Building a Greek corpus for Textual Entailment

Evi Marzelou, Maria Zourari, Voula Giouli, Stelios Piperidis
Institute for Language and Speech Processing,
ILSP /ATHENA
Athens, Greece
{emarzel; mariaz; voula; spip@ilsp.gr}

Abstract

The paper reports on completed work aimed at the creation of a resource, namely, the Greek Textual Entailment Corpus (GTEC) that is appropriate for guiding training and evaluation of a system that recognizes Textual Entailment in Greek texts. The corpus of textual units was collected in view of a range of NLP applications, where semantic interpretation is of paramount importance, and it was manually annotated at the level of Textual Entailment. Moreover, a number of linguistic annotations were also integrated that were deemed useful for prospect system developers. The critical issue was the development of a final resource that is re-usable and adaptable to different NLP systems, in order to either enhance their accuracy or to evaluate their output. We are hereby focusing on the methodological issues underpinning data selection and annotation. An initial approach towards the development of a system catering for the automatic Recognition of Textual Entailment in Greek is also presented and preliminary results are reported.

1. Introduction

Over the recent years, there has been a growing research interest towards textual entailment recognition, the latter seen as an umbrella topic encapsulating semantic variability problems faced by a wide range of NLP applications with a strong text understanding dimension. To this end, building a system which, given two text fragments, will be able to recognize whether the meaning of one text can be inferred (entailed) from the other, has become a challenging task. The methods and different approaches that have been suggested, all concerning the English language, have made extensive use of corpora exclusively collected and/or fabricated for this purpose. Building on established methodologies, we have built a Greek corpus that is being used for training and evaluation of a system aimed at Recognizing Textual Entailment (RTE) in Greek texts. This paper reports on completed work involving the collection and annotation of the Greek Textual Entailment Corpus (GTEC), and it presents preliminary investigations in the direction of developing an RTE system that deals with entailment in Greek texts.

This paper is organised as follows: section 2 gives an overview of the data comprising the GTEC, whereas section 3 discusses the methodologies adopted for the selection of appropriate textual units. Corpus annotation is being described in section 4, the focus being on techniques for validation and consistency assurance. Initial considerations towards the development of a system that recognises textual entailment in Greek along with preliminary results are presented in section 5; final conclusions and future considerations are outlined in section 6.

2. Corpus Description

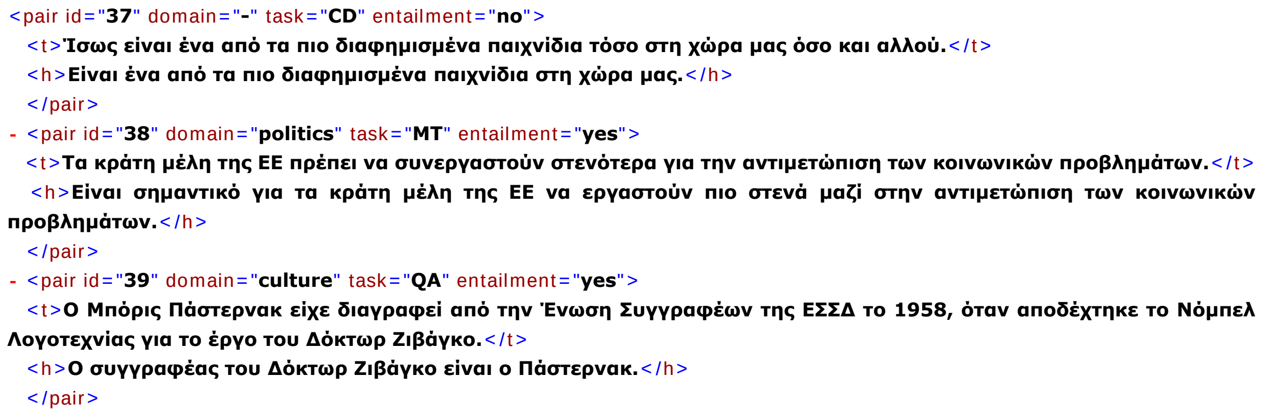
RTE corpora referred to in the literature [Dagan et al. 2005] [Dagan et al. 2006] typically consist of pairs of textual units, namely the entailing "Text" (T) and the entailed "Hypothesis" (H) that are relevant to a range of Natural Language Processing (NLP) applications. In our work, *T* is a textual unit that is typically made up of one or more sentences while *H* usually contains one simpler sentence. This is based on the logical assumption that *T* must enclose more information than *H* in order to entail it. We say that *T* entails *H* if and only if the meaning of *H* can be inferred from the meaning of *T*.

Our corpus dataset consists of 600 *T-H* pairs equally divided in three subsets each one representing the output of a certain application. Each portion of the dataset intends to include *T-H* examples that correspond to success and failure cases of the actual applications. The applications chosen for the needs of the current work are: Question Answering (QA), Comparable Documents (CD) and Machine Translation (MT). Moreover, since the applications at hand are, in most cases, domain-oriented, we initially opted for gathering textual data pertaining to specific subject fields (i.e., law, politics, travel, etc). However, our intention was to build an up-to-date resource, which would cover as many thematic domains as possible, and therefore, the corpus was further enriched with texts representing general language.

Raw data have been manually annotated for Textual Entailment. Each *T-H* pair appears within a single <*pair*> element, with the following attributes:

- ♦ *id*: a numeral identifier, unique for each T-H pair.
- ♦ *task*: the acronym of the application setting from which the pair has been generated, with one of the following values "QA", "CD", "MT".
- ♦ *domain*: the specific subject field that the initial documents pertain to. Possible values in the present implementation are "law", "politics", travel", "finance", "sports", "culture", and "general".
- ♦ *entailment*: the gold standard entailment annotation, being "YES", "NO" or "UNKNOWN".

T-H pairs have further been coupled with annotations at various levels of linguistic analysis that will be presented in section 4. The GTEC is represented in XML format (Figure 1), with the different levels of annotation stored in separate XML files.



```
<pair id="37" domain="-" task="CD" entailment="no">
  <t>Ίσως είναι ένα από τα πιο διαφημισμένα παιχνίδια τόσο στη χώρα μας όσο και αλλού.</t>
  <h>Είναι ένα από τα πιο διαφημισμένα παιχνίδια στη χώρα μας.</h>
  </pair>
- <pair id="38" domain="politics" task="MT" entailment="yes">
  <t>Τα κράτη μέλη της ΕΕ πρέπει να συνεργαστούν στενότερα για την αντιμετώπιση των κοινωνικών προβλημάτων.</t>
   <h>Είναι σημαντικό για τα κράτη μέλη της ΕΕ να εργαστούν πιο στενά μαζί στην αντιμετώπιση των κοινωνικών
προβλημάτων.</h>
  </pair>
- <pair id="39" domain="culture" task="QA" entailment="yes">
  <t>Ο Μπόρις Πάστερνακ είχε διαγραφεί από την Ένωση Συγγραφέων της ΕΣΣΔ το 1958, όταν αποδέχτηκε το Νόμπελ
Λογοτεχνίας για το έργο του Δόκτωρ Ζιβάγκο.</t>
  <h>Ο συγγραφέας του Δόκτωρ Ζιβάγκο είναι ο Πάστερνακ.</h>
  </pair>
```

Figure 1. XML Representation

3. Corpus creation

Selection of appropriate applications was the starting point for the development of the GTEC. Corpus development with a view to RTE was systematically tackled for the first time in the PASCAL Challenge Workshops. In this framework, textual data were collected corresponding to NLP applications for which RTE was considered a significant element. To this end, RTE 1 datasets pertained to seven applications: Information Retrieval (IR), Comparable Documents (CD), Reading Comprehension (RC), Question Answering (QA), Information Extraction (IE), Machine Translation (MT), and Paraphrase Acquisition (PP). The results of the first challenge led to the reduction of the applications under study in subsequent challenges, by eliminating the applications that rendered the best results. Therefore, consecutive Challenges (RTE 2 and RTE 3) catered only for IR, QA, IE, and Summarization (SUM). The latter was added as an application relative to CD, yet quite distinct from it, since a machine's output was compared to a text that was produced manually.

With the above considerations taken into account and on the basis of the similarity that some of the above applications appear to have, the GTEC was collected with respect to three different text processing applications: QA, CD, and MT. More precisely, we clustered the above eight applications into two separate groups based on the theoretical assumption that (a) the first group (comprising IR, IE, and QA) has roughly to do with extraction of information either in the form of a sentence or as a template filling procedure, while (b) the second group (that comprises CD, RC, SUM, PP, and MT) has comparable corpora creation as a common denominator. Regarding the first group, we focused on QA on the basis of observations of the RTE datasets, according to which QA pairs are similar to those pertaining to the IR and IE tasks, while, at the same time, they are the most representative ones. Furthermore, CD and MT were chosen as the most representative applications of the second group with MT further engaging the rather strict notion of parallel documents. The dataset was then, manually, collected separately for each application.

3.1 CD dataset selection

Candidate *T-H* pairs pertaining to the CD task were selected from a variety of sources: (a) proper comparable documents, (b) texts and their summaries, and (c) news headlines. Different methodologies were employed accordingly. A cluster of comparable news articles pertaining to politics, economics, sports, culture etc. referring to the same topic or subject matter were initially collected from online sources (electronic newspaper or magazine editions, portals, etc.). These were further coupled with existing comparable documents that were collected at the Institute for Language and Speech Processing in the framework of national and EU-funded projects. Once the cluster of comparable documents was intact, selection of the candidate *T* and *H* pairs was performed with manual alignment of sentences that present high lexical similarity (overlap).

Similarly, *T-H* pairs pertinent to the CD dataset were extracted from texts and their corresponding summaries on the basis of the assumption that a text and its summary consist a pair of comparable documents. To this end, a number of online texts and their summaries were collected from online resources, and the *T-H* pairs were extracted by applying sentence alignment methods on the basis of lexical overlap (Table 1). *T* might extend beyond the sentence limits and was selected from the original text, whereas *H* was a single sentence obtained from the summary.

T	Ο Ερρίκος ο Θαλασσοπόρος είναι περισσότερο γνωστός για την οργάνωση των εξερευνήσεων και την οικονομική ενίσχυση των θαλασσοπόρων. (*Henry the Navigator is most famous for the organisation and the foundation of discovery voyages.*)
H	Ο Ερρίκος ο Θαλασσοπόρος ήταν οργανωτής και χρηματοδότης εξερευνητικών ταξιδιών. (*Henry the Navigator was the organiser and the sponsor of discovery voyages.*)

Table 1. Example of T-H pair (CD)

Finally, the so-called *title-lead paragraph* technique, proposed in the literature was also exploited (see Table 2). Based on the observation that the title of a news article is most of the times a partial paraphrase of the first paragraph, conveying thus a comparable meaning, (Bayer et al., 2005) proved it to be a fruitful methodology for automatically acquiring training data that are sufficient for large-scale statistical models. To this end, we collected a number of news stories from various web sites, and their titles and corresponding lead paragraphs were extracted forming the *H* and *T* text fragments of the candidate pairs respectively.

T	Ο εκπρόσωπος του αμερικανικού υπουργείου Εξωτερικών Ρίτσαρντ Μπάουτσερ ανακοίνωσε ότι ο Κόλιν Πάουελ θα επισκεφτεί την ερχόμενη εβδομάδα το Χαρτούμ. (*The representative of the American ministry of Foreign Affairs Richard Baucher announced that Colin Powell will visit next week Khartoum*)
H	Στο Σουδάν μεταβαίνει ο Πάουελ (*Powell visits Soudan.*)

Table 2. Example of Title-lead paragraph (CD)

3.2 MT dataset selection

The *T* and *H* pairs for the MT setting were selected from alternative translations of the same source document, that is, human translations as opposed to translations produced automatically. The methodology adopted involved (a) the manual inspection of a variety of multilingual web sites that pertain to the domains of interest, and (b) the identification of candidate sources for the collection of parallel English – Greek documents. The requirement to be met was that the source documents should be in English, whereas the Greek texts should be produced by human translators. This approach differs from the one followed in the RTE challenges, since the two text fragments are translations of a unique source text and not translations of different but comparable text sources as in the RTE. In this way, we believe that we are closer to an MT evaluation scenario.

$T_{machine}$	Το νησί Aegina είναι μόνο μια ώρα μακρυά από τον Πειραιά. (*The island of Aegina is only an hour away from Piraeus.*)
H_{human}	Το νησί της Αίγινας απέχει μόλις μία ώρα από τον Πειραιά. (*The island of Aegina abstains hardly one hour from Piraeus.*)

Table 3. Example of T-H pair (MT)

The website of the European Union proved to be an invaluable source with this respect. Once the documents were collected, the English text was translated into Greek via the SYSTRAN[1] translation engine. The *T* and *H* pairs for the MT dataset were then selected from the sentence aligned human and machine translations (see Table 3). Since in the MT setting, however, both *T* and *H* are translations of the same text fragment, they should ideally express the same meaning regardless of differences in surface structure. Practically, whether entailment holds or not, depends on these differences. Moreover, pairs in the MT dataset that were judged as true seemed to exhibit a relationship broader than entailment. Since *T* and *H* fragments are, to a large extent, paraphrases of each other, semantic equivalence seemed more appropriate for this task. This observation resulted into an apparent deviation from the official RTE guidelines, as we started wondering about which of the two translations (human or machine) should consist the *T* text fragment and which the *H* one. We thus decided to examine both options by including in the MT dataset 100 pairs with *T* being the machine translation and *H* the human one ($T_{machine} \rightarrow H_{human}$), while another 100 pairs were added where *T* and *H* were the human and machine translations respectively ($T_{human} \rightarrow H_{machine}$). Annotators' judgments showed that in the MT task, textual entailment was conceived to be bidirectional, that is, *T* and *H* fragments have to entail each other, for a machine translation to be considered acceptable.

3.3 QA dataset selection

Lack of relevant systems, capable of providing true and false entailment pairs with respect to the task at hand, forced us to adopt manual selection of the QA dataset. The process consisted in three stages. Initially, a set of quizzes found over the Internet was chosen to serve as an appropriate source for extracting questions that could be fed as an input to a QA system. From the initial 500 questions, a subset of 200 questions was finally selected, that were either factual, or relative to definitions. Moreover, the answers to the selected questions had to be relevant to entities such as persons, temporal or local complements etc.

T	Το όνομα “μινωικός” προέρχεται από τον βασιλέα Μίνωα και δόθηκε από τον Άρθουρ Έβανς, τον αρχαιολόγο που ανέσκαψε το ανάκτορο της Κνωσού. (*The name “minoikos” from king «Minos», was given by Arthour Evans, the archaeologist who excavated the palace of Knossos.*)
H	Ο αρχαιολόγος Άρθουρ Έβανς ανακάλυψε την Κνωσό. (*The archaeologist Arthour Evans discovered Knossos.*)

Table 4. Example of T-H pair (QA)

At the next stage, following the official RTE specifications, the questions selected were turned into affirmative sentences with the correct answer “plugged in”, thus forming the *H* text fragments. Finally,

[1] www.systran.com

key-words of the *H* text fragments were identified (mainly names of persons, locations, and organizations) and were subsequently used for querying using Google. The answers returned to the set of queries corresponding to each *H* text fragment formed a pool from which the most appropriate text segment *T* was extracted. Two selection criteria were taken into account: (a) *T* should contain the answer either as its main proposition or as a dependent one, and (b) *T* should exhibit some sort of lexical or syntactic similarity with *H* (Table 4).

3.4 Populating the negative pairs

A corpus to be considered appropriate for the RTE task should ideally contain both positive and negative *T-H* pairs so that Textual Entailment could be soundly described and systematized on the grounds of efficient data. The afore-mentioned methodologies for obtaining the GTEC exhibit a bias towards producing positive *T-H* pairs, and we were forced to populate the corpus with more negative data. It would be easier to collect such pairs, were corresponding NLP applications existed for Greek, since the failures of these systems would provide them.

To accommodate for this shortcoming, a closer inspection of the English RTE data was performed, that confirmed our initial assumptions with respect to semantic entailment and surface structure, the focus being on negative entailment. This resulted in a more formal description of both *true* and *false* pairs. Textual entailment between two text fragments *T* and *H holds* either if:

(a) *H* conveys a meaning semantically *identical* to the meaning of *T* or to a sub-meaning of *T*, expressed in either the same or equivalent lexical items (paraphrase or partial paraphrase), or

(b) *H* conveys a meaning semantically equivalent to the main meaning of *T* or to a sub-meaning of *T,* expressed in different lexical items, which however obey to deeper semantic relations (strict entailment).

On the contrary, textual entailment between two text fragments *T* and *H does not hold* when the two text fragments are semantically different. Closer inspection of the latter showed that negative *T-H* pairs that are meaningful for the challenge at hand, have to exhibit one of the following properties:

- the information contained in *T* is a subset of the information contained in *H*: i.e. when at least one event or participant mentioned in the H text fragment, is neither referred in the T text fragment nor can be entailed by it. (see Table 5);
- *T* and *H* text fragments bare different information, *their* being:

incompatible: i.e. The event described in *H* is the same as the event described in *T*, but there exists at least one participant that is different (see Table 6).

contradictive: i.e. The events described in *H* and *T* are not only different from one another but also contradictory (see Table 7).

irrelevant: i.e. The event described in *H* is different from the event described in *T*, yet there exists at least one participant that is the same, (see Table 8).

T	Ο Περσέας σκότωσε τη Μέδουσα. (*Perseus killed Medusa.*)
H	Ο Περσέας σκότωσε τη Μέδουσα και της έκοψε το κεφάλι. (*Perseus killed Medusa by cutting her head.*)

Table 5. T is H' s subset

T	Ο Πρωθυπουργός επισκέπτεται το Λονδίνο. *(The prime minister visited London.)*
H	Ο Πρωθυπουργός επισκέπτεται το Παρίσι. (*The prime minister visited Paris.*)

Table 6. T and H are incompatible

T	Το ποσοστό ανεργίας αυξάνεται κάθε χρόνο. *(The percentage of unemployment increases every year.)*
H	Χρόνο με το χρόνο το ποσοστό ανεργίας μειώνεται. *(The percentage of unemployment decreases every year.)*

Table 7. T and H are contradictive

T	Ο κατά συρροή δολοφόνος συνελήφθη χθες το βράδυ. (*The serial-killer was arrested last night.)*
H	Ο κατά συρροή δολοφόνος σκοτώθηκε σήμερα το πρωί. (*The serial-killer was killed this morning.*)

Table 8. T and H are irrelevant

Further selection of negative pairs was performed on the basis of the afore-mentioned considerations, so that both positive and negative pairs participate equally to the datasets pertaining to all the tasks (Table 9).

	yes	no	unknown	sum
CD	105	93	2	200
MT	111	87	2	200
QA	108	91	1	200
				600

Table 9: The GTEC's pairs

4. Corpus Annotation

After corpus collection, the raw data of the GTEC were annotated for Textual Entailment by expert and non-expert human annotators. The comparative analysis of the answers supplied for each *T-H* pair, resulted in the standardization of the gold entailment annotation.

A set of initial guidelines was given to a team of expert linguists who had studied RTE thoroughly. According to the guidelines, coders had to annotate the GTEC, that is to decide whether the entailment relationship between *T* and *H* holds, assigning a value of "yes" or "no".

The reliability and soundness of annotations relevant to the RTE task in the GTEC, was further assessed by means of quantitative analysis. Average inter-annotator agreement that amounts to 0,784 was calculated among experts using the Kappa statistic.

Furthermore, a comparative analysis of these annotations was the basis for the gold annotation corpus and resulted in the identification of difficult or ambiguous cases, in which the agreement among experts was low. In order to resolve ambiguities and check annotation validity we further explored laymen's opinion on a subset of the corpus.

4.1 Validation of Annotations: Obtaining the Gold Standard Corpus

Textual entailment is by definition based on logical inference. However, the transition from the level of meaning to the language level creates ambiguities, due to the fact that subjectivity and pragmatics interfere. As in every annotation task that bares a high level of subjectivity and ambiguity, an experiment was conducted in order to resolve ambiguities and to ensure consistency in annotating the GTEC. To this end and to obtain the gold entailment annotation, RTE experts' and laymen's opinion was explored. Laymen's group comprises linguists, engineers, translators and journalists, so as to acquire a view of language comprehension as objective and integral as possible.

The experiment was performed on a subset of the corpus (450 *T-H* pairs) in two rounds. The first dataset consisted of *T-H* pairs that presented lexical and structural similarity (paraphrases), while the second one aimed at examining laymen's opinion on cases involving strict entailment. To minimize discrepancies resulting from the somewhat vague concepts of "common human understanding of language" and "common background knowledge", annotators were also supplied with supplementary notes where needed. Moreover, a field for adding comments on problematic and difficult to resolve cases was also catered for in the annotation interface.

As soon as the annotation process was completed, the majority of the entailment pairs (61,5%) were easily disambiguated, i.e., those for which a consensus above a certain threshold (90%-100%) was attested. Another 37,4% required further consideration and experts' annotation needed to be taken into account, since it was obvious that laymen's annotation was hasty or due to insufficient guidelines, background knowledge, etc. After multiple passes over the data, the value "unknown" was assigned to *T-H* pairs where disambiguation was still impossible, i.e. where the observed agreement between experts and laymen was equal to or less than 50%. These T-H pairs can be safely excluded when accuracy is calculated, since no "yes" or "no" value can be assigned to them. In the table below annotator agreement is scaled from the highest (90%-100%) to lowest (≤ 50%).

Annotator Agreement	Corpus percentage
90%-100%	61,5%
60%-80%	30,4%
≤ 50%	7%

Table 10: Annotator Agreement

A closer inspection of the data and the comments provided by laymen in the relative slot was performed for detecting sources of problematic cases for both experts and laymen. This qualitative analysis became, also, the basis of our preliminary study for the development of an RTE system for the Greek language. Case analysis showed that certain phenomena, such as ellipsis of basic syntactic constituents (e.g. verb, subject) or semantically important complements confuse annotators and influence their judgment causing, thus, a significant reduction in accuracy. Additionally, most of the annotators commented, as one might expect, on the awkward and, in some cases ungrammatical output of MT systems. Furthermore, they seemed to wonder about the boundaries of the knowledge that is considered to be common and also the level on which they should be influenced by it. In the enrichment of the corpus such cases should either be excluded or resolved beforehand.

4.2 Annotation at various linguistic levels

To render the corpus a useful resource to prospective system developers, further annotations were integrated semi-automatically via an existing pipeline of shallow processing tools for the Greek language. These include:

- Handling and tokenization; following common practice, the tokenizer makes use of a set of regular expressions, coupled with precompiled lists of abbreviations, and a set of simple heuristics (Papageorgiou et al., 2002) for the recognition of word and sentence boundaries, abbreviations, digits, and simple dates.
- POS-tagging & lemmatization; a tagger that is based on Brill's TBL architecture (Brill, 1997), modified to address peculiarities of the Greek language (Papageorgiou et al., 2000) was used in order to assign morphosyntactic information to tokenised words. Further more, the tagger uses a PAROLE-compliant tagset of 584 different part-of-speech tags (Lambropoulou et al., 1996). Following POS tagging, lemmas are retrieved from ILSP's Greek morphological lexicon.
- Named Entity Recognition was then performed using MENER (Maximum Entropy Named Entity Recognizer), a system compatible with the ACE (Automatic Content Extraction) schema, catering for the recognition and classification of the following types of NEs: person (PER), organization (ORG), location (LOC) and geopolitical entity (GPE) (Giouli et al., 2006).
- Coreference Resolution. Two forms of anaphora are covered: intra-sentential, where coreferring expressions occur in the same sentence, and

inter-sentential, where the pronoun refers to an entity mentioned in a previous sentence. The coreference resolution component is aimed at the creation of coreferential chains and is based on the work of (Lappin & Leass, 1994).

- Deep semantic analysis is performed by means of a dependency parser that was employed for the syntactic representation of text fragments. The current implementation aimed at the syntactic analysis of EL data, exploits the MaltParser platform [Nivre et al., 2006], via which a memory-based dependency parser for Greek was trained on the Greek Dependency Treebank. The latter comprises data annotated at ILSP at several linguistic levels [Prokopidis et al., 2005].

Annotations at the afore-mentioned levels of linguistic analysis were applied automatically and were hand validated by expert linguists.

5. Towards a system for RTE in Greek

The corpus has been developed to serve as a resource for guiding training and testing of a system that recognizes Textual Entailment in Greek documents. Presently, we are exploring the role of syntax in RTE together with lexical analysis and shallow semantics (synonyms, antonyms).

To cope with the fact that high word overlap between T and H is not always a positive evidence for textual entailment, and since the appropriate lexical resources, such as WordNet, FrameNet, etc, are either sparse or non-existent for the Greek language, we opted for a system, which takes into account not only word overlap but also further lexical evidence such as lexical paraphrases, syntactic structure similarity based on dependency relations and shallow semantic analysis.

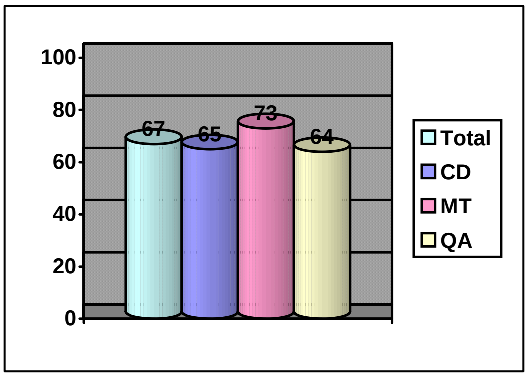


Figure 2: System Performance: Accuracy

Initially, the system uses extended lexical resources for spotting synonyms, antonyms and identifying structures in *T* and *H* sharing the same semantic content. Paraphrase acquisition is further enhanced with limited pattern recognition and the application of appropriate transformations. Where no other lexical relation is found, the system makes use of latent semantic analysis in order to assign scores or penalties accordingly. Finally, the system makes use of syntactic labels and dependency relations among constituents for the final estimation of similarity among *T* and *H* structures. Basically, our system "rewards" lexical similarity only when syntactic equivalence is also attested. The system is currently under development.

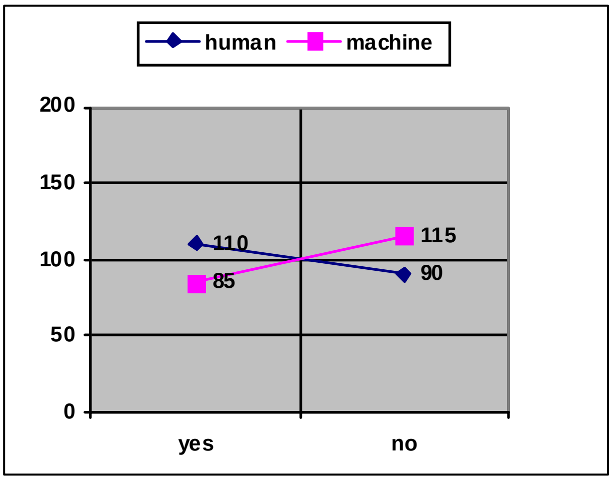


Figure 3: System performance: Recall

System performance was evaluated with the Precision and Recall metrics. Initial results show that a system, which takes syntactic analysis into consideration, can achieve satisfactory results. More precisely, the accuracy of the system reaches 67%, while recall amounts to 87%. As shown in figure 3, the system has a slight tendency to annotate the T-H pairs falsely as negative returning 25 false negatives.

6. Conclusions and Future work

A corpus has been described that was developed in view of training and evaluation of a system aimed at RTE in Greek, and the construction, annotation and evaluation phases have been presented. In order to cover efficiently various linguistic phenomena, the GTEC textual data are of different degrees of difficulty and pertain to various subject fields. Focusing on a limited set of applications seemed a good starting point for tackling textual entailment, whereas proper treatment of the remaining or even new applications, if necessary, remains to be seen in the future. Moreover, once a robust methodology for corpus creation and validation has been defined, enriching the corpus with new *T-H* pairs seems feasible.

The GTEC is currently being exploited for the development of an RTE system tailored to entailment recognition in Greek. It is expected to confirm the feasibility of handling textual entailment exploiting only shallow linguistic features. Our intention, however, is to investigate the role of deep semantic analysis in the RTE task and, also, to observe the system's behaviour when plugged in a real NLP application.

7. References

Dagan, I., Glickman, O., Magnini B. (2005). The PASCAL Recognising Textual Entailment Challenge. In Proceedings of the 1st Recognising Textual Entailment

Challenge.

Dagan, I., Bar-Haim R., Dolan B., Ferro L., Giampiccolo D., Magnini B. Szpektor I. (2006). The Second PASCAL Recognising Textual Entailment Challenge. In Proceedings of the 2nd Recognising Textual Entailment Challenge.

Papageorgiou H., Prokopidis P., Giouli V., Demiros I., Konstantinidis A., Piperidis St. (2002). Multi-level XML-based Corpus Annotation. In Proceedings of Proceedings of the 3nd Language and Resources Evaluation Conference.

Bayer S., Burger J., Ferro L., Henderson J., Yeh A. (2005). MITRE's Submissions to the EU Pascal RTE Challenge, In Proceedings of the 1st Recognising Textual Entailment Challenge.

Papageorgiou H., Prokopidis P., Giouli V., S. Piperidis. (2000). A Unified POS Tagging Architecture and its Application to Greek. In Proceedings of the 2nd Language and Resources Evaluation Conference, Athens, Greece, pp 1455-1462.

Lambropoulou, P., E. Mantzari, and M. Gavriilidou. (1996). Lexicon-Morphosyntactic Specifications: Language Specific Instantiation (Greek). PP-PAROLE, MLAP 63-386 report.

Voula Giouli, Alexis Konstandinidis, Elina Desypri, Harris Papageorgiou, (2006). Multi-domain Multi-lingual Named Entity Recognition: Revisiting & Grounding the resources issue. In Proceedings of LREC 2006.

Lappin, S. and H. J. Leass. (1994). An Algorithm for Pronominal Anaphora Resolution. In Computational Linguistics 20 (4), 535-561.

Prokopidis, P., Desypri, E., Koutsombogera, M., Papageorgiou, H., and Piperidis, S. (2005). Theoretical and practical issues in the construction of a Greek Dependency Treebank. In Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005). Barcelona, Spain, pp. 149–160.

Nivre, J., J. Hall and J. Nilsson (2006) MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006), May 24-26, 2006, Genoa, Italy, pp. 2216-2219