# Monolingual Corpus-based MT using Chunks

**Stella Markantonatou[1], Sokratis Sofianopoulos[2], Vassiliki Spilioti[3], Yiorgos Tambouratzis[4], Marina Vassiliou[5], Olga Yannoutsou[6], Nikos Ioannou[7]**

Machine Translation Department, Institute for Language & Speech Processing

6 Artemidos & Epidavrou Str., Paradissos Amaroussiou, Athens, GREECE 151 25

[1]marks, [2]s_sofian, [3]vspiliot, [4]giorg_t, [5]mvas, [6]olga@{ilsp.gr}, [7]nick.ioannou@gmail.com

## Abstract

In the present article, a hybrid approach is proposed for implementing a machine translation system using a large monolingual corpus coupled with a bilingual lexicon and basic NLP tools. In the first phase of the METIS system, a source language (SL) sentence, after being tagged, lemmatised and translated by a flat lemma-to-lemma lexicon, was matched against a tagged and lemmatised target language (TL) corpus using a pattern matching algorithm. In the second phase, translations are generated by combining sub-sentential structures. In this paper, the main features of the second phase are discussed while the system architecture and the corresponding translation approach are presented. The proposed methodology is illustrated with examples of the translation process.

**Keywords:** MT, monolingual corpus, chunks, METIS-II

## 1    Introduction

In this article we present on-going work on a hybrid approach for implementing a machine translation system which uses a large monolingual corpus coupled with a bilingual lexicon, a tagger, a lemmatiser and a chunker. Translating without bilingual parallel corpora has been the focus of the METIS[1] projects. In the first phase of the METIS system (Dologlou et al., 2003 and Ioannou, 2003), a source language (SL) clause was tagged, lemmatised and translated by a flat lemma-to-lemma lexicon. The string resulting from these procedures was matched against a tagged and lemmatised target language

corpus using a pattern matching algorithm. Results of adequate quality were received, only when a similar clause did exist in the TL corpus. However, even for very large corpora this proved to be unlikely. The next step was to attempt to generate a translation by combining translations of the chunks of the SL clause.

In the present paper, we first present the main features of our approach and then the architecture of the system. Finally, we use concrete examples to illustrate the translation process.

## 2    The main features of METIS

Resources have been one of the major problems in MT regardless of the approach, whether RBMT, EBMT, SMT or other: lexica, grammars/parsers, parallel corpora are some of the required resources. EBMT (Nagao, 1984) and statistics-based approaches (Brown et al., 1990) originally aimed at avoiding the problem of great expenditure resources in human expertise. The argument, however, was proven to be weak in two respects. First, from the days of early SMT (Brown et al., 1990), it was admitted that some amount of linguistic knowledge was necessary. This wisdom does not seem to have been altered much by today, at least as regards the need for bilingual lexica (Brown et al., 1990 and Popovic et al., 2005). Second, all corpus-based approaches rely on large bitexts (McTait, 2003) in order to produce reasonable results, and such bitexts are rare, may be of questionable linguistic quality (Al-Onaizan, 2000), and are usually confined to a sublanguage, while their register identity is a parameter rather difficult to control. The approach selected for METIS is innovative, exactly because it relies on a monolingual corpus, still a relatively low-cost and easy-to-construct resource, whose quality and register type are more controllable issues than in the case of bitexts.

Working at sub-sentential level has been proposed as a promising way of achieving better exploitation of the linguistic knowledge in a corpus (Cranias, 1997). A variety of ways of fragmenting

---

sentences for MT purposes have been proposed ranging from the exploitation of highly structured representations of linguistic knowledge (Way, 2003) to the establishment of string correspondences with little/trivial linguistic knowledge representation adhered to them (Brown et al., 1990 and McTait, 2003). However, any method relying on the combination of sub-sentential strings faces the problem of boundary friction, while 'more linguistic' methods are reported to be less affected by it than 'less linguistic' ones (Way, 2003).

The hybrid approach described here presupposes work at sub-sentential level and freely draws on the EBMT, RBMT and SMT paradigms. It aims to be modular, language-independent and with a small number of language-pair specific tools and resources being added to the core engine. In order to illustrate its principles, the Greek (SL) to English (TL) language pair was selected by ILSP within the METIS projects.

## 3    A Methodology for Implementing the Machine Translation Task

In order to translate with a monolingual corpus, we have defined a sequence of steps shown in Figure 1 where different colours signal the two main parts of the system architecture. The first part (white-coloured entities) consists of processes that are performed initially so as to obtain a translation. The second part (grey-coloured entities) consists of processes performed only when the first part results are of a non-satisfactory quality. The source sentence and the target corpus are annotated before the sentence matching algorithm applies. The overall translation process comprises the following steps:

1. Annotation of the TL corpus (off-line)
2. Annotation of the SL sentence (on-the-fly)
3. Exploitation of the TL corpus to create the best translation (on-the-fly)
4. Synthesising the translation output (on-the-fly)

### 3.1    Annotation of the TL Corpus

In order to be searched efficiently for candidate translations of SL sentences, the TL corpus is annotated. For the purposes of METIS-II, the British National Corpus (BNC)[2] has been selected as the TL corpus, because it has been established as the largest, general-purpose balanced corpus for this language. Annotation is performed off-line and only

once: BNC is tagged with the CLAWS5[3] tagset (it actually comes with a large part of it golden-tagged as standard) and is lemmatised with a purpose-built lemmatiser[4]. It is then exhaustively annotated with a purpose-built tool for clauses containing a finite verb (non-finite clauses such as gerunds or infinitival clauses are not considered: [*Walking the dog I met Iris*] [*who wanted to pick flowers*]). Clauses are then annotated for VGs, NPs, PPs (at the moment) with the ShaRPa 2.0 chunker (Vandeghinste, 2005).

To ensure a fast and efficient search for a best match, clauses are indexed according to their finite verb and chunks are classified into sets according to their label (sets of NPs, PPs etc.) and their head.

### 3.2    Annotation of the SL Sentence

The SL sentence is annotated with the linguistic information necessary to guide the matching algorithm before being fed to the matching algorithm. First, it is tagged and lemmatised with a PAROLE compatible ILSP tool (Labropoulou et al., 1996). It is then annotated for finite clauses and their constituent chunks with the ILSP chunker (Boutsis et al., 2000). The output of the chunker consists of a sequence of labelled chunks and the words contained in each chunk. A purpose-made script marks the respective heads. Next, two flat bilingual lexica are sequentially applied on the tagged-lemmatised string; first the Expression Lexicon, which contains the translations for multi-word units and second, the Word Lexicon with single-word units. The output of the lookup is a list of sets of TL lemmata (each list containing all possible translations for a given term in the source language) with PoS information for the Word lexicon, while word forms are maintained in the Expression one, (ILSP: Internal Document, Specifications for METIS lexicon, 2004).

Up to this point only basic resources have been used for both the SL and the TL. Apart from the bilingual lexica, they are all monolingual general purpose NLP tools not dedicated solely to MT. In our case, bilingual lexica have been constructed by drawing on existing resources, which after being checked for consistency and accuracy, were homogenised to fit to the system's requirements.

### 3.3    Employing Mapping Rules

The system, as presented in Figure 1, allows for the possibility of employing a limited set of mapping rules aimed to map the string obtained by the

---

[2] http://www.natcorp.ox.ac.uk/index.html

[3] http://www.comp.lancs.ac.uk/ucrel/claws5tags.html
[4] http://iai.iai.uni-sb.de/~carl/metis/lemmatiser

lemma-to lemma-translation onto a string which is closer to what we expect to find in the target language. Analogies respected, this process has been shown to greatly enhance the translation quality in rule-based systems (Dyvik, 1995). Mapping rules will not be used to deal with local problems but rather to accommodate significant linguistic differences across a given language pair. Subsets of these rules may be (re)used for any pair of languages presenting the same typological differences. As an indicative example we use NP order, which the pattern matching algorithm treats in a way that makes sure that Modern Greek NP nominatives correspond to preverbal English NPs (typical features of subject NP in Modern Greek and English respectively). This case obviously reflects the typological difference between languages which use case and languages which employ strict word order to mark functional relations.

## 3.4 The Sentence Matching and the Synthesising Algorithm

All steps up to this point belong to the annotation stage. The material collected during the SL sentence annotation phase is input to the Sentence Matching Algorithm, which compares this information with the corresponding information retrieved from BNC.

As a first step, the algorithm, which examines both the sentence structure (in terms of number and types of chunks) and sentence contents (in terms of lemmata and tags within each chunk), searches the BNC for a very similar sentence. If one exists, it is retrieved and sent to the synthesising algorithm. If, however, no candidate sentence has a very high similarity to the input, the phrase matching algorithm searches within the BNC to retrieve chunks originating in different sentences in order to replace the mismatching chunks of the best-matching sentence.

In the unlikely case that no overall structure is found, the system attempts to modify the structure and provide translations for as many phrasal parts of the SL sentence as possible by searching again within the BNC for appropriate chunks, extracted from different sentences.

The synthesising algorithm combines the essential parts of the best-matching sentence (the '*framework*', see Section 4) with the material from other BNC sentences to generate a sentence of satisfactory quality.

In the most general case the pattern matching based search algorithm yields a set of fragments (chunks and sets of chunks), which are fed to the synthesising algorithm. The latter roughly comprises two tasks: (a) the modification and rearrangement of the retrieved chunks, so that they can be meaningfully combined into a sentence and (b) the handling of morphological phenomena. Task (a) draws mainly on a number of synthesis rules, while for task (b) a morphological generator is employed [see footnote 4].

Below, we present in more detail the rationale and the practical steps taken at the matching phase.

## 4 The Matching Procedure: rationale

The mechanism employed for making the SL and the TL languages "meet" relies on the already mentioned notion of a clause '*framework*' (Section 3.4), which represents the main clause structure with the verb head-lexicalised. We thus seek to retrieve from the monolingual corpus clauses that contain the TL verb[5], which is the exact translation of the SL verb (the lexicon may provide more than one such solutions), in a context consisting of the same amount of referential expressions.

The idea behind this requirement is that sentences express events with a certain number of participants. The event is basically denoted by the verb while the participants mainly by referential expressions, embedded within some grammatical information functor, call it Case (from a purely morphological point of view) or Preposition or both. For instance, the Modern Greek sentence

| O | Petros | mpike | sto | | dhomatio |
|---|--------|-------|-----|---|----------|
| The- | Peter- | enter- | in-the- | - | room-Acc |
| Nom | Nom | 3rd-SG-Past | Pr | | |

'Peter entered the room'

denotes an event with two participants, one embedded under the Nominative Case and the other one under a preposition and the Accusative Case. Its English correspondent differs from it as regards the grammatical functor of the second referential expression.

For our approach, it is important that, although we avail ourselves to no information about the subcategorisation preferences of the verbs involved, we end up with the proper verb and the proper referential expressions embedded under the proper functor.

---

[5] One could look for families of verbs occurring in the same syntactic environment. We would first like to exhaust the present approach and then move to a more abstract description of phrase structure.

To this end, our pattern-matching algorithm generalises over these two types of grammatical functor, Case and Prepositions. Thus, while the matching algorithm takes care of the essential cross-language information (the verb predicate and the amount of referential expressions), the grammatical particularities of either language are supplied by their well-formed strings (this viewed as mapping from the SL->TL implies that the corpus plays the role of the supplier of grammatical information about the TL). In the example above, our algorithm will select a TL sentence with the verb 'enter' in the appropriate grammatical context, which is not a one-to-one copy of the SL grammatical context.

Of course, the assumption underlying this approach is that verbal expressions are translated to verbal expressions and referential expressions to referential ones. This might be a strong hypothesis; however, it is considerably less strong than requiring grammatical equivalence across language pairs.

On a similar par, that of generalising linguistic patterns at the matching phase, we have chosen to work with lemma-to-lemma bilingual lexica rather than looking for tokens in the TL corpus. Morphological information is, in general, relatively simple to incorporate at the end of the overall translation procedure.

Having said the above, it must be noted that all SL information is kept as default information, overwritten only by corpus information. For instance, when no framework is found containing all the appropriate chunks, an appropriate one is introduced by directly mapping information from the SL onto the TL.

We now proceed to present the matching procedure step by step.

### 4.1 Matching Step by Step

**Step 1:** As explained before, clauses from the BNC are retrieved, based on the main verb and the number of chunks. For each different translation of the SL verb a different set of clauses is created.

The multiple translations provided by the lexicon are reduced by calculating the relative frequencies of co-occurrence of chunk heads (i.e. verbs with nouns, verbs with prepositions, prepositions with their noun complements) within the BNC. Consequently the number of combinations the system has to check against the BNC material is reduced. The alternative candidate combinations are checked and ranked in the following way:

Initially, the relative frequency $R_{((i,j),(a,b))}$, where $(i,j)$ denotes the $j$-th translation of the $i$-th chunk-

head in relation to a $a$-th translation of a $b$-th chunk-head $((a,b))$, is calculated:

$$R_{((i,j),(a,b))} = \frac{c_{\{j,b\}}}{\sum_{b=1}^{v} c_{\{j,b\}}} \qquad \text{(eq. 1)}$$

where $C\{j,b\}$ is the number of co-occurrences of the $(i,j)$ lemma with the $(a,b)$ lemma and $v$ is the number of translations provided by the lexicon for the $a$-th chunk-head.

Then, every possible combination is determined by (eq.2):

$$\prod_{\substack{i=1 \\ 1 \le j \le trj \\ 1 \le b \le trb}}^{\mu\text{-}1} R_{((i,j),(i+1,b))} \qquad \text{(eq. 2)}$$

where $\mu$ is the number of the chunks in the sentence, and $trj$, $trb$ are the numbers of translations for the $i$-th and $(i+1)$-th chunk-head, respectively. The combination with the higher score is chosen.

**Step 2:** For each translation of the SL clause, which has scored high, a comparison is run between the SL clauses and the BNC clauses. The search originates within the class of clauses containing the given verb. If no matches ('good frameworks') are found, searching has failed (at this level of development of the system). The result of each comparison is a score for the SL clause and TL clause pair, based on general chunk information, such as the number of chunks in the clause, chunk labels and chunk heads, using a pattern recognition-based method. The formula for calculating the score is

$$ClauseScore = \sum_{n-1}^{m} \left\{ ocf_n \times \frac{ChunkScore_n}{\sum_{n=1}^{m} ocf_n} \right\} \text{(eq.3)}$$

where $m$ is the number of chunks in the SL clause and $ocf$ is the overall cost factor of each chunk (based on the chunk type).

Chunk scores are calculated by combining the partial scores obtained after comparing the chunk label as well as the tag and the lemma of the chunk head. Given that not all chunk types are of the same significance, we need to introduce a series of weights. The formula for calculating the score for each chunk is the following:

$$ChunkScore_n = (1 - tcf_n - lcf_n) \times LabelComp_n + \\ + tcf_n \times TagComp_n + lcf_n \times LemmaComp_n$$

where *tcf* is the tag cost factor, *lcf* the lemma cost factor and *(1-tcf-lcf)* the chunk label cost factor.

**Step 3:** In the third step of the algorithm, the comparison is more detailed and involves comparing the tokens contained in each chunk. The SL chunks are checked against the respective chunks in the BNC clause, again using a pattern recognition-based method. At the end of this step a second score is given to each clause pair (and to each chunk of the clause) in a similar way to the second step.

The final score for each pair is the product of the clause scores obtained at steps 2 and 3. Final scores are calculated for each chunk as well. The BNC clause of the comparison pair with the highest clause score will serve as the best-matching and form the archetype of the translation. The chunk comparison pairs of the clause are then classified on the basis of their final score. Chunks scoring higher than A% will be used in the final translation without any changes. Chunks scoring between A% and B% (A > B) will be used in the final translation after modifications are made. Finally, chunks with a score lower than B% are not considered eligible candidate translations. To translate these SL chunks, we need to search the BNC again for chunks based on chunk label and head token information. Values A and B are entered as parameters to the system, so that the translator can tune the precision of the final translation.

## 4.2 Example of the Translation Process

The process proposed for translating a sentence with the approach presented so far is summarised in Table 1 where rows are numbered.

In (1), the SL string is a Modern Greek declarative sentence with a VSO word order.

In (2), (3) & (4), the results of tagging, lemmatising and chunking the SL sentence are shown.

In (5), the result of the dictionary look-up is shown. All possible translations are managed through the relative frequency of co-occurrence algorithm.

In (6) the chunks from the SL string are copied on the lemma-to lemma string.

In (7), the core engine searches and finds a similar string in terms of chunks and lexical heads. Furthermore, by applying the NP order mapping rule (Section 3.3), the algorithm has established an implicit link between the NPs in the SL and the TL so that TL 'cuban officers' is linked to the SL 'american officer' and TL 'continuous animosity' is linked to SL '{constant, continuous, unabated}, {tension, intensity}'.

In (8) the found BNC chunks are shown. As, in this example, the sentences are isomorphic, they coincide in terms of the number and type of chunks.

In (9) the retrieved string after synthesising appears.

## 4.3 Experimental Results

In Table 2, the translation results obtained from the prototype for a sample experiment are briefly presented. For this experiment, a simple sentence was used (Row 1). The results of the analysis of the sentence are shown in Rows 2 to 5, while the reference translation is shown in lemmatised form in Row 6. The experiment was carried out using a prototype of the system running under Java. The monolingual corpus consisted of 1,703,551 sentences, and the translation process was completed in 31.44 seconds on a Dell 670 Precision workstation. The top 12 sentences retrieved from the corpus as candidate translations are shown in the bottom part of Table 2, ranked according to their overall score, together with their associated scores. As can be seen, the score for step 2 is generally higher than that for step 3. In certain cases, the score of step 3 is higher for a lower-ranked sentence, though the overall score agrees to a large extent with that of step2. The system is successful in retrieving the sentences with the highest similarity to the SL sentence (sentences 1 to 6). Lower-ranked sentences seem to indicate a decreasing similarity to the reference translation. The exact ranking depends on the exact values of the weights, which are currently being fine-tuned.

## 5 Future Work

In the present article we have described a methodology for a machine translation system employing a limited set of resources. The approach exploits sub-sentential structure information and is based on searching and retrieving the most appropriate translation from a large monolingual corpus. It is self-evident that the accuracy and quality of the retrieved translations is heavily dependent upon the size and coverage of the given corpus.

Currently, we are experimenting on the optimisation of the proposed algorithm along the following lines:

∗ Extending the corpus indexing scheme, in order to accelerate the search process and improve its effectiveness

∗ Narrowing down the search space

∗ Exploring further the issue of synthesising the final translation from multiple segments (chunks/clauses)

∗ Studying the issue of automatic evaluation (METEOR, NIST, Papineni et al., 2002) of the output of the algorithm.

## References

Y. Al-Onaizan, U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, & K. Yamada. 2000. *Translating with Scarce Resources.* American Association for Artificial Intelligence Conference (AAAI'00), 30 July – 3 August, Austin, Texas, pages 672-678 (http://www.isi.edu/natural-language/projects/rewrite).

S. Boutsis, P. Prokopidis, V. Giouli & S. Piperidis. 2000. *A Robust Parser for Unrestricted Greek Text.* In "Proceedings of the Second International Conference on Language Resources and Evaluation", 31 May-2 June, Athens, Greece, Vol. 1, pp. 467-482.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer & P. S. Roosin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79-85.

L. Cranias, H. Papageorgiou & S. Piperidis. 1997. Example Retrieval from a Translation Memory. *Natural Language Engineering*, 3:255-277.

I. Dologlou, S. Markantonatou, G. Tambouratzis, O. Yannoutsou, A. Fourla & N. Ioannou. 2003. *Using Monolingual Corpora for Statistical Machine Translation.* In "Proceedings of EAMT/CLAW 2003", Dublin, Ireland, 15-17 May.

H. Dyvik. 1995. Exploiting Structural Similarities in Machine Translation. *Computers and the Humanities*, 28:225-234.

ILSP Internal Document 2004. *Specifications for METIS lexicon.*

N. Ioannou. 2003. METIS: *Statistical Machine Translation Using Monolingual Corpora.* In "Proceedings of the Workshop on Text Processing for Modern Greek: From Symbolic to Statistical Approaches" (held in conjunction with the 6th

International Conference of Greek Linguistics), Rethymno, Greece, 20 September. ISBN:960-88268-0-2.

P. Labropoulou, E. Mantzari & M. Gavrilidou. 1996. *Lexicon-Morphosyntactic Specifications: Language Specific Instantiation* (Greek), PP-PAROLE, MLAP 63-386 report.

METEOR: http://www-2.cs.cmu.edu/~banerjee/MT/METEOR/

K. McTait. 2003. *Translation Patterns, Linguistic Knowledge and Complexity in EBMT.* In "Recent Advances in Example-Based Machine Translation", M. Carl and A. Way (eds.) Kluwer Academic Publishers, pp. 307-338.

M. Nagao. 1984. *A Framework of a Mechanical Translation between Japanese and English by Analogy Principle.* In "Artificial and Human Intelligence", A. Elithorn and R. Banerji (eds). North-Holland.

NIST: http://www.nist.gov/speech/tests/mt/

K. A. Papineni, S. Roukos, T. Ward & W. J. Zhu. 2002. *Bleu, a method for automatic evaluation of Machine Translation.* In "Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics", Philadelphia (USA), pages 311-318.

M. Popovic and H. Ney. 2005. *Exploiting Phrasal Lexica and Additional Morpho-syntactic Language Resources for Statistical Machine Translation with Scarce Training Data.* EAMT 10th Annual Conference, 30-31 May, Budapest, Hungary.

V. Vandeghinste 2005. *Manual for ShaRPa 2.0.* Internal Report, Centre for Computational Linguistics, K.U.Leuven.

A. Way 2003. *Translating with Examples: The LFG-DOT Models of Translation,* In "Recent Advances in Example-Based Machine Translation", M. Carl and A. Way (eds.). Kluwer Academic Publishers, pages 443-472.

```
┌─────────────────────┐                              ╭─────────────────────────╮
│   Source Sentence   │─────────────────────────────▶│   TAGGER/LEMMATISER     │
└─────────────────────┘                              ╰─────────────────────────╯
                                              ┌──────────────────────────────────┐
                                              │     Tagged/Lemmatised Sentence    │
                                              └──────────────────────────────────┘
                                                        ⬡ Expression Lexicon
                                              ┌──────────────────────────────────┐
                                              │  Sentence with detected expressions│
                                              └──────────────────────────────────┘
                                                        ⬡ Word Lexicon
       ╭─────────────────────────╮            ┌──────────────────────────────────┐
       │    CHUNKER/PARSER        │            │ Lexicon-translated string of words│
       ╰─────────────────────────╯            └──────────────────────────────────┘
                                                        ⬡ Mapping Rules I
        ┌──────────┐   ┌──────────────┐        ┌──────────────────────────────────┐
        │ Sentence │   │  in-sentence │        │Linguistically-processed string of │
        │ Structure│   │   Phrasal    │        │            words                  │
        └──────────┘   │   Content    │        └──────────────────────────────────┘
                       └──────────────┘
  ⬡ Mapping        ╭──────────────────────────╮         ⬡ Processed
    Rules II       │ SENTENCE MATCHING ALGOR. │           BNC sentences
                   ╰──────────────────────────╯
              ┌──────────┐  ┌──────────┐                  ⬡ Processed
              │Translation│  │Best trans│                   BNC phrases
              │ Structure │  │ Sentence │
              └──────────┘  └──────────┘
  ⬡ Synthesis       YES  ◇ Adequate   NO    ┌───────────────┐
    Rules               Translation          │  Phrase 1..n  │
                         Quality ?           └───────────────┘
                                             ╭───────────────────────────╮
                                             │  PHRASE MATCHING ALGOR.   │
                                             ╰───────────────────────────╯
  ╭──────────────────────────╮               ┌───────────────────────────┐
  │ SYNTHESISING  ALGORITHM  │               │ Translation of Phrase 1..n│
  ╰──────────────────────────╯               └───────────────────────────┘
                              ┌──────────────────────────────────────────┐
                              │  Translations 1-k of source sentence     │
                              └──────────────────────────────────────────┘
```
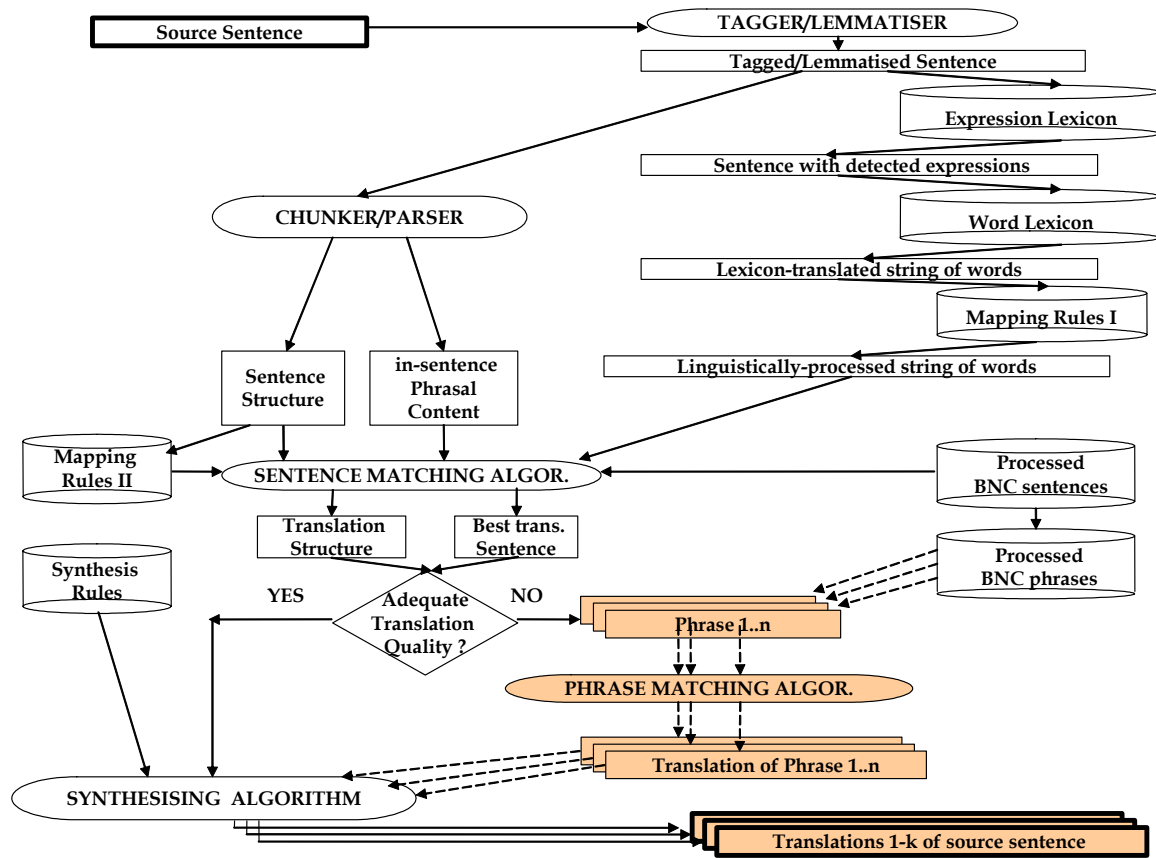
Figure 1: General architecture of the proposed methodology

| 1 | περιγράφουν | Αμερικανοί | αξιωματούχοι | τη | διαρκή | ένταση | μεταξύ | Ελλάδας | και | Τουρκίας |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | **Vb** | **Aj** | **No** | **At** | **Aj** | **No** | **AsPp** | **No** | **Cj** | **No** |
| 3 | περιγράφω | Αμερικανός | αξιωματούχος | ο | διαρκής | ένταση | μεταξύ | Ελλάδα | και | Τουρκία |
| 4 | **VG** | **NP** | | **NP** | | **PP** | | | | |
| 5 | describe | American | officer official | the | constant continuous unabated | tension intensity | between mean-while | Greece | and | Turkey |
| 6 | **VG** | **NP** | | **NP** | | **PP** | | | | |

| | | | | **Searching for match in pre-processed BNC** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 7 | cuban | officers | describe | the | continuous | animosity | between | Greece | and | Turkey |
| 8 | **NP** | | **VG** | **NP** | | **PP** | | | | |
| 9 | american | officers | describe | the | continuous | tension intensity | between | Greece | and | Turkey |

Table 1: An example of the translation approach

| level | Sentence | | | | | | | Score (step 2) | Score (step3) | Overall Score |
|---|---|---|---|---|---|---|---|---|---|---|
| **SL string (1)** | H | γυναίκα | έχασε | έναν | αδελφό | στον | πόλεμο | | | |
| **Tags** | **At** | **No** | **Vb** | **Card** | **No** | | **AsPp** | | | |
| **Lemmata (3)** | O | γυναίκα | χάνω | ένας | αδελφός | στου | πόλεμο | | | |
| **SL string chunked (4)** | NP | | VG | NP | | PP | | | | |
| **Lemma-to-lemma (5)** | The | woman wife lady | lose miss misplace | a one | brother | in during | war battle | | | |
| **Reference translation (6)** | The | woman | lose | a | brother | in | war | | | |
| **Retrieved sentences from pre-processed corpus** | | | | | | | | | | |
| **Retrieved sentence 1** | The poor woman | | lost | her older brother | | in war | | 100 | 95.9 | 95.9 |
| **Retrieved sentence 2** | The woman | | lost | her brother | | during the great war | | 94.9 | 88.2 | 83.8 |
| **Retrieved sentence 3** | The woman | | lost | a dog | | in war | | 93.3 | 88.6 | 79.9 |
| **Retrieved sentence 4** | Both women | | lost | their husbands | | in the war | | 93.3 | 82.4 | 76.9 |
| **Retrieved sentence 5** | The man | | lost | a brother | | in war | | 88.2 | 78.5 | 69.3 |
| **Retrieved sentence 6** | The brother | | lost | his wife | | in war | | 86.6 | 74.9 | 64.9 |
| **Retrieved sentence 7** | The woman | | lost | an apple | | in the kitchen | | 86.6 | 73.2 | 63.4 |
| **Retrieved sentence 8** | Britain | | lost | a lot | | in that war too | | 86.6 | 71.8 | 62.2 |
| **Retrieved sentence 9** | The brother | | lost | her | | in the war | | 83.8 | 72.2 | 60.5 |
| **Retrieved sentence 10** | He | | lost | two sons | | in the Great war | | 83.8 | 66.1 | 55.3 |
| **Retrieved sentence 11** | They both | | lost | their husbands | | in the war | | 81.0 | 68.0 | 55.1 |
| **Retrieved sentence 12** | Pitch Barratt Developments | | lost | 9p | | to 173p | | 80.0 | 61.2 | 48.9 |

Table 2: Translation results generated by the prototype for a sample sentence