1998). Corpus Linguistics. W3-Corpora project. [On-line]. Accessed on 11/10/06 at URL: http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/types/annotated.html

van der Eijk Pim. (1993). Automating the acquisition of bilingual terminology. In *6th Conference of the European Chapter of the Association for Computational Linguistics (ACL)*, (pp113-. 119). Utrecht University, Netherlands.

Zhou Yilu, Qin Jialun, Chen Hsinchun & Nunamaker F. Jay. (2005). Multilingual web retrieval: An experiment on a multilingual business intelligence portal. HICSS.

Appendix A

```
807 .  .
482 and      και
367 ,  ,
193 The      Το
186 The      Η
174 ARTICLE  ΑΡΘΡΟ
172 and and  και
164 Council  Συμβούλιο
147 1  1
146 2  2
134 or ή
130 en el
115 and      και και
107 of των
106 Constitution/ Constitution/
 96 Commission    Επιτροπή
 95 3  3
 93 of της
 93 Article  Article
 91 ,  , ,
 90 European Ευρωπαϊκό
 90 , ,      ,
 89 = ]      id
 86 Union    Ένωση
 84 [ id     [ = ]
 79 by από
 78 III-     III-
 75 Union    Ένωσης
 72 to να
 72 framework     πλαίσιο
 71 The      Οι
 69 not      δεν
 67 and and and   και
 66 I- I-
 63 for      για
 59 the      το
 59 Parliament    Κοινοβούλιο
 59 0001     0001
 58 may      μπορεί
 54 be από
 54 Council  Συμβουλίου
 54 . .      .
 53 The      Ο
 52 States   μέλη
 51 The      Τα
 50 in τις
 49 is είναι
 48 and and  και και
 48 Commission    Επιτροπής
 47 accordance    σύμφωνα
 45 on για
 42 Member   τα
```

40 measures μέτρα
40 Parliament Κοινοβουλίου
40 4 4
38 , a ,
38 , , , ,
37 with με
37 to την
37 European Ευρωπαϊκού
36 their τους
36 of of των
35 the του
35 of of της
35 between μεταξύ
34 right δικαίωμα
34 for τη
34 , , , ,
31 that ότι
31 Member τα κράτη
31 European Ευρωπαϊκής
30 of των των
30 It Η
30 Article άρθρου
30 , , , ,
29 to to να
29 the the του
29 has έχει
29 a , ,
27 which οποία
27 adopt εκδίδει
26 Constitution Σύνταγμα
26 .]
25 or or ή
25 Member κράτη
25 Article άρθρο
25 ( ) η
24 in and και
24 in με
24 In [
24 II- II-
23 act αποφασίζει
23 Council Council Συμβούλιο
22 the the το
22 be . .
22 and and and and και
22 Union shall Ένωσης
22 SECTION ΤΜΗΜΑ
22 III- shall III-
21 shall . .
21 majority πλειοψηφία
21 be by από
21 Member States κρατών μελών
21 If Εάν
21 . . .
20 shall Union Ένωσης
20 establish θεσπίζει
20 and and and και και

20 and και και και
20 This Το
20 In Το
20 European Ευρωπαϊκός
20 European Ευρωπαϊκή
19 within εντός
19 relating σχετικά
19 objectives στόχων
19 The The Το
19 III- III-
19 5 5
19 3 2
18 of of of της
18 members μέλη
18 also θα
18 A H
18 6 6
18 2 3
18 . στο .
17 to να να
17 other άλλων
17 of της της
17 in in τις
17 implementation εφαρμογή
17 common κοινή
17 be πρέπει
17 Member κράτους
17 Constitution Συντάγματος
16 this παρόντος
16 shall III- III-
16 qualified ειδική
16 paragraph παραγράφου
16 countries χώρες
16 and την
16 The The Η
16 In Οι
16 III- el
16 European European Ευρωπαϊκό
16 Council European Συμβούλιο
16 , , , , ,
15 particular ιδίως
15 may μπορούν
15 laws νόμος-
15 its αυτό
15 have ότι
15 cooperation συνεργασία
15 and in και
15 and and και και και
15 EU ΕΕ
15 Court Justice Δικαστήριο
15 AND ΚΑΙ
15))
14 with τις
14 unanimously ομόφωνα
14 to στην
14 shall or ή

14 prejudice επιφύλαξη
14 policy πολιτική
14 for for για
14 conditions όρους
14 common κοινής
14 are οι
14 any κάθε
14 and and την
14 and την και
14 The [
14 , , , , ,
13 with με με
13 shall Ένωσης
13 or . .
13 or ή ή
13 of of of των
13 institutions θεσμικά όργανα
13 c γ
13 any δεν
13 State μέλους
13 III III
13 : :
13 7 7
12 shall ή
12 proposal μετά πρόταση
12 principle αρχή
12 policy πολιτικής
12 criteria κριτήρια
12 cooperation συνεργασίας
12 cases περιπτώσεις
12 by by από
12 and and and and and και
12 The It Η
12 The . .
12 The Για
12 Court Δικαστήριο
12 , , a ,
11 will θα
11 under υπό
11 to to την
11 their τους τους
11 that . .
11 procedure διαδικασία
11 or shall ή
11 legislation including ήταν
11 laws νόμος
11 for για τη
11 decisions αποφάσεις
11 as πρέπει
11 Minister Foreign Affairs Υπουργός Εξωτερικών
11 European Ευρωπαϊκό Συμβούλιο
11 Constitution/ ΑΡΘΡΟ
11 Commission Commission Επιτροπή
11 Bank Τράπεζας
11 . να .
11 , τον ,

11 , , , , ,
10 rules κανόνες
10 principles αρχές
10 position θέση
10 paragraph παράγραφος
10 of της των
10 level επίπεδο
10 law νόμος
10 for για για
10 en III-
10 each κάθε
10 by από από
10 and και και και και
10 an μια
10 all όλα
10 action δράσης
10 Without Με
10 Union Union Ένωσης
10 TITLE ΤΙΤΛΟΣ
10 Member State κράτος μέλος
10 In Κατά
10 European European Ευρωπαϊκού
10 Bank Τράπεζα
10 Articles άρθρων
10 4 3
10 , , τον

Appendix B

Evaluation of Greek-English Dictionary: Empirical Study Phase I

Konstantinos Charitakis
e-mail: kons-cha [at] dsv.su.se

Following are listed five different tables. Each table has eight (8) columns and around a hundred (100) rows. The first row contains the index number of each translation pair. The second and third columns named *English* and *Greek* respectively contain the suggested English and Greek word translations. The rest five (5) columns named *Accurate*, *Somewhat Correct*, *Undecided*, *Somewhat Incorrect* and *Wrong*, contain the possible five choices A, B, C, D, and E including a checkbox field. You are asked to fill the appropriate checkbox for each suggested translation pair.

After completing the evaluation form electronically you should send back to Konstantinos by email, the following:

- A Filled-in [DICTIONARY_EVALUATION – NAME_SURNAME].DOC file where [NAME_SURENAME] should be your own name and surname, e.g "Dictionary Evaluation - Konstantinos Charitakis.doc"

Instructions

Choose only one of the five choices (A, B, C, D and E) and fill **only one** checkbox field for each suggested translation pair according to the following:

A. Accurate – if the suggested translation is accurate.
B. Somewhat correct – if the suggested translation is correct but not accurate.
C. Undecided – if the person evaluating cannot make a decision about the translation pair.
D. Somewhat incorrect – if the translation is not correct but similar to the correct.
E. Wrong – if the suggested translation is just plain wrong.

You can browse through and check or uncheck a box either by clicking in it using the mouse, or you can browse through using the keyboard's arrow buttons and check/uncheck a box by pressing the space bar.

Please use the test area to get familiar with checking the checkbox fields before you start with the evaluation.

----------**Test area** ----------

A☐ B☐ C☐ D☐ E☐

A☐ B☐ C☐ D☐ E☐

A☐ B☐ C☐ D☐ E☐

Examples for each choice:

A. Accurate – the suggested translation is accurate.
e.g. "Council" = "Συμβούλιο"
"internal market" = "εσωτερικής αγοράς"
"with third countries" = "με τρίτες χώρες"

B. Somewhat correct – the suggested translation is correct but not accurate.
e.g. "areas" = "τομείς"
"Court Justice" = "Δικαστήριο"

C. Undecided – person evaluating cannot make a decision about the translation pair.

D. Somewhat incorrect – the translation is not correct but similar to the correct.
e.g. "not any" = "δεν"
"acting majority" = "αποφασίζει πλειοψηφία"

E. Wrong - the suggested translation is just plain wrong.
e.g. "future" = "πρόθεση"
"employment" = "πρέπει"

Important notice

1. Always have in mind the domain of the parallel texts from which the dictionary has been derived from. The texts are taken from the web portal of the European Union.
2. **Only one** possible choice (A, B, C, D or E) is allowed for each translation pair.

Please evaluate and judge responsibly, your evaluation will seriously influence my work. If you have any doubts or questions you can always contact me for further explanations.

	English	Greek	Accurate	Somewhat Correct	Undecided	Somewhat Incorrect	Wrong
12.	.	.	A☐	B☐	C☐	D☐	E☐
13.	and	και	A☐	B☐	C☐	D☐	E☐
14.	ARTICLE	ΑΡΘΡΟ	A☐	B☐	C☐	D☐	E☐
15.	and	και	A☐	B☐	C☐	D☐	E☐
16.	Council	Συμβούλιο	A☐	B☐	C☐	D☐	E☐
17.	and	και	A☐	B☐	C☐	D☐	E☐
18.	Constitution/	Constitution/	A☐	B☐	C☐	D☐	E☐
19.	3	3	A☐	B☐	C☐	D☐	E☐
20.	of	της	A☐	B☐	C☐	D☐	E☐
21.	=]	id	A☐	B☐	C☐	D☐	E☐
22.	by	από	A☐	B☐	C☐	D☐	E☐
23.	III-	III-	A☐	B☐	C☐	D☐	E☐
24.	Union	Ένωσης	A☐	B☐	C☐	D☐	E☐
25.	framework	πλαίσιο	A☐	B☐	C☐	D☐	E☐
26.	The	Οι	A☐	B☐	C☐	D☐	E☐
27.	not	δεν	A☐	B☐	C☐	D☐	E☐
28.	I-	I-	A☐	B☐	C☐	D☐	E☐

29.	for	για	A☐	B☐	C☐	D☐	E☐
30.	0001	0001	A☐	B☐	C☐	D☐	E☐
31.	may	μπορεί	A☐	B☐	C☐	D☐	E☐
32.	Council	Συμβουλίου	A☐	B☐	C☐	D☐	E☐
33.	States	μέλη	A☐	B☐	C☐	D☐	E☐
34.	is	είναι	A☐	B☐	C☐	D☐	E☐
35.	Commission	Επιτροπής	A☐	B☐	C☐	D☐	E☐
36.	accordance	σύμφωνα	A☐	B☐	C☐	D☐	E☐
37.	Member	τα	A☐	B☐	C☐	D☐	E☐
38.	European	Ευρωπαϊκού	A☐	B☐	C☐	D☐	E☐
39.	their	τους	A☐	B☐	C☐	D☐	E☐
40.	of	των	A☐	B☐	C☐	D☐	E☐
41.	of	της	A☐	B☐	C☐	D☐	E☐
42.	right	δικαίωμα	A☐	B☐	C☐	D☐	E☐
43.	European	Ευρωπαϊκής	A☐	B☐	C☐	D☐	E☐
44.	to	να	A☐	B☐	C☐	D☐	E☐
45.	the	του	A☐	B☐	C☐	D☐	E☐
46.	adopt	εκδίδει	A☐	B☐	C☐	D☐	E☐
47.	Member	κράτη	A☐	B☐	C☐	D☐	E☐
48.	( )	η	A☐	B☐	C☐	D☐	E☐
49.	in and	και	A☐	B☐	C☐	D☐	E☐
50.	In	[	A☐	B☐	C☐	D☐	E☐
51.	II-	II-	A☐	B☐	C☐	D☐	E☐
52.	SECTION	ΤΜΗΜΑ	A☐	B☐	C☐	D☐	E☐
53.	shall .	.	A☐	B☐	C☐	D☐	E☐
54.	majority	πλειοψηφία	A☐	B☐	C☐	D☐	E☐
55.	be by	από	A☐	B☐	C☐	D☐	E☐
56.	Member States	κρατών μελών	A☐	B☐	C☐	D☐	E☐
57.	If	Εάν	A☐	B☐	C☐	D☐	E☐
58.	establish	θεσπίζει	A☐	B☐	C☐	D☐	E☐
59.	This	Το	A☐	B☐	C☐	D☐	E☐
60.	European	Ευρωπαϊκός	A☐	B☐	C☐	D☐	E☐
61.	European	Ευρωπαϊκή	A☐	B☐	C☐	D☐	E☐
62.	relating	σχετικά	A☐	B☐	C☐	D☐	E☐
63.	objectives	στόχων	A☐	B☐	C☐	D☐	E☐
64.	also	θα	A☐	B☐	C☐	D☐	E☐
65.	A	Η	A☐	B☐	C☐	D☐	E☐
66.	.	στο .	A☐	B☐	C☐	D☐	E☐
67.	other	άλλων	A☐	B☐	C☐	D☐	E☐
68.	of	της	A☐	B☐	C☐	D☐	E☐
69.	in	τις	A☐	B☐	C☐	D☐	E☐
70.	implementation	εφαρμογή	A☐	B☐	C☐	D☐	E☐
71.	be	πρέπει	A☐	B☐	C☐	D☐	E☐
72.	Member	κράτους	A☐	B☐	C☐	D☐	E☐
73.	Constitution	Συντάγματος	A☐	B☐	C☐	D☐	E☐
74.	this	παρόντος	A☐	B☐	C☐	D☐	E☐
75.	countries	χώρες	A☐	B☐	C☐	D☐	E☐
76.	and	την	A☐	B☐	C☐	D☐	E☐
77.	The	Η	A☐	B☐	C☐	D☐	E☐
78.	III-	el	A☐	B☐	C☐	D☐	E☐
79.	European	Ευρωπαϊκό	A☐	B☐	C☐	D☐	E☐
80.	,	,	A☐	B☐	C☐	D☐	E☐
81.	laws	νόμος-	A☐	B☐	C☐	D☐	E☐
82.	cooperation	συνεργασία	A☐	B☐	C☐	D☐	E☐
83.	and in	και	A☐	B☐	C☐	D☐	E☐
84.	and	και	A☐	B☐	C☐	D☐	E☐
85.	EU	ΕΕ	A☐	B☐	C☐	D☐	E☐
86.	Court Justice	Δικαστήριο	A☐	B☐	C☐	D☐	E☐
87.	with	τις	A☐	B☐	C☐	D☐	E☐
88.	unanimously	ομόφωνα	A☐	B☐	C☐	D☐	E☐
89.	for	για	A☐	B☐	C☐	D☐	E☐

90.	common	κοινής	A☐	B☐	C☐	D☐	E☐
91.	any	κάθε	A☐	B☐	C☐	D☐	E☐
92.	and	την και	A☐	B☐	C☐	D☐	E☐
93.	,	,	A☐	B☐	C☐	D☐	E☐
94.	with	με	A☐	B☐	C☐	D☐	E☐
95.	shall	Ένωσης	A☐	B☐	C☐	D☐	E☐
96.	of	των	A☐	B☐	C☐	D☐	E☐
97.	any	δεν	A☐	B☐	C☐	D☐	E☐
98.	III	III	A☐	B☐	C☐	D☐	E☐
99.	:	:	A☐	B☐	C☐	D☐	E☐
100.	proposal	μετά πρόταση	A☐	B☐	C☐	D☐	E☐
101.	criteria	κριτήρια	A☐	B☐	C☐	D☐	E☐
102.	cooperation	συνεργασίας	A☐	B☐	C☐	D☐	E☐
103.	by	από	A☐	B☐	C☐	D☐	E☐
104.	The	Για	A☐	B☐	C☐	D☐	E☐
105.	Court	Δικαστήριο	A☐	B☐	C☐	D☐	E☐
106.	, a	,	A☐	B☐	C☐	D☐	E☐
107.	under	υπό	A☐	B☐	C☐	D☐	E☐
108.	to	την	A☐	B☐	C☐	D☐	E☐
109.	their	τους	A☐	B☐	C☐	D☐	E☐
110.	procedure	διαδικασία	A☐	B☐	C☐	D☐	E☐
111.	or shall	ή	A☐	B☐	C☐	D☐	E☐

	English	**Greek**	**Accurate**	**Somewhat Correct**	**Undecided**	**Somewhat Incorrect**	**Wrong**
1.	working	8	A☐	B☐	C☐	D☐	E☐
2.	work	»	A☐	B☐	C☐	D☐	E☐
3.	which	μπορούν	A☐	B☐	C☐	D☐	E☐
4.	to	την	A☐	B☐	C☐	D☐	E☐
5.	three	Εξάλλου	A☐	B☐	C☐	D☐	E☐
6.	these	εκ	A☐	B☐	C☐	D☐	E☐
7.	the	που του	A☐	B☐	C☐	D☐	E☐
8.	the	του ,	A☐	B☐	C☐	D☐	E☐
9.	that .	.	A☐	B☐	C☐	D☐	E☐
10.	that	ότι δεν	A☐	B☐	C☐	D☐	E☐
11.	that	ισχύουν	A☐	B☐	C☐	D☐	E☐
12.	such	εν λόγω	A☐	B☐	C☐	D☐	E☐
13.	social	κοινωνικής	A☐	B☐	C☐	D☐	E☐
14.	should	πρέπει	A☐	B☐	C☐	D☐	E☐
15.	shall Member States	δεν κρατών μελών	A☐	B☐	C☐	D☐	E☐
16.	shall	κρατών μελών	A☐	B☐	C☐	D☐	E☐
17.	security	ασφάλεια	A☐	B☐	C☐	D☐	E☐
18.	second	δεύτερη	A☐	B☐	C☐	D☐	E☐
19.	same	βάση	A☐	B☐	C☐	D☐	E☐
20.	promoting	διαβούλευση	A☐	B☐	C☐	D☐	E☐
21.	positions	θέσεων	A☐	B☐	C☐	D☐	E☐
22.	policies	πολιτική	A☐	B☐	C☐	D☐	E☐
23.	paragraph	θεσπίζει	A☐	B☐	C☐	D☐	E☐
24.	own resources	ιδίων πόρων	A☐	B☐	C☐	D☐	E☐
25.	opinions	γνώμες	A☐	B☐	C☐	D☐	E☐
26.	of	της των	A☐	B☐	C☐	D☐	E☐
27.	of	των αφορά	A☐	B☐	C☐	D☐	E☐
28.	of	της προς	A☐	B☐	C☐	D☐	E☐
29.	not	εθνικά	A☐	B☐	C☐	D☐	E☐
30.	necessary	=	A☐	B☐	C☐	D☐	E☐
31.	liberalisation	ελευθέρωση	A☐	B☐	C☐	D☐	E☐
32.	legislation	εξέδωσε	A☐	B☐	C☐	D☐	E☐
33.	law	ευρωπαϊκός νόμος	A☐	B☐	C☐	D☐	E☐
34.	law	δικαίου	A☐	B☐	C☐	D☐	E☐
35.	its	πρωτοβουλία	A☐	B☐	C☐	D☐	E☐

36.	its	Ευρώπης	A☐	B☐	C☐	D☐	E☐
37.	issues	τη	A☐	B☐	C☐	D☐	E☐
38.	is on	για	A☐	B☐	C☐	D☐	E☐
39.	governing exercise	διέπουν άσκησή	A☐	B☐	C☐	D☐	E☐
40.	functioning	ευρωπαϊκή	A☐	B☐	C☐	D☐	E☐
41.	full	Συμβούλιο	A☐	B☐	C☐	D☐	E☐
42.	from	επίπεδο	A☐	B☐	C☐	D☐	E☐
43.	free	ελεύθερη	A☐	B☐	C☐	D☐	E☐
44.	for is	για	A☐	B☐	C☐	D☐	E☐
45.	employment	πρέπει	A☐	B☐	C☐	D☐	E☐
46.	does	όλες	A☐	B☐	C☐	D☐	E☐
47.	decision	ευρωπαϊκή απόφαση	A☐	B☐	C☐	D☐	E☐
48.	currency	καταστατικό	A☐	B☐	C☐	D☐	E☐
49.	coordinate	συντονίζουν	A☐	B☐	C☐	D☐	E☐
50.	cooperate	συνεργάζονται	A☐	B☐	C☐	D☐	E☐
51.	between	μετά	A☐	B☐	C☐	D☐	E☐
52.	as	κάθε	A☐	B☐	C☐	D☐	E☐
53.	approach	υλοποίηση	A☐	B☐	C☐	D☐	E☐
54.	and in	και	A☐	B☐	C☐	D☐	E☐
55.	already	τον	A☐	B☐	C☐	D☐	E☐
56.	aim	νέων	A☐	B☐	C☐	D☐	E☐
57.	actions	δράσεων	A☐	B☐	C☐	D☐	E☐
58.	acting majority	αποφασίζει πλειοψηφία	A☐	B☐	C☐	D☐	E☐
59.	a ,	,	A☐	B☐	C☐	D☐	E☐
60.	[Article	ΑΡΘΡΟ III-	A☐	B☐	C☐	D☐	E☐
61.	Union	Ένωση	A☐	B☐	C☐	D☐	E☐
62.	The	Η Το	A☐	B☐	C☐	D☐	E☐
63.	Statute	Οργανισμό	A☐	B☐	C☐	D☐	E☐
64.	Parliament	Κοινοβουλίου	A☐	B☐	C☐	D☐	E☐
65.	November	Απριλίου	A☐	B☐	C☐	D☐	E☐
66.	Member all	τα	A☐	B☐	C☐	D☐	E☐
67.	Member States shall	κρατών μελών	A☐	B☐	C☐	D☐	E☐
68.	January stood 1, 4 %	υιοθέτηση	A☐	B☐	C☐	D☐	E☐
69.	It shall act	ΑΡΘΡΟ III- 1	A☐	B☐	C☐	D☐	E☐
70.	In European	Το Ευρωπαϊκό	A☐	B☐	C☐	D☐	E☐
71.	Commission Community	Επιτροπής	A☐	B☐	C☐	D☐	E☐
72.	Central	Κεντρικής	A☐	B☐	C☐	D☐	E☐
73.	Article	Το	A☐	B☐	C☐	D☐	E☐
74.	Any	Κάθε	A☐	B☐	C☐	D☐	E☐
75.	10(	10	A☐	B☐	C☐	D☐	E☐
76.	,	, ορισμένες	A☐	B☐	C☐	D☐	E☐
77.	)	αυτόν	A☐	B☐	C☐	D☐	E☐
78.	ordinary	ευρωπαϊκοί συνήθη	A☐	B☐	C☐	D☐	E☐
79.	out	γίνει	A☐	B☐	C☐	D☐	E☐
80.	for	τον για περίοδο	A☐	B☐	C☐	D☐	E☐
81.	necessary measures	μέτρα	A☐	B☐	C☐	D☐	E☐
82.	between	μεταξύ	A☐	B☐	C☐	D☐	E☐
83.	an	γνώμη	A☐	B☐	C☐	D☐	E☐
84.	three months	τριών μηνών	A☐	B☐	C☐	D☐	E☐
85.	which	νέα	A☐	B☐	C☐	D☐	E☐
86.	report	έκθεση	A☐	B☐	C☐	D☐	E☐
87.	recommendation	σύσταση	A☐	B☐	C☐	D☐	E☐
88.	will	θα	A☐	B☐	C☐	D☐	E☐
89.	may	ΑΡΘΡΟ	A☐	B☐	C☐	D☐	E☐
90.	the	το του	A☐	B☐	C☐	D☐	E☐
91.	Court	Δικαστηρίου	A☐	B☐	C☐	D☐	E☐
92.	Commission .	.	A☐	B☐	C☐	D☐	E☐

93.	competent	αρμόδιες	A☐	B☐	C☐	D☐	E☐
94.	be as	πρέπει	A☐	B☐	C☐	D☐	E☐
95.	consent	έγκριση	A☐	B☐	C☐	D☐	E☐
96.	products	προϊόντα	A☐	B☐	C☐	D☐	E☐
97.	They	Τα	A☐	B☐	C☐	D☐	E☐
98.	Council	Συμβουλίου	A☐	B☐	C☐	D☐	E☐
99.	categories	κατηγορίες	A☐	B☐	C☐	D☐	E☐
100.	WITH THIRD COUNTRIES	ΜΕ ΤΡΙΤΕΣ ΧΩΡΕΣ	A☐	B☐	C☐	D☐	E☐

	English	Greek	Accurate	Somewhat Correct	Undecided	Somewhat Incorrect	Wrong
1.	rules	κανόνες	A☐	B☐	C☐	D☐	E☐
2.	principles	αρχές	A☐	B☐	C☐	D☐	E☐
3.	of	της των	A☐	B☐	C☐	D☐	E☐
4.	law	νόμος	A☐	B☐	C☐	D☐	E☐
5.	by	από	A☐	B☐	C☐	D☐	E☐
6.	and	και	A☐	B☐	C☐	D☐	E☐
7.	all	όλα	A☐	B☐	C☐	D☐	E☐
8.	action	δράσης	A☐	B☐	C☐	D☐	E☐
9.	Without	Με	A☐	B☐	C☐	D☐	E☐
10.	TITLE	ΤΙΤΛΟΣ	A☐	B☐	C☐	D☐	E☐
11.	Member State	κράτος μέλος	A☐	B☐	C☐	D☐	E☐
12.	In	Κατά	A☐	B☐	C☐	D☐	E☐
13.	European	Ευρωπαϊκού	A☐	B☐	C☐	D☐	E☐
14.	to	την	A☐	B☐	C☐	D☐	E☐
15.	subject	άρθρα	A☐	B☐	C☐	D☐	E☐
16.	referred Article	άρθρου	A☐	B☐	C☐	D☐	E☐
17.	protection	προστασία	A☐	B☐	C☐	D☐	E☐
18.	proposal	πρόταση	A☐	B☐	C☐	D☐	E☐
19.	for	τη	A☐	B☐	C☐	D☐	E☐
20.	development	ανάπτυξη	A☐	B☐	C☐	D☐	E☐
21.	competition	ανταγωνισμού	A☐	B☐	C☐	D☐	E☐
22.	areas	τομείς	A☐	B☐	C☐	D☐	E☐
23.	all	όλες	A☐	B☐	C☐	D☐	E☐
24.	a ,	,	A☐	B☐	C☐	D☐	E☐
25.	Subsection	Υποτμήμα	A☐	B☐	C☐	D☐	E☐
26.	States	στα μέλη	A☐	B☐	C☐	D☐	E☐
27.	State	κράτος	A☐	B☐	C☐	D☐	E☐
28.	European .	.	A☐	B☐	C☐	D☐	E☐
29.	Council	Συμβουλίου	A☐	B☐	C☐	D☐	E☐
30.	Articles	άρθρα	A☐	B☐	C☐	D☐	E☐
31.	third	τρίτες	A☐	B☐	C☐	D☐	E☐
32.	provisions	διατάξεις	A☐	B☐	C☐	D☐	E☐
33.	organisation	οργάνωση	A☐	B☐	C☐	D☐	E☐
34.	on	για	A☐	B☐	C☐	D☐	E☐
35.	national	εθνικές	A☐	B☐	C☐	D☐	E☐
36.	measures	μεταξύ	A☐	B☐	C☐	D☐	E☐
37.	may	δύναται	A☐	B☐	C☐	D☐	E☐
38.	is	είναι	A☐	B☐	C☐	D☐	E☐
39.	indents	επιτύχει	A☐	B☐	C☐	D☐	E☐
40.	in on	με	A☐	B☐	C☐	D☐	E☐
41.	freedom	ελευθερία	A☐	B☐	C☐	D☐	E☐
42.	average inflation rate	νομίσματος	A☐	B☐	C☐	D☐	E☐
43.	after consulting	μετά διαβούλευση	A☐	B☐	C☐	D☐	E☐
44.	after	λάβει	A☐	B☐	C☐	D☐	E☐
45.	The	Ο	A☐	B☐	C☐	D☐	E☐
46.	Statute ESCB	τιμή αναφοράς	A☐	B☐	C☐	D☐	E☐
47.	States	Τα μέλη	A☐	B☐	C☐	D☐	E☐
48.	Right	Δικαίωμα	A☐	B☐	C☐	D☐	E☐

49.	Everyone	Κάθε πρόσωπο	A☐	B☐	C☐	D☐	E☐
50.	European	Συμβούλιο	A☐	B☐	C☐	D☐	E☐
51.	where	ότι	A☐	B☐	C☐	D☐	E☐
52.	reading	ανάγνωση	A☐	B☐	C☐	D☐	E☐
53.	principles	αρχών	A☐	B☐	C☐	D☐	E☐
54.	period severe tensions central rate	νομοθεσία κεντρικής τράπεζας 107	A☐	B☐	C☐	D☐	E☐
55.	paragraph	παράγραφο	A☐	B☐	C☐	D☐	E☐
56.	opinion	γνώμη	A☐	B☐	C☐	D☐	E☐
57.	obtaining consent	αφού έγκριση	A☐	B☐	C☐	D☐	E☐
58.	may	μπορεί	A☐	B☐	C☐	D☐	E☐
59.	law	νόμο	A☐	B☐	C☐	D☐	E☐
60.	interests	συμφερόντων	A☐	B☐	C☐	D☐	E☐
61.	initiative	πρωτοβουλία	A☐	B☐	C☐	D☐	E☐
62.	include	περιλαμβάνει	A☐	B☐	C☐	D☐	E☐
63.	in	με	A☐	B☐	C☐	D☐	E☐
64.	in and	και	A☐	B☐	C☐	D☐	E☐
65.	functioning	λειτουργία	A☐	B☐	C☐	D☐	E☐
66.	external	εξωτερικής	A☐	B☐	C☐	D☐	E☐
67.	d	δ	A☐	B☐	C☐	D☐	E☐
68.	as	κάθε	A☐	B☐	C☐	D☐	E☐
69.	activities	Επιτροπή	A☐	B☐	C☐	D☐	E☐
70.	act	πράξη	A☐	B☐	C☐	D☐	E☐
71.	Union	Ένωση	A☐	B☐	C☐	D☐	E☐
72.	States	μέλη	A☐	B☐	C☐	D☐	E☐
73.	European Council	Συμβούλιο	A☐	B☐	C☐	D☐	E☐
74.	European	Ευρωπαϊκό Κοινοβούλιο	A☐	B☐	C☐	D☐	E☐
75.	Court Justice	Δικαστηρίου	A☐	B☐	C☐	D☐	E☐
76.	Council	Συμβούλιο	A☐	B☐	C☐	D☐	E☐
77.	ARTICLE	ΑΡΘΡΟ	A☐	B☐	C☐	D☐	E☐
78.	with	με	A☐	B☐	C☐	D☐	E☐
79.	will	θα	A☐	B☐	C☐	D☐	E☐
80.	two	δύο	A☐	B☐	C☐	D☐	E☐
81.	the	της	A☐	B☐	C☐	D☐	E☐
82.	simple	απλή	A☐	B☐	C☐	D☐	E☐
83.	should be	πρέπει	A☐	B☐	C☐	D☐	E☐
84.	shall be .	.	A☐	B☐	C☐	D☐	E☐
85.	services	υπηρεσιών	A☐	B☐	C☐	D☐	E☐
86.	rights	δικαιώματα	A☐	B☐	C☐	D☐	E☐
87.	referred paragraph	παραγράφου	A☐	B☐	C☐	D☐	E☐
88.	of	και	A☐	B☐	C☐	D☐	E☐
89.	not any	δεν	A☐	B☐	C☐	D☐	E☐
90.	level	κάθε	A☐	B☐	C☐	D☐	E☐
91.	in	τις	A☐	B☐	C☐	D☐	E☐
92.	field	τομέα	A☐	B☐	C☐	D☐	E☐
93.	convergence Treaty	σύγκλισης	A☐	B☐	C☐	D☐	E☐
94.	citizens	πολιτών	A☐	B☐	C☐	D☐	E☐
95.	by be	από	A☐	B☐	C☐	D☐	E☐
96.	areas	κράτη	A☐	B☐	C☐	D☐	E☐
97.	and	και	A☐	B☐	C☐	D☐	E☐
98.	Union	Ένωσης	A☐	B☐	C☐	D☐	E☐
99.	In	Στο	A☐	B☐	C☐	D☐	E☐
100.	I	Ι	A☐	B☐	C☐	D☐	E☐

	English	Greek	Accurate	Somewhat Correct	Undecided	Somewhat Incorrect	Wrong
1.	years	έτη	A☐	B☐	C☐	D☐	E☐
2.	with	με	A☐	B☐	C☐	D☐	E☐
3.	will	ότι	A☐	B☐	C☐	D☐	E☐
4.	will	με	A☐	B☐	C☐	D☐	E☐
5.	weeks	εβδομάδων	A☐	B☐	C☐	D☐	E☐
6.	was	μιας	A☐	B☐	C☐	D☐	E☐
7.	transport	μεταφορών	A☐	B☐	C☐	D☐	E☐
8.	trade	εμπόριο	A☐	B☐	C☐	D☐	E☐
9.	to	να την	A☐	B☐	C☐	D☐	E☐
10.	to	να	A☐	B☐	C☐	D☐	E☐
11.	the	του	A☐	B☐	C☐	D☐	E☐
12.	the	που	A☐	B☐	C☐	D☐	E☐
13.	the	που το	A☐	B☐	C☐	D☐	E☐
14.	that	ότι	A☐	B☐	C☐	D☐	E☐
15.	such	λόγω	A☐	B☐	C☐	D☐	E☐
16.	strengthening	τον	A☐	B☐	C☐	D☐	E☐
17.	shall members	μελών	A☐	B☐	C☐	D☐	E☐
18.	regulation	κανονισμού	A☐	B☐	C☐	D☐	E☐
19.	referred this	παρόντος	A☐	B☐	C☐	D☐	E☐
20.	production	παραγωγής	A☐	B☐	C☐	D☐	E☐
21.	potential	τον	A☐	B☐	C☐	D☐	E☐
22.	policies	πολιτικών	A☐	B☐	C☐	D☐	E☐
23.	out	έκθεση	A☐	B☐	C☐	D☐	E☐
24.	or	ή	A☐	B☐	C☐	D☐	E☐
25.	or	εάν	A☐	B☐	C☐	D☐	E☐
26.	of	της	A☐	B☐	C☐	D☐	E☐
27.	of	και	A☐	B☐	C☐	D☐	E☐
28.	of	της των	A☐	B☐	C☐	D☐	E☐
29.	of	στον της	A☐	B☐	C☐	D☐	E☐
30.	new	νέο	A☐	B☐	C☐	D☐	E☐
31.	necessary	Οικονομική	A☐	B☐	C☐	D☐	E☐
32.	member	μέλη	A☐	B☐	C☐	D☐	E☐
33.	meaning	έννοια	A☐	B☐	C☐	D☐	E☐
34.	market	αγορά	A☐	B☐	C☐	D☐	E☐
35.	laid ,	,	A☐	B☐	C☐	D☐	E☐
36.	international	διεθνών	A☐	B☐	C☐	D☐	E☐
37.	internal market	εσωτερικής αγοράς	A☐	B☐	C☐	D☐	E☐
38.	in and	και	A☐	B☐	C☐	D☐	E☐
39.	in .	.	A☐	B☐	C☐	D☐	E☐
40.	implementing	εφαρμογής	A☐	B☐	C☐	D☐	E☐
41.	has been the its	του το	A☐	B☐	C☐	D☐	E☐
42.	further	ΕΕ	A☐	B☐	C☐	D☐	E☐
43.	fully	πλήρως	A☐	B☐	C☐	D☐	E☐
44.	for on	για	A☐	B☐	C☐	D☐	E☐
45.	for .	να .	A☐	B☐	C☐	D☐	E☐
46.	following	Κοινοβούλιο	A☐	B☐	C☐	D☐	E☐
47.	financial	δημοσιονομικό	A☐	B☐	C☐	D☐	E☐
48.	en	Article	A☐	B☐	C☐	D☐	E☐
49.	employment	απασχόληση	A☐	B☐	C☐	D☐	E☐
50.	does not	δεν	A☐	B☐	C☐	D☐	E☐
51.	directly	άμεσα	A☐	B☐	C☐	D☐	E☐
52.	directive	ακόμα	A☐	B☐	C☐	D☐	E☐
53.	derogation	παρέκκλιση	A☐	B☐	C☐	D☐	E☐
54.	decision establishing	ευρωπαϊκή απόφαση	A☐	B☐	C☐	D☐	E☐
55.	decision	απόφαση	A☐	B☐	C☐	D☐	E☐
56.	consist	απαρτίζεται	A☐	B☐	C☐	D☐	E☐
57.	assisted	επικουρείται	A☐	B☐	C☐	D☐	E☐
58.	as	εντός	A☐	B☐	C☐	D☐	E☐
59.	as	δύο	A☐	B☐	C☐	D☐	E☐

60.	apply	έχουν	A☐	B☐	C☐	D☐	E☐
61.	any	δεν	A☐	B☐	C☐	D☐	E☐
62.	and in	και	A☐	B☐	C☐	D☐	E☐
63.	and	και	A☐	B☐	C☐	D☐	E☐
64.	agreement	μη	A☐	B☐	C☐	D☐	E☐
65.	affected	χαρακτήρα	A☐	B☐	C☐	D☐	E☐
66.	action	ειδικότερα	A☐	B☐	C☐	D☐	E☐
67.	a to	την	A☐	B☐	C☐	D☐	E☐
68.	a ,	,	A☐	B☐	C☐	D☐	E☐
69.	[	ΤΙΤΛΟΣ	A☐	B☐	C☐	D☐	E☐
70.	[	[=	A☐	B☐	C☐	D☐	E☐
71.	Union	Ένωση	A☐	B☐	C☐	D☐	E☐
72.	Union shall	Ένωσης	A☐	B☐	C☐	D☐	E☐
73.	Treaty	-	A☐	B☐	C☐	D☐	E☐
74.	This Commission	Επιτροπή	A☐	B☐	C☐	D☐	E☐
75.	This	Στην	A☐	B☐	C☐	D☐	E☐
76.	These	Το	A☐	B☐	C☐	D☐	E☐
77.	The In	Το	A☐	B☐	C☐	D☐	E☐
78.	The Council	Το	A☐	B☐	C☐	D☐	E☐
79.	The	[Η	A☐	B☐	C☐	D☐	E☐
80.	State	κράτος μέλος	A☐	B☐	C☐	D☐	E☐
81.	Regarding fulfilment mentioned four Article 109 j(	υψηλό βαθμό σταθερής γνώμονα τέσσερα	A☐	B☐	C☐	D☐	E☐
82.	Parliament its	Κοινοβούλιο	A☐	B☐	C☐	D☐	E☐
83.	Member	κράτος μέλος	A☐	B☐	C☐	D☐	E☐
84.	It The	Η	A☐	B☐	C☐	D☐	E☐
85.	In	7	A☐	B☐	C☐	D☐	E☐
86.	If	Το	A☐	B☐	C☐	D☐	E☐
87.	I-	I-	A☐	B☐	C☐	D☐	E☐
88.	Fund	Ταμείο	A☐	B☐	C☐	D☐	E☐
89.	European	Ευρωπαϊκή	A☐	B☐	C☐	D☐	E☐
90.	European	Ευρώπης	A☐	B☐	C☐	D☐	E☐
91.	Council	Το Συμβούλιο	A☐	B☐	C☐	D☐	E☐
92.	Constitution/	[id	A☐	B☐	C☐	D☐	E☐
93.	Committee	Επιτροπή	A☐	B☐	C☐	D☐	E☐
94.	-	Στην	A☐	B☐	C☐	D☐	E☐
95.	, a	,	A☐	B☐	C☐	D☐	E☐
96.	,	τον ,	A☐	B☐	C☐	D☐	E☐
97.	( )	η	A☐	B☐	C☐	D☐	E☐
98.	( )	)	A☐	B☐	C☐	D☐	E☐
99.							

	English	**Greek**	**Accurate**	**Somewhat correct**	**Undecided**	**Somewhat Incorrect**	**Wrong**
1.	which	οποία	A☐	B☐	C☐	D☐	E☐
2.	were	2005	A☐	B☐	C☐	D☐	E☐
3.	to	να	A☐	B☐	C☐	D☐	E☐
4.	to . for	στην σε την	A☐	B☐	C☐	D☐	E☐
5.	to	να σε	A☐	B☐	C☐	D☐	E☐
6.	this	τρίτων	A☐	B☐	C☐	D☐	E☐
7.	this	σκοπό	A☐	B☐	C☐	D☐	E☐
8.	this	αυτό	A☐	B☐	C☐	D☐	E☐
9.	third	τρίτων	A☐	B☐	C☐	D☐	E☐
10.	the	το του	A☐	B☐	C☐	D☐	E☐
11.	the of	του	A☐	B☐	C☐	D☐	E☐
12.	the by	το	A☐	B☐	C☐	D☐	E☐
13.	that	προϊόντων	A☐	B☐	C☐	D☐	E☐
14.	territory	επικράτεια	A☐	B☐	C☐	D☐	E☐

15.	take	υπόψη	A☐	B☐	C☐	D☐	E☐
16.	shall .	στο .	A☐	B☐	C☐	D☐	E☐
17.	shall	καθορίζει	A☐	B☐	C☐	D☐	E☐
18.	shall	ΑΡΘΡΟ III-	A☐	B☐	C☐	D☐	E☐
19.	second	δεύτερο	A☐	B☐	C☐	D☐	E☐
20.	referred	προβλέπεται	A☐	B☐	C☐	D☐	E☐
21.	reference four	χαμηλότερο	A☐	B☐	C☐	D☐	E☐
22.	recommendations	συστάσεις	A☐	B☐	C☐	D☐	E☐
23.	pursuant	βάσει	A☐	B☐	C☐	D☐	E☐
24.	prohibited	απαγορεύονται	A☐	B☐	C☐	D☐	E☐
25.	programmes	προγραμμάτων	A☐	B☐	C☐	D☐	E☐
26.	policy	πολιτικής	A☐	B☐	C☐	D☐	E☐
27.	particular	εν λόγω	A☐	B☐	C☐	D☐	E☐
28.	or	ή	A☐	B☐	C☐	D☐	E☐
29.	or	είτε	A☐	B☐	C☐	D☐	E☐
30.	on for	για	A☐	B☐	C☐	D☐	E☐
31.	of	των	A☐	B☐	C☐	D☐	E☐
32.	of	και	A☐	B☐	C☐	D☐	E☐
33.	objectives	στόχοι	A☐	B☐	C☐	D☐	E☐
34.	not	δεν	A☐	B☐	C☐	D☐	E☐
35.	necessary	αναγκαίες	A☐	B☐	C☐	D☐	E☐
36.	member ERM last two years not not bilateral against State'	εθνική καταστατικού 108 συνθήκης ΕΣΚΤ	A☐	B☐	C☐	D☐	E☐
37.	majority	πλειοψηφία	A☐	B☐	C☐	D☐	E☐
38.	limits	ορίων	A☐	B☐	C☐	D☐	E☐
39.	legislative	νομοθετική	A☐	B☐	C☐	D☐	E☐
40.	is for	για	A☐	B☐	C☐	D☐	E☐
41.	internal market	εσωτερική αγορά	A☐	B☐	C☐	D☐	E☐
42.	instruments	(	A☐	B☐	C☐	D☐	E☐
43.	industry	θα	A☐	B☐	C☐	D☐	E☐
44.	in	με	A☐	B☐	C☐	D☐	E☐
45.	in	και	A☐	B☐	C☐	D☐	E☐
46.	for	για τη	A☐	B☐	C☐	D☐	E☐
47.	for	τη	A☐	B☐	C☐	D☐	E☐
48.	devalued own initiative	Συμπεριλαμβανομένου εθνικής	A☐	B☐	C☐	D☐	E☐
49.	decisions	ευρωπαϊκές αποφάσεις	A☐	B☐	C☐	D☐	E☐
50.	data	δεδομένων	A☐	B☐	C☐	D☐	E☐
51.	component	όλου αριθμού	A☐	B☐	C☐	D☐	E☐
52.	citizens	πολίτες	A☐	B☐	C☐	D☐	E☐
53.	bodies	οργανισμών	A☐	B☐	C☐	D☐	E☐
54.	be from	από	A☐	B☐	C☐	D☐	E☐
55.	be .	στο .	A☐	B☐	C☐	D☐	E☐
56.	be	από	A☐	B☐	C☐	D☐	E☐
57.	authorities	αρχών	A☐	B☐	C☐	D☐	E☐
58.	at	ΕΕ	A☐	B☐	C☐	D☐	E☐
59.	assistance	αρχές	A☐	B☐	C☐	D☐	E☐
60.	areas	επίπεδο	A☐	B☐	C☐	D☐	E☐
61.	and	και την	A☐	B☐	C☐	D☐	E☐
62.	an	υπηρεσιών	A☐	B☐	C☐	D☐	E☐
63.	all the	το έτος που	A☐	B☐	C☐	D☐	E☐
64.	aid	βοήθειας	A☐	B☐	C☐	D☐	E☐
65.	agreement	συμφωνία	A☐	B☐	C☐	D☐	E☐
66.	adopted	θεσπίζονται	A☐	B☐	C☐	D☐	E☐
67.	access	πρόσβαση	A☐	B☐	C☐	D☐	E☐
68.	Within	Στο	A☐	B☐	C☐	D☐	E☐
69.	Where	Όταν	A☐	B☐	C☐	D☐	E☐
70.	Union'	Ένωση	A☐	B☐	C☐	D☐	E☐
71.	Treaty	Συνθήκης	A☐	B☐	C☐	D☐	E☐

72.	This	0001	A☐	B☐	C☐	D☐	E☐
73.	These	Οι	A☐	B☐	C☐	D☐	E☐
74.	The	Οι	A☐	B☐	C☐	D☐	E☐
75.	The In	Η	A☐	B☐	C☐	D☐	E☐
76.	The	Το Ευρωπαϊκό Συμβούλιο	A☐	B☐	C☐	D☐	E☐
77.	State	κράτος μέλος	A☐	B☐	C☐	D☐	E☐
78.	PROVISIONS	ΔΙΑΤΑΞΕΙΣ	A☐	B☐	C☐	D☐	E☐
79.	On	Στο	A☐	B☐	C☐	D☐	E☐
80.	Minister Foreign Affairs	Υπουργό Εξωτερικών	A☐	B☐	C☐	D☐	E☐
81.	Member States	κρατών μελών	A☐	B☐	C☐	D☐	E☐
82.	Member States	κρατών	A☐	B☐	C☐	D☐	E☐
83.	Member	τα κράτη	A☐	B☐	C☐	D☐	E☐
84.	Member	κράτη	A☐	B☐	C☐	D☐	E☐
85.	January	Ιανουάριο	A☐	B☐	C☐	D☐	E☐
86.	It	Για	A☐	B☐	C☐	D☐	E☐
87.	In Treaty	-	A☐	B☐	C☐	D☐	E☐
88.	In	Πρέπει	A☐	B☐	C☐	D☐	E☐
89.	General	Γενικό	A☐	B☐	C☐	D☐	E☐
90.	Freedom	Ελευθερία	A☐	B☐	C☐	D☐	E☐
91.	Every	Κάθε	A☐	B☐	C☐	D☐	E☐
92.	European	Ευρωπαϊκό	A☐	B☐	C☐	D☐	E☐
93.	European	Ευρωπαϊκή	A☐	B☐	C☐	D☐	E☐
94.	European	Το Ευρωπαϊκό	A☐	B☐	C☐	D☐	E☐
95.	EDPS	ΕΕΠΔ	A☐	B☐	C☐	D☐	E☐
96.	Council	Συμβούλιο	A☐	B☐	C☐	D☐	E☐
97.	Commission adopted	Επιτροπή	A☐	B☐	C☐	D☐	E☐
98.	Commission	Επιτροπή	A☐	B☐	C☐	D☐	E☐
99.	By way derogation	Κατά παρέκκλιση	A☐	B☐	C☐	D☐	E☐
100.	Article	Άρθρο	A☐	B☐	C☐	D☐	E☐

Note:

- Don’t forget to save the results after completing the evaluation.
File -> Save As…-> [DICTIONARY_EVALUATION – NAME_SURNAME].DOC

- Send the filled form to: kons-cha [at] dsv.su.se

Thank you for your time!