Abstract
This paper describes a method for the automatic construction of a translation lexicon from hybrid parallel texts. The method features mainly statistical techniques, while effort has been made to invoke the least required language specific information in an attempt to implement as language independent a method as possible. The method presupposes parallel texts and identifies translational equivalences at the word or multi-word unit level for those cases that such an equivalence holds true. Parallel texts are first statistically aligned at sentence level and then lemmatised and tagged with their part-of-speech categories. Noun phrase grammars operating on tag sequences extract noun phrases on both sides of the parallel corpus and statistical evaluation yields the most coherent, statistically relevant multi-word-units on either side. Translation candidates of word- or multi-word units are then evaluated by a similarity metric defined by the co-occurrence frequency and independent frequencies of the units. The method has been tested on a small English-Greek corpus consisting of texts relevant to software systems, yielding approximately 94% accurate translations, while proposing translations for approximately 50% of all occurrences of content words of the English side and 40% of the Greek side of the corpus.

1 Introduction
The availability of parallel texts in electronic form has given rise to a wide range of applications aiming at the elicitation of linguistic resources such as translation dictionaries, transfer grammars and retrieval of translation examples [Dagan et al., 1991], [Matsumoto et al., 1993] or even the building of fully-blown machine translation systems [Brown et al., 1990].

Statistical processing has proved powerful for extraction of translation equivalences at sentence [Gale and Church, 1991b] and intra-sentence level. The purpose of this paper is to describe a technique for extracting a translation lexicon at the word and multi-word unit level employing statistical techniques coupled with shallow linguistic processing that caters for the translation unit variation from single word to a sequence of words, thus achieving better performance. [Brown, 1988] uses a probabilistic measure to estimate word similarity of two languages in the context of statistically-based machine translation.[Kay, 1991] presents an algorithm for aligning bilingual texts on the basis of internal evidence only. This algorithm can be used to produce both sentence alignments and word alignments. Processing is performed in many iterations and each new iteration uses the results of the previous one in order to calculate more accurate word and sentence correspondences. In each iteration, processing consists of calculating correspondences between sentences on the basis of their relative positions, and then calculating word correspondences on the basis of words’ co-occurrences in related sentences. [Kay and Roescheisen, 1993] proposed the use of the Dice coefficient as the similarity measure between words of two languages in an attempt to secure the correctness of the alignment of parallel texts at sentence level. [Kitamura and Matsumoto, 1995] have used the same Dice coefficient to calculate the word similarity between Japanese-English parallel corpora. Single word correspondences have also been investigated by [Gale and Church, 1991] using a statistical evaluation of contingency tables. [Boutsis and Piperidis, 1996] describe a method for extracting single word equivalences based on a parallel corpus statistically aligned at sentence level and employing a similarity metric along the lines of the Dice coefficient with comparable performance.

Collocational correspondences have been studied by [Smadja, 1992] and [Smadja et al., 1996], in an attempt to find translation patterns for continuous and discontinuous collocations in English and French. Meaningful collocations are first extracted in the source language while their corresponding French ones are found by calculating the mutual information between instances of the English collocation and various single word candidates in English-French aligned corpora.

More recent work has broadened the scope identifying correspondences between word sequences. [Kupiec, 1993] proposes a method for extracting translation patterns of noun phrases from English-French parallel corpora. The corpus is tagged at pos level and then finite-state recognisers specified by regular expressions defined in terms of part-of-speech categories detect noun phrases on
either side. Probabilities of correspondences are then calculated using an iterative EM-like algorithm. [Kumano and Hirakawa, 1994] presuppose an ordinary bilingual dictionary and non-parallel corpora, attempting to find bilingual correspondences in a Japanese-English setting at word, noun phrase and unknown word level. Extending previous work [Kitamura and Matsumoto, 1996] apply the Dice coefficient on word sequence correspondence extraction.

In this paper we propose a method that processes bilingual parallel texts which are first aligned at sentence level. It then uses statistical and linguistic techniques and examines word and multi-word patterns at both language-sides of the corpus in order to extract word and multi-word associations between the two texts. The basic assumption we make is that pairs of words which are translations of each other usually appear in corresponding places in the source and target texts. Corresponding places means corresponding sentences or corresponding groups of sentences. Of course, during the translation process the order of the words does not remain the same nor is the correspondence of the words one to one. In addition, the impact of the context is quite strong and many words change their meaning in different contexts. The proposed method can deal effectively with words and word sequences that are translated with some reasonable consistency from one language into another. Since terms are usually translated in such a way, texts containing terminology from the same field are expected to perform well regarding the automatic extraction of word equivalences.

2 Method Overview
Extraction of bilingual word correspondences is performed through a set of pipelined standalone procedures. The output of one module is channelled to the next. Processing takes at the input the parallel texts in both languages and gives at the output the identified multilingual equivalences. During the first stage parallel texts are aligned at the sentence level using batch statistical techniques. Tagging and lemmatization are performed next in order to annotate wordforms with part of speech information and subsequently reduce each wordform to the proper lemma. Noun phrase extraction and statistical evaluation of the identified compounds take place at the third stage. The identified compound terms are brought to a canonical form and are highlighted in the text. During combinatorial processing lexical units are associated with their possible translations and a metric is calculated, expressing the validity of the formed associations. At the final stage a rule based selection takes place, giving at the output word correspondences that may be of type “1-1” or “n-m”.

3 Test Corpus
The corpus used to develop and test the proposed algorithms consists of text from the HP-VUE software platform documentation set.

The Greek text contains 35726 wordforms and the English text 28872. The number of different words is 4512 for the Greek text and 3219 for the English text. The richer morphology of the Greek language accounts for the approximately 30% difference between these two figures. In addition, the Greek text contains 2588 lemmas and the English text 2111 lemmas.

As far as corpus characteristics are concerned, translation quality from source into target language is obviously of importance. Free translation inserts noise that is expected to greatly affect performance. It is also of importance that terms are consistently translated throughout the text since the method reaches its limits when having to deal with many alternative translations for each word. Since terminology is usually translated in a consistent way, technical texts from the same domain, which is the case for our test corpus, should perform well when processed with the method.

4 Description of the Method
4.1 Sentence alignment
Sentence alignment is the first step towards the creation of the probabilistic lexicon. The bilingual corpus is reorganized into a bitext of mutual translations.

In general the appearance of an alignment depends on its resolution, i.e. on the nature of the units on which the segmentation is done. The resolution of alignment can vary from low to high: section, paragraph, sentence, phrase. The type of alignment at this stage, takes the sentence to be the segmentation unit.

The alignment algorithm is based on a simple statistical model of character lengths [Gale and Church, 1991b]. The
The grammar governing English noun phrase structure is as follows:

- Texts are treated independently of the other.
- Stage is monolingual so that each language side of the bilingual corpus is statistically processed in order to test termhood and grammatical patterns. Extracted noun phrases are to identify noun phrases following characteristic for terms.
- Part of speech linguistic knowledge is exploited in order to detect tagging errors.

The English tagger first assigns all the possible word-sense to the Greek language. It employs stochastic information to disambiguate those that are not found in the lexicon, a morphological analysis is carried out to predict their possible tags. The tagger then associates them against a lexicon of word forms. In the case of words not found in the lexicon (dictionary lookup). In the case of multiple tagging, the tagger selects the correct attribute by context-based rules. The tagset used has been devised for the tagging, the tagger selects the correct attribute by context-based rules.

The success rate of this method is considerably high as it computes almost 95% of the couples of the correct alignments in batch mode.

4.2 POS Tagging and Lemmatization

Both English and Greek texts are analyzed morphosyntactically. The words in the parallel sentences are tagged with their corresponding lemmas and part-of-speech (pos) categories. The corpus is thus represented as a bitext of tagged mutual translations where every word is accompanied by its corresponding lemma and part-of-speech tag.

Tagging for Greek is based on a morphological lexicon and a rule-based module for disambiguation purposes. It endows the words of the text with the characteristic found in the lexicon (dictionary lookup). In the case of multiple tagging, the tagger selects the correct attribute by context-based rules. The tagset used has been devised for the morphological annotation of Greek corpora and conforms to the guidelines set up by TEI and NERC, trying, at the same time, to capture the morphological particularities of the Greek language.

The English tagger first assigns all the possible word-class tags to the lexical items in a given text by checking them against a lexicon of word forms. In the case of words not found in the lexicon, a morphological analysis is carried out to predict their possible tags. The tagger then employs stochastic information to disambiguate those that have multiple tags. The sentence, as a sequence of tags, is confronted with a set of context-frame rules to filter out tagging errors.

4.3 Noun Phrase Extraction

Part of speech linguistic knowledge is exploited in order to identify noun phrases following characteristic for terms, grammatical patterns. Extracted noun phrases are statistically processed in order to test termhood and identify compounds likely to represent valid terms of the technical domain covered by the texts. Processing at this stage is monolingual so that each language side of the texts is treated independently of the other.

The grammar governing English noun phrase structure is as follows:

\[ \text{ng} \rightarrow [n]^+, \text{adjg} \rightarrow [adj]^+, \]  
\[ \text{np} \rightarrow (\text{adjg}) \text{ng} \rightarrow \text{np prep np} \]

Greek noun phrases are described by:

\[ \text{ng} \rightarrow [n]^+, \text{adjg} \rightarrow [adj]^+, \]  
\[ \text{np} \rightarrow (\text{adjg}) \text{ng} \rightarrow \text{np det case genitive}) \text{np case genitive}. \]

Enclosure in ( ) denotes optionality and [ ]+ indicates repetition of the enclosed expression one or more times.

In case of overgeneration where several noun phrases of different lengths can be extracted from the same text chunk, all alternatives are considered. That is all possible noun phrases, ranging from single nouns, minimal noun phrases, to maximal noun phrases, are thus identified.

After noun phrases have been extracted in accordance to the above grammar patterns, additional statistical processing evaluates the validity of the extracted compounds and tests their coherence. Cases of compounds with four or more words are not treated.

For each compound \( w_{1}^{n} = w_{1} \ldots w_{n} \), \( P(w_{1}^{n}) \) is computed with the following formula for n-gram probability estimation:

\[ P(w_{1}^{n}) = P(w_{1}^{n-1}) \cdot P(w_{1}^{n-1}) \]  
\[ = \frac{C(w_{1}^{n-1})}{\sum_{w} C(w_{1}^{n-1})} \cdot P(w_{1}^{n-1}) \]

where \( C(w) \) is the number of times string \( w \) occurs in the text. Most probable compounds are selected and channelled to the next stages.

4.4 Combinatorial Processing

The basic idea for the extraction of bilingual word equivalences is that pairs of words which are translations of each other usually appear in corresponding sentences or groups of sentences of the source and target text. Let’s assume, for instance, that within a sublanguage, word \( x \) is consistently translated into word \( y \). It follows that, in most cases, whenever \( x \) is encountered \( y \) is expected to occur in the corresponding sentence of the target language text. The correspondences between sentences have already been computed at the text alignment stage.

Let’s assume the source sentence \( S_{j,1}^{m} = s_{1j}, s_{2j}, \ldots, x, \ldots, s_{y} \) and the target sentence \( T_{j,1}^{m} = t_{1j}, t_{2j}, \ldots, y, \ldots, t_{mj} \).

The mapping function \( \varphi_{j} \) gets non zero values equal to 1 only for:

\[ \varphi_{j}(s_{1j},t_{1j}), \varphi_{j}(s_{2j},t_{2j}), \ldots, \varphi_{j}(s_{y},t_{y}), \ldots, \varphi_{j}(s_{m},t_{mj}) \]

In addition, let \( \Phi(x,y) \) be:

\[ \Phi(x,y) = \sum_{j} \varphi_{j}(x,y) \]

so that \( \Phi(x,y) \) expresses the association between \( x \) and \( y \) in the setting of the bilingual corpus, calculating the number of their simultaneous occurrences. Considering that \( x \) and \( y \) do not always appear in the same context we realise that \( \Phi \) should have a local maximum at point \( (x,y) \). In this way the correct translation \( x-y \) emerges under the
Note also that:
\[ \forall j, x, y \left[ \varphi_j(x) \geq \varphi_j(x, y), \varphi_j(y) \geq \varphi_j(x, y) \right], \]
where \( \varphi_j(x) \) is \( x \)'s frequency in the \( j \) sentence in the corpus, denoting that if a translation unit \( x \) is translated into a unit \( y \), the number of \( x \) occurrences in a given sentence of the source text is not always the same as the number of \( y \) occurrences in the corresponding sentence of the target text.

In this step, the following issues should also be dealt with:

- Functional Words. Since we are not up to get equivalences between functional words, we form exclusion lists in order to minimise computer processing time and size of generated combinations.
- Multi-Word-Units. Since noun phrases likely to represent multi-word-units have been identified in the previous stage, processing in this stage is aware of their existence and treats them as one unit.

### 4.5 Metric Calculation

After possible associations between the words of the two texts have been established, a mechanism to filter out the best ones is necessary. To this end, frequencies of words and multi-words of the source and target texts, and the association function \( \Phi \) are taken into account in order to calculate a metric that will be used to select the most reliable translations for each word.

A metric popular in lexicography is mutual information:
\[
MI(x, y) = \log \frac{P(x, y)}{P(x) P(y)},
\]
this metric however is not expected to perform well, in the context of this application. This is because, in practice, it tends to associate infrequent words with frequent ones: for a given \( x \), the fraction \( P(x,y)/P(y) \) gets its maximum value for those \( y \) that have in the corpus frequency \( \varphi(y) = \Phi(x,y) \) (since \( \forall x, y \left[ \varphi(y) \geq \varphi(x, y) \right] \)). This can be easily the case for an infrequent \( y \) that happens to occur always in the context of \( x \).

Given the considerations outlined above, the following observation is valid and can guide the formulation of a new metric:

"If \( x-y \) is a valid bilingual word equivalence, then the frequencies \( \varphi(x) \) and \( \varphi(y) \) should have similar values. The association function \( \Phi(x,y) \) should be of the same numeric order too."

We introduce the metric:
\[
M(x, y) = \frac{\left( \varphi(x) - m \right)^2 + \left( \varphi(y) - m \right)^2 + \left( \Phi(x, y) - m \right)^2}{m} \]  \tag{6},
\[
\text{where } m = \frac{\varphi(x) + \varphi(y) + \Phi(x, y)}{3} \]  \tag{7}.

This metric yields small values when \( \varphi(x) \), \( \varphi(y) \) and \( \Phi(x,y) \) are close to one another. When \( \varphi(x) \) and \( \varphi(y) \) are not close, or even when \( \varphi(x) \) and \( \varphi(y) \) are similar but \( \Phi(x,y) \) differs substantially, the metric produces large values. This enables us to choose best translation candidates. For each word, top candidates only are channelled to the next stage.

For example, the Greek trigram "ΑΡΧΕΙΟ-ΒΑΣΗ-ΔΕΔΟΜΕΝΟ" appearing 20 times in the corpus has been associated with the words of the following table:

<table>
<thead>
<tr>
<th>Score</th>
<th>Score % Variation</th>
<th>( \Phi(x,y) )</th>
<th>Target Word</th>
<th>Target Word Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.205</td>
<td>0.000000</td>
<td>15.00</td>
<td>DATABASE-FILE</td>
<td>17</td>
</tr>
<tr>
<td>0.690</td>
<td>70.246948</td>
<td>7.00</td>
<td>HP-VUE-SYNTAX</td>
<td>13</td>
</tr>
<tr>
<td>0.720</td>
<td>4.162519</td>
<td>8.00</td>
<td>CONVERT</td>
<td>26</td>
</tr>
<tr>
<td>0.994</td>
<td>27.614494</td>
<td>3.00</td>
<td>SEARCH</td>
<td>23</td>
</tr>
<tr>
<td>1.049</td>
<td>5.238557</td>
<td>2.00</td>
<td>LOCATE</td>
<td>20</td>
</tr>
<tr>
<td>1.055</td>
<td>0.519161</td>
<td>2.00</td>
<td>DIALOG-BOX</td>
<td>16</td>
</tr>
<tr>
<td>1.065</td>
<td>0.954310</td>
<td>2.00</td>
<td>APPEAR</td>
<td>15</td>
</tr>
<tr>
<td>1.135</td>
<td>6.140458</td>
<td>1.00</td>
<td>EASY</td>
<td>20</td>
</tr>
<tr>
<td>1.143</td>
<td>0.695868</td>
<td>3.00</td>
<td>EXISTING</td>
<td>9</td>
</tr>
<tr>
<td>1.157</td>
<td>0.580395</td>
<td>2.00</td>
<td>NEW-ACTION</td>
<td>11</td>
</tr>
</tbody>
</table>

The first column contains M-metric scores calculated for "ΑΡΧΕΙΟ-ΒΑΣΗ-ΔΕΔΟΜΕΝΟ" and the corresponding target word of the 4th column. The second column contains the percentage of score difference between a row and the previous one. The next column contains the value of the \( \Phi \) function for "ΑΡΧΕΙΟ-ΒΑΣΗ-ΔΕΔΟΜΕΝΟ" and the corresponding target word. Last column presents target words frequencies.

In the above example, we observe that score’s value gets its minimum for the case of the English word "DATABASE-FILE". The percentage of score difference between the first and the second line is \((0.690-0.205)/0.690 = 100\text{%/69.04}\%\); it is high enough to allow us to decide that "ΑΡΧΕΙΟ-ΒΑΣΗ-ΔΕΔΟΜΕΝΟ" = "DATABASE-FILE" is a valid bilingual unit equivalence.

However, for most units, we cannot reach such a conclusion. One such example is the English word "AUDIO". This word appears 14 times throughout the corpus and is translated:

- 7.00 times: "ΗΧΟΣ"  
- 4 times: "ΗΧΗΠΙΚΟΣ"  
- 4 times: "AUDIO" (foreign word)

once it is not translated at all

Obviously, "AUDIO" is not translated consistently. We expect that the calculated scores will not allow the identification of any translation for this word.

In fact, the table of the best candidates looks like:
same English unit. word, is mapped to a Greek unit that is mapped to the produced only when one English unit, single or multi-
it should be noted that a valid bilingual equivalence is sent to the output. If testing of expressions (8) - (10) fails for either the source or target word, latter test succeeds as well, a valid bilingual word checked for mapping to the current source word. If this case, the table associated with the target word is later), a one-to-one correspondence is likely to exist. In specific predetermined value ranges (to be discussed of the last section.

4.6 Rule-Based Selection

The purpose of this stage is to select for as many words and compounds of the source text as possible, a word or compound of the target text so that the formed bilingual pair can be a valid bilingual equivalence.

This stage takes as input two indexed files containing the output of the previous stage i.e. the best associations for every Greek and English element. The structure of the information passed to this stage is displayed in the tables of the last section.

Although in general translation correspondences can be one-to-one or many-to-many, yet the algorithm allows for one-to-one correspondences only. Correspondences containing more than one words on either side of the relation can be produced only by translating the multi-words identified during the noun phrase parsing phase. The procedure for unit correspondence extraction is as follows: for each word of the source language (Greek), the table containing the associated target language (English) words is recalled from the database. The table is examined and selection is guided by the following expressions for each bilingual pair:

\[
\text{Source word’s frequency} / \text{pair’s frequency} \quad (8) \\
\text{Target word’s frequency} / \text{pair’s frequency} \quad (9) \\
\text{Pair’s score} / \text{next pair’s score} \quad (10)
\]

If these expressions evaluated for the first pair lie within specific predetermined value ranges (to be discussed later), a one-to-one correspondence is likely to exist. In this case, the table associated with the target word is recalled, and using the same procedure the target word is checked for mapping to the current source word. If this latter test succeeds as well, a valid bilingual word equivalence is sent to the output. If testing of expressions (8) - (10) fails for either the source or target word, processing continues with the next source word.

It should be noted that a valid bilingual equivalence is produced only when one English unit, single or multi-word, is mapped to a Greek unit that is mapped to the same English unit.

<table>
<thead>
<tr>
<th>Score</th>
<th>Score % Variation</th>
<th>(x,y)</th>
<th>Target Word</th>
<th>Target Word Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.765</td>
<td>0.0000000</td>
<td>4.00</td>
<td>ΗΧΩΣ</td>
<td>14</td>
</tr>
<tr>
<td>0.918</td>
<td>16.666672</td>
<td>5.00</td>
<td>ΗΧΗΤΙΚΟΣ</td>
<td>5</td>
</tr>
<tr>
<td>1.059</td>
<td>1.892294</td>
<td>2.00</td>
<td>ΥΠΟΛΟΓΙΣΤΗΣ</td>
<td>19</td>
</tr>
<tr>
<td>1.096</td>
<td>3.382034</td>
<td>1.00</td>
<td>ΣΗΜΕΙΩΣΗ</td>
<td>13</td>
</tr>
<tr>
<td>1.110</td>
<td>1.313506</td>
<td>1.00</td>
<td>ΕΝΤΑΣΗ</td>
<td>11</td>
</tr>
<tr>
<td>1.113</td>
<td>0.239927</td>
<td>4.00</td>
<td>ΑΥΔΙΟ</td>
<td>4</td>
</tr>
<tr>
<td>1.127</td>
<td>1.259023</td>
<td>1.00</td>
<td>ΔΙΑΡΚΕΙΑ</td>
<td>17</td>
</tr>
<tr>
<td>0.159</td>
<td>1.442590</td>
<td>1.00</td>
<td>ΕΙΤΡΑΦΗ</td>
<td>19</td>
</tr>
</tbody>
</table>

Value ranges for expressions (8) and (9) have been selected to be Expr(8) < 2, Expr(9) < 2. This is because two words that are meant to be translations of one another, can be expected to be translated that way at least half of their occurrences in the texts.

To determine the allowable value of Expr(10), the following model-training process has been followed: For a number of randomly selected words, the tables associated with them have been inspected and the cases of correct translations have been located. On the basis of this information, the accepted range for expression (10) has been calculated: Expr(10) > 0.2. Training stops when new data does not substantially alter model parameters.

5 Results

After the method has been applied to the corpus described in the “Corpus” section, 282 Greek to English translations have been produced. The success rate of the list was 93.6%. The 282 translations account for 10972 occurrences in the Greek side of the text and 11640 in the English side, that is approximately 40.5% and 50.1% respectively of all the content word occurrences in the texts, excluding the functional words. Some of the alphabetically first correspondences are the following (underlined pairs are the wrong ones):

ΑΓΝΟΩ <-> IGNOR
ΑΚΡΑΙΟΣ <-> INTEGER
ΑΝΑΓΚΑΙΟΣ <-> NECESSARY
ΑΝΑΓΝΩΡΙΖΩ <-> IDENTIFY
ΑΝΑΖΗΤΗΣΗ <-> SEARCH
ΑΝΑΘΕΤΩ <-> ASSIGN
ΑΝΑΛΥΣΗ <-> RESOLUTION
ΑΝΑΜΕΣΑ <-> BETWEEN
ΑΝΑΘΕΡΩ <-> REFER
ΑΝΑΦΟΡΑ <-> REFERENCE
ΑΝΕΞΑΡΤΗΣΗ <-> REGARDLESS
ΑΝΟΙΓΩ <-> OPEN
ΑΝΤΙΠΡΑΦΗ <-> CUSTOM

...
These results are quite encouraging, given that the method depends on light statistical processing and takes advantage of shallow linguistic information.

6 Conclusion

This paper has presented a method for the automatic construction of a translation lexicon from parallel texts featuring hybrid techniques, statistical and linguistic, that enable tackling the translation unit variation problem. Such methods prove extremely useful for building translation lexica of specific domains based on the actual usage of translation units in parallel corpora. The method presented here has been applied on a small corpus of software documentation manual, but it can equally well be applied on other domain corpora of similar closure. While the work described here has specifically addressed Greek-English corpora, the underlying principles are general enough to be able to be applied in other language pairs.

7 Acknowledgments

This research has been conducted in the framework of the LE-I 2238 AVENTINUS project funded by the Commission of the European Union. The authors would like to thank NCSR Demokritos and HP Hellas for providing the Greek and English version of the VUE corpus. Special thanks are due to Byron Georgantopoulos for his help in the tagging task and to Karine Wagner for her assistance in the corpus normalisation.

8 References


