Decomposing Semantic Inferences

Elena Cabrio, Bernardo Magnini

Linguistics **I**ssues in **L**anguage **T**echnology – LiLT

The Semantics of Entailment

Edited by Cleo Condoravi and Annie Zaenen

CENTER FOR THE STUDY
OF LANGUAGE
AND INFORMATION

Contents

Decomposing Semantic Inferences

Elena Cabrio[1] and Bernardo Magnini[2]

Beside formal approaches to semantic inference that rely on logical representation of meaning, the notion of Textual Entailment (TE) has been proposed as an applied framework to capture major semantic inference needs across applications in Computational Linguistics. Although several approaches have been tried and evaluation campaigns have shown improvements in TE, a renewed interest is rising in the research community towards a deeper and better understanding of the core phenomena involved in textual inference. Pursuing this direction, we are convinced that crucial progress will derive from a focus on decomposing the complexity of the TE task into basic phenomena and on their combination. In this paper, we carry out a deep analysis on TE data sets, investigating the relations among two relevant aspects of semantic inferences: the logical dimension, i.e. the capacity of the inference to prove the conclusion from its premises, and the linguistic dimension, i.e. the linguistic devices used to accomplish the goal of the inference. We propose a decomposition approach over TE pairs, where single linguistic phenomena are isolated in what we have called *atomic inference pairs*, and we show that at this granularity level the actual correlation between the linguistic and the logical dimensions of semantic inferences emerges and can be empirically observed.

1 Introduction

The ability to carry out semantic inferences is pervasive in our capacity to understand natural languages. In particular, we show a crucial skill in establishing meaningful relations among different pieces of text in order to reconstruct their connections: as an example, the meaning of one portion of text can be expressed by another portion of text (i.e.

[1]INRIA Sophia Antipolis, France.
[2]Fondazione Bruno Kessler, Trento, Italy.

LiLT Volume 9, Issue 1, July 2013.
The Semantics of Entailment.

paraphrasing), it can be contained (i.e. entailed) by the other, it can be interpreted as the cause or the effect, or it can express the fact that it temporally precedes or follows the other. From a computational perspective, it seems difficult for any automatic system not to aim at replicating some degree of human semantic inferencing.

While the logical nature of such semantic inferences has been the subject of a huge amount of literature in the area of Philosophy of Language, it is only in the recent years that this topic has produced new trends of investigation in Computational Linguistics. A relevant achievement has been the focus on automatically recognizing "textual inferences" as the main research goal, which has let to the set-up of a general framework of research, independent from the actual methods used to address the problem. Focusing on the discovery of semantic relations among two portions of text has in fact opened the way to a number of new approaches and techniques, as well as to the development of several annotated data sets.

The renaissance of interest around semantic inferences in Computational Linguistics is well shown by several initiatives. Among them, the Recognizing Textual Entailment initiative (RTE) (Dagan et al., 2009), started in 2005 with the organization of the RTE series of evaluation campaigns[3], the semantic text similarity task at Semeval[4], and the recognition of causal relations[5]. A common feature of the above mentioned initiatives is that they all define semantic inferences as a direct relation among two portions of text. This distinguishes them from several annotation tasks (e.g. Part of Speech Tagging, Named Entity Recognition, Semantic Role Labeling), where the goal is the detection of linguistic phenomena within a single portion of text. The text-based approach to inferences has also made it easier to integrate several current research tools for text annotation in the service of inference detection.

As mentioned, establishing the inference tasks at the level of text, thus independently from the actual method implemented, has opened the door to a new research stream. New initiatives are pursing this approach to create shared and open platforms.[6] A relevant effect of this text-based view on semantic inferences is that much more annotated material is currently available for investigating the linguistic phenomena underlying semantic inferences. In addition, several approaches are now using such data sets for training automatic systems based on machine learning algorithms.

[3] `http://www.nist.gov/tac/2011/RTE/`
[4] `http://www.cs.york.ac.uk/semeval-2013/task6/`
[5] `http://www.cs.york.ac.uk/semeval-2012/task7/`
[6] `http://www.excitement-project.eu/`

While this paper takes advantage of the text-based framework in semantic inferences, and builds on top of the impressive progress in this area, we think that a deeper analysis of the current available data sets is still required, as it may bring new insight for further technological developments. Specifically, we notice that most of the current annotated data sets for the Textual Entailment task have been mainly developed according to applications criteria (e.g. in RTE-1-4 pairs are selected from relevant application domains; RTE-5-6 mainly serve summarization purposes; AVE[7] data sets (Peñas et al., 2008) come from Question Answering, etc.). Although this may serve the purpose of creating training material for specific application scenarios, overall, less attention has been paid to the analysis of the linguistic phenomena underlying textual inferences and the way they interact with different types of inferences. A consequence of the current lack of analysis is that it is not fully clear what a system can actually learn from the available data sets.

In the light of the above considerations, the purpose of this paper is to carry out a deep analysis of Textual Entailment (TE) data sets. We investigate the relations among two relevant aspects of semantic inferences: the *logical* dimension, i.e. the capacity of the inference to prove the conclusion from its premises, and the *linguistic* dimension, i.e. the linguistic devices that are used to accomplish the goal of the inference.

With respect to other studies - see, for instance, Garoufi (2007) and Sammons et al. (2010) - that have annotated and investigated TE datasets, we take a data oriented and neutral approach. As an example, we do not assign a polarity to single linguistic phenomena, and we do not impose specific categorizations on positive and negative entailment, rather we expect to derive such distinctions from observations.

According to this perspective, we aim at understanding whether there are regularities (i.e. relevant patterns) that might be learned combining the two dimensions. In the paper we show that the sparseness of the linguistic phenomena in current data sets and their distribution in positive and negative pairs, actually constitute an intrinsic limitation to supervised approaches to TE. Given this, we plead for a *decomposition framework* of semantic inferences in order to facilitate both a deeper understanding of the distribution of the phenomena that contribute to the inference, and to simplify the computational complexity of the problem. In this framework systems can learn from *specialized data sets*, covering both the most relevant phenomena underlying inferences and

[7] `http://nlp.uned.es/clef-qa/ave/`

the different nature of the inferences.

In the paper we systematically analyze a data set of TE pairs according to two relevant dimensions: *(i)* the nature of the inference, using the traditional logical view on arguments (Section 3); *(ii)* the linguistic phenomena involved in the inference (Section 4). In both sections we first provide the necessary background, and then we apply the analysis to a TE data set that we use throughout the paper. Section 5 presents a novel approach aiming at producing inference data sets where single linguistic phenomena are isolated one at a time. Through the decomposition of an initial RTE pair we obtain all the *atomic* pairs involved in the inference process, each tagged with the corresponding phenomenon. We show that the fine-grained analysis allowed by atomic pairs is a powerful investigation tool, which sheds new light on the relations between the polarity of a certain linguistic phenomenon and the occurrence of that phenomenon in both positive and negative pairs. Such analysis provides evidence that current RTE data sets offer a limited capacity to discriminate features that may support learning algorithms, particularly because the polarity of several linguistic phenomena correlates poorly with their distribution in positive and negative pairs. Finally, we conclude the paper recommending a systematic development of *specialized* data sets of atomic pairs and learning approaches over them.

2 Inference data sets

This section first presents the current status of RTE data sets, then describes other data sets used by the community for semantic inferences, and finally introduces the data set we have used for the analysis carried out in this paper.

In 2005, the PASCAL Network of Excellence started an attempt to promote a generic evaluation framework covering semantic-oriented inferences needed for practical applications, launching the Recognizing Textual Entailment challenge (Dagan et al., 2005), (Dagan et al., 2006), (Dagan et al., 2009), with the aim of setting a unifying benchmark for the development and evaluation of methods that typically address similar problems in different, application-oriented, manners. As many of the needs of several Natural Language Processing (NLP) applications can be cast in terms of TE, the goal of the evaluation campaign is to promote the development of general entailment recognition engines, designed to provide generic modules across applications. Since 2005, such initiative has been repeated yearly[8], asking the participants to develop

[8] `http://aclweb.org/aclwiki/index.php?title=Recognizing_Textual_Entailment`

a system that, given two text fragments (the *text* T and the *hypothesis* H), can determine whether the meaning of one text is entailed, i.e. can be inferred, from the other. Example 1.1 represents a positive example pair (i.e. *entailment*), where the entailment relation holds between T and H (pair 10, RTE-4 test set). For pairs where the entailment relation does not hold between T and H, systems are required to make a further distinction between pairs where the entailment does not hold because the content of H is contradicted by the content of T (i.e. *contradiction*, see Example 1.2 - pair 6, RTE-4 test set), and pairs where the entailment cannot be determined because the truth of H cannot be verified on the basis of the content of T (i.e. *unknown*, see Example 1.3 - pair 699, RTE-4 test set).

(1.1) T: *In the end, defeated, Anthony committed suicide and so did Cleopatra, according to legend, by putting an asp to her breast.*
H: *Cleopatra committed suicide.*

(1.2) T: *Reports from other developed nations were corroborating these findings. Europe, New Zealand and Australia were also beginning to report decreases in new HIV cases.*
H: *AIDS victims increase in Europe.*

(1.3) T: *Proposals to extend the Dubai Metro to neighbouring Ajman are currently being discussed. The plans, still in the early stages, would be welcome news for investors who own properties in Ajman.*
H: *Dubai Metro will be expanded.*

In line with the rationale underlying the RTE challenges, T-H pairs are collected from several application scenarios (e.g. Question Answering, Information Extraction, Information Retrieval, Summarization), reflecting the way by which the corresponding application could take advantage of automated entailment judgment. In the collection phase, each pair of the data set is judged by three annotators, and pairs on which the annotators disagree are discarded. The obtained data set is split into training and test data sets (note that most of the participating systems implement Machine Learning approaches requiring training data), containing on average about 1000 pairs each. The distribution according to the three-way annotation, both in the individual setting and in the overall data sets, is: 50% *entailment*, 35% *unknown*, and 15% *contradiction* pairs.[9]

[9]Since RTE-6, the task has been partially changed, and consists in finding all the sentences that entail a given H in a given set of documents about a topic (i.e.

Entailment in RTE pairs is defined as the inference a speaker with basic knowledge of the world would make. Entailments are therefore dependent on linguistic knowledge, and may also depend on some world knowledge - see the controversy between Zaenen et al. (2005) and Manning (2006). Partially guided by reasons of convenience for the task definition, some assumptions have been defined by the organizers of the challenge, for instance, the a priori truth of both T and H, and the sameness of meaning of entities mentioned in T and H. From a human perspective, the inference required are fairly superficial, since generally no long chains of reasoning are involved. However some pairs are designed to trick simplistic approaches.

Since the goal of RTE data sets is to collect inferences needed by NLP applications while processing real data, the example pairs are very different from a previous resource built to address natural language inference problems, i.e. the FraCas test suite (Cooper et al., 1996). This resource includes 346 problems, containing each one or more premises and one question (i.e. the goal of each problem is expressed as a question). With respect to RTE pairs, here the problems are designed to focus on a broader range of semantic and inferential phenomena, including quantifiers, plurals, anaphora, ellipsis and so on, as shown in Example 1.4 (fracas-022: monotonicity, upwards on second argument)[10].

(1.4) P1: *No delegate finished the report on time.*
Q: *Did no delegate finish the report?*
H: *No delegate finished the report.*
Answer: *unknown*
Why: *can't drop adjunct in negative context*

Even if the FraCas test suite is much smaller when compared to the number of annotated pairs in RTE data sets, and it is less natural-seeming (i.e. it provides textbook examples of semantic phenomena,

the corpus). This task is situated in the summarization application setting, where *i)* H's are based on Summary Content Units (Nenkova et al., 2007) created from human-authored summaries for a corpus of documents about a common topic, and *ii)* the entailing sentences (T's), are to be retrieved in the same corpus from which the summaries were made. Data sets for this task are therefore very different from the previous edition of the challenge, since there are no predefined T-H pairs.

[10]In the example, P and Q are respectively the premises and the question from the original source problem. The H element contains a sentence which is, as nearly as possible, the declarative equivalent to the question posed in the Q element. B. MacCartney (Stanford University) converted FraCas questions into declarative hypothesis: `http://www-nlp.stanford.edu/~wcmac/downloads/fracas.xml`

quite different from the kind of inferences that can be found in real data), it is worth mentioned here.

Another available inference data set that we are aware of is the Microsoft Research Paraphrase Corpus[11], that contains 5800 pairs of sentences which have been extracted from news sources on the web, and then manually annotated as paraphrase/semantic equivalence. Moreover, other inference data sets have been built to train automatic systems in the following NLP challenges: *i)* for the Answer Validation Exercise (AVE) at the Cross-Language Evaluation Forum (CLEF), systems have to consider triplets (Question, Answer, Supporting Text) and decide whether the Answer to the Question is correct and supported or not according to the given Supporting Text. Resources containing such triplets have been built for training and testing the participating systems, both for Spanish and for English languages[12]; *ii)* for the Semantic Textual Similarity task at Semeval 2012[13], where systems are asked to examine the degree of semantic equivalence between two sentences, the data set comprises pairs of sentences drawn from the publicly available data sets used in training (e.g. Microsoft Paraphrase, WMT2008 development data set - Europarl section[14], pairs of sentences where the first comes from Ontonotes and the second from a WordNet definition, and so on). In both competitions, most of the approaches implement Machine Learning methods, that try to exploit training set data for learning.

Since the work we present in this paper focuses in particular on Textual Entailment, the data we consider for our analysis include a sample of pairs extracted from RTE-5 data set (Bentivogli et al., 2009b). More specifically, in order to compare our results with the literature, we created our reference data joining the data sets annotated by Sammons et al. (2010) (composed of 210 pairs from RTE-5 test set: 107 *entailment*, 37 *contradiction*, 66 *unknown*) and by Bentivogli et al. (2010) (composed of 90 pairs from RTE-5: 30 *entailment*, 30 *contradiction*, 30 *unknown*). Since the two data sets have a lot of pairs in common, joining the two results in 243 pairs, divided into 117 positive (i.e. *entailment*), and 126 negative (i.e. 51 *contradiction* and 75 *unknown*) pairs. With respect to RTE-5 sub tasks (IE, IR and QA), such pairs are distributed as follows: 91 QA, 74 IE and 75 IR. From now on, we consider

[11] http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/

[12] http://nlp.uned.es/clef-qa/ave/

[13] http://www.cs.york.ac.uk/semeval-2012/task6/

[14] http://www.statmt.org/wmt08/shared-evaluation-task.html

this data set as the reference data for our study (we will refer to it as "RTE-5-SAMPLE"), on which the annotation and the experiments described in the next sections are carried out.

3 Analyzing semantic inferences by their logical nature

TE can be seen as the capacity to capture the strength of an inference (i.e. how much the conclusion can be inferred from the premises). We have found appropriate for our purposes the four validity criteria described in (Nolt et al., 1998): *truth of premises*, *validity and inductive probability*, *relevance*, *requirement of total evidence*. In our analysis, we apply such criteria to a sample of RTE pairs, aiming at understanding whether there are regularities (i.e. relevant patterns) that might be learned combining the *logical* dimension with the *linguistic* dimension of semantic inferences.

3.1 Semantic inferences as logical arguments

The main purpose of an argument is to demonstrate that a conclusion is true or at least likely to be true. It is therefore possible to judge an argument with respect to the fact that it accomplishes or fails to accomplish this purpose. In Nolt et al. (1998), four criteria for making such judgments are examined: *i)* whether the premises are true; *ii)* whether the conclusion is at least probable, given the truth of the premises; *iii)* whether the premises are relevant to the conclusion; and *iv)* whether the conclusion is vulnerable to new evidence.[15]

The motivations for criterion 1 (i.e. *truth of premises*) are related to the fact that if any of the premises of an argument is false, it is not possible to establish the truth of its conclusion. Often the truth or falsity of one or more premises is unknown, so that the argument fails to establish its conclusion "so far as we know". In such cases, we may suspend the judgment until relevant information that would allow us to correctly apply criterion 1 is acquired. Criterion 1 is a necessary - but not sufficient - condition for establishing the conclusion, i.e. the truth of the premise does not guarantee that the conclusion is also true.

In a good argument, the premises must adequately support the conclusion, and the second and third criteria (i.e. *validity and inductive probability*, and *relevance*, respectively) are thought to assess this aspect. In particular, the goal of criterion 2 is to evaluate the arguments with respect to the probability of the conclusion, given the truth of the premises. According to this parameter, arguments are classified into

[15] In Section 3.2 examples for each criterion are presented and discussed.

three categories:

- *deductive arguments*, whose conclusion follows *necessarily* from their basic premises (i.e. it is impossible for their conclusion to be false while the basic premises are true);
- *inductive arguments*, whose conclusion does not necessarily follow from their basic premises (i.e. there is a certain probability that the conclusion is true if the premises are, but there is also a probability that it is false)[16];
- *abductive arguments*, where the reasoning goes from data description of something to a hypothesis that accounts for the reliable data and seeks to explain relevant evidence. From an observable Q and a general principle P $\supset$ Q we conclude that P must be the underlying reason that Q is true. We assume P because Q is true (Hobbs, 2008).

Given a set of premises, the probability of a conclusion is called *inductive probability*, and it is measured on a scale from 0 to 1. The inductive probability of a deductive argument is maximal, i.e. equal to 1, while the inductive probability of an inductive argument is (typically) less than 1. Although deductive arguments provide the greatest certainty (inductive probability = 1), in practice we must often settle for inductive reasoning, that allows for a range of inductive probabilities and varies widely in reliability. When the inductive probability of an argument is high, the reasoning of the argument is said to be *strong* or *strongly inductive*. On the contrary, it is said to be *weak* or *weakly inductive* when the inductive probability is low. There is no clear distinction line between strong and weak inductive reasoning, since these definitions can be context-dependent.

The inductive probability of an inductive argument depends on the relative strengths of its premises and conclusion. Nolt et al. (1998) claim that the strength of a statement is determined by what the statement says, i.e. the more it says, the stronger it is (regardless of the truth of its content). The truth of a strong statement is proved only under specific circumstances, while the truth of a weak statement can be verified under a wider variety of possible circumstances because its content is less specific.

For these reasons, the strength of a statement is approximately inversely related to its *a priori* probability, i.e. the probability prior or in the absence of evidence: the stronger the statement is, the less inherently likely it is to be true, while the weaker it is, the more probable it is.

[16]Nolt et al. (1998) highlight the fact that in the literature the distinction between inductive and deductive argument is not universal, and slightly different definitions can be found in some works.

Inductive arguments can be divided into two types: *i)* the *Humeian* arguments (after the philosopher David Hume who was the first to study them) require the presupposition that the universe or some aspects of it is or is likely to be uniform or law like (e.g. *generalization*, *analogy* and *causality*); and *ii)* the *statistical* arguments, which do not require this presupposition, and the conclusions are supported by the premises for statistical or mathematical reasons (e.g. *statistical syllogism* and *statistical generalization*).

Criterion 3 claims that any argument which lacks relevance (regardless of its inductive probability) is useless for demonstrating the truth of its conclusion (it is said to commit a *fallacy of relevance*).

One of the most important differences between inductive and deductive arguments concerns their vulnerability to new evidence, meaning that deductive arguments remain deductive when new premises are added, while the inductive probability of inductive arguments can be strengthened or weakened by the introduction of new information. For this reason, the criterion of *total evidence condition* stipulates that if an argument is inductive its premises must contain all known evidence that is relevant to the conclusion. Inductive arguments which fail to meet this requirement are said to commit the *fallacy of suppressed evidence*, that can be committed either intentionally or unintentionally.

3.2 Validation criteria applied to RTE pairs

In the light of the definitions provided in the previous section, we annotated our RTE-5-SAMPLE data set with respect to the argument evaluation criteria described in Section 3.1. In general, in TE we assume the fact that: *i)* if T and H refer to an entity x, the reference is the same (reinforcing the relevance criterion), and *ii)* T (i.e. the premise) is assumed to be true (criterion 1 is always satisfied).

According to the second evaluation criterion (i.e. validity and inductive probability), TE pairs are annotated as *deductive* (Example 1.5, pair id=414), *inductive* (Example 1.6, pair id=194), *abductive* (Example 1.7, pair id=224) or *not valid* (i.e. invalid argument, contradiction) (Example 1.8, pair id=11). Inductive arguments have also been annotated according to the subcategories of inductive reasoning following Nolt et al. (1998), i.e. *statistical syllogism*, *statistical generalization* (both statistical arguments), *inductive generalization*, *simple induction*, *analogy* and *causality* (i.e. Humeian arguments).

(1.5) *T: On February 24th the Swedish Royal Court announced that the Crown Princess Victoria was to be married in 2010 to her boyfriend and former fitness trainer Daniel Westling. Victoria,*

31, and Daniel, 35, have been in an relationship for 7 years. Since the wedding is to be held in the summer of 2010 [...]
H: Princess Victoria will get married in 2010.

(1.6) *T: SEOUL, South Korea - North Korea's state news agency says that leader Kim Jong Il observed the launch of the country's satellite. The Korean Central News Agency says in a reported dated Sunday that Kim visited the General Satellite Control and Command Center and observed the liftoff. North Korea launched a rocket Sunday that flew over Japan. [...]*
H: Kim Jong-il is the leader of North Korea.

(1.7) *T: Secretary of State of the Vatican City, Cardinal Tarcisio said that the Pope apologized for the way his remarks made during a speech at the University of Regensburg in Germany on September 12 2006 were interpreted saying, "the Holy Father is very sorry that some passages of his speech may have appeared offensive to Muslims and were interpreted in a way he hadn't intended them to be. [...]"*
H: The Pope works with Cardinal Tarcisio.

(1.8) *T: A Soyuz capsule carrying a Russian cosmonaut, an American astronaut and U.S. billionaire tourist Charles Simonyi has docked at the international space station. Russian cosmonaut Gennady Padalka manually guided the capsule to a stop ahead of schedule Saturday two days after blasting off from the Baikonur cosmodrome in Kazakhstan. [...]*
H: Charles Simonyi is a Russian cosmonaut.

With respect to criterion 3, (i.e. relevance) a pair is annotated as *not relevant* when such criterion is not satisfied, meaning that the text does not contain enough information to infer the truth of the hypothesis (a *fallacy of relevance* is committed), as in Example 1.9 (pair id=100).

(1.9) *T: A South Korean official expressed doubts over United Nations Secretary-General Kofi Annan's apparent support for a permanent Security Council seat for Japan, and attention has been drawn to widespread mistrust of Japan by Chinese - although the Chinese government has not commented directly against Japan.*
H: China won't receive money from Japan.

With respect to criterion 4 (i.e. total evidence condition), a pair is annotated as *lack of total evidence* when it commits the *fallacy of suppressed evidence*, i.e. some information is omitted in the premises

due to lack of knowledge (Example 1.10, pair id=49). When pairs are annotated as *deductive*, *inductive* and *abductive*, we verify that criteria 3 and 4 are satisfied.

(1.10) *T: The earthquake happened at 0332 (0132 GMT), hours after a 4.6-magnitude tremor shook the area but caused no reported damage. Thousands of the city's 70,000 residents ran into the streets in panic during the 30 second tremor. A student dormitory was said to be one of the buildings badly damaged. [...] One student told Rai state TV that he managed to escape the building before the roof collapsed.*
H: A powerful earthquake strikes central Italy.

To assess the validity of the proposed annotation, a subset of RTE-5-SAMPLE (i.e. 90 pairs from RTE-5: 30 *entailment*, 30 *contradiction*, 30 *unknown*, Bentivogli et al. (2010)) has been independently annotated by another annotator with linguistic skills. To measure the inter-rater agreement we calculate the Cohen's kappa coefficient (Carletta, 1996), that is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. More specifically, Cohen's kappa measures the agreement between two raters who each classifies N items into C mutually exclusive categories. The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}, \tag{1.11}$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement then $\kappa = 1$. If there is no agreement among the raters other than what would be expected by chance (as defined by Pr(e)), $\kappa = 0$. For NLP tasks, the inter-annotator agreement is considered as significant when $\kappa > 0.6$. We applied the formula 1.11 to our data considering the six possible annotation tags listed above (i.e. *deductive*, *inductive*, *abductive*, *not valid*, *not relevant*, *lack of total evidence*), and the inter-annotator agreement results in $\kappa = 0.75$. As a rule of thumb, this is a satisfactory agreement. A closer look at the annotations produced by the two raters brings to light that while annotating a pair as *deductive* is straightforward, tagging a pair with respect to criteria 3 and 4 (i.e. as either *not relevant* or *lack of total evidence*) is not trivial, resulting in the highest disagreement between the annotators. Table 1 provides the results of the

annotation process, as resulting after a reconciliation phase carried out by the annotators.

Argument types		**RTE pairs**			
		TOT	**Ent**	**Contr**	**Unk**
Deductive		86	86	0	0
Inductive	statistical syllogism	31	0	0	0
	statistical generalization		2	0	1
	inductive generalization		5	0	2
	simple induction		11	1	2
	analogy		1	0	3
	causality		2	0	1
Abductive		22	10	0	12
not valid		47	0	47	0
not relevant		21	0	0	21
lack of total evidence		36	0	3	33
TOTAL		**243**	**117**	**51**	**75**

TABLE 1 Distribution of inferential phenomena in RTE-5-SAMPLE.

The four criteria for argument evaluation that we have applied to TE pairs have highlighted that Textual Entailment involves both deductive, inductive and abductive arguments, the first ones prevailing numerically on the other two (as can be seen in Table 1, 73% of the positive entailment pairs are deductive arguments). In particular, positive *entailment* pairs can be deductive arguments, inductive arguments with a strong inductive probability or abductive arguments. On the contrary, (almost) all *contradiction* pairs are invalid arguments (the premises do not support the conclusion). *Unknown* pairs can be either inductive arguments with a low inductive probability (i.e. 12%), abductive arguments (i.e. 16%), arguments committing the fallacy of relevance (i.e. 28%), or arguments committing the fallacy of suppressed evidence (44%). In general, abductive arguments are very infrequent in RTE data set, and can result both in entailment or in unknown pairs.

As introduced in Section 3.1, relevance is an essential criterion, even if simplifying assumptions have been made by RTE organizers (i.e. the same meaning of entities mentioned in T and H is assumed). The criterion of total evidence relates to the problem of background knowledge, since incomplete arguments require new evidence both to validate or invalidate the conclusion. The motivation underlying the proposal of a generic framework to model language variability has been source of misunderstandings, since the definition of TE does not set a clear distinction line between linguistic knowledge and world knowledge that is involved in such kind of reasoning. In the Recognizing Textual Entailment challenge, strategies to deal with this issue have been outlined, partially guided by reasons of convenience for the task definition. They

will be discussed in the next section.

4 Analyzing semantic inferences by linguistic and knowledge phenomena

This section analyses semantic inferences according to the linguistic and background knowledge phenomena present in both the premises and the conclusion of an argument, that are required to support the reasoning process. The goal is twofold: on one side, we aim at providing a fine-grained and data-driven classification of the linguistic and knowledge phenomena underlying the inference process. On the other hand, showing the distribution of such phenomena in real data gives indications on the expected capabilities of Textual Entailment systems.

4.1 Phenomena identification and classification

In line with the TE framework, addressing the inference task at a textual level opens different and new challenges from those encountered in formal deduction systems, where the arguments are already expressed in some formal meaning representation (e.g. first order logic) in the input. To identify implications in natural language sentences, automatic systems are therefore asked to deal with inductive reasoning, lexical semantic knowledge, and variability of linguistic expressions (Bos and Markert, 2006). Indeed, language variability manifests itself at different levels of complexity, and involves almost all linguistic phenomena of natural languages, including lexical, syntactic and semantic variation.

Although different levels of granularity can be used to define the inference sub-problems, we decided to group the phenomena using both fine-grained categories and broader categories (Bentivogli et al., 2010). Macro categories are defined referring to widely accepted linguistic categories in the literature (Garoufi, 2007), and to the inference types typically addressed in RTE systems: lexical, syntactic, lexical-syntactic, discourse and reasoning. Each macro category includes fine-grained phenomena, listed below. This list is not exhaustive and reflects the phenomena we detected in the sample of RTE-5 pairs we analyzed.[17]

- *lexical:* identity, format[18], acronymy, demonymy, synonymy, semantic opposition, hyperonymy, geographical knowledge;
- *lexical-syntactic:* nominalization/verbalization, causative, paraphrase, transparent heads;

[17] A definition of the listed phenomena, and examples for each category are available here: `http://www-sop.inria.fr/members/Elena.Cabrio/resources.html`

[18] Normalization of temporal or spatial expressions.

- *syntactic:* negation, modifier, argument realization, apposition, list, coordination, active/passive alternation;
- *discourse:* coreference, apposition, zero anaphora, ellipsis, statements;
- *reasoning:* apposition, modifiers, genitive, relative clause, elliptic expressions, meronymy, metonymy, membership/representativeness, reasoning on quantities, temporal and spatial reasoning, all the general inferences using background knowledge.

Some phenomena (e.g. apposition) can be classified in more than one macro category, according to their specific occurrence in the text. For instance, in Example 1.12 the apposition is considered as syntactic, while in Example 1.13 the apposition is classified into the category reasoning.

(1.12) T: *The government of Niger and Tuareg rebels of the Movement of Niger People for Justice (MNJ) have agreed to end hostilities [...].*
H: *MNJ is a group of rebels.*

(1.13) T: *Ernesto, now a tropical storm, made landfall along the coastline of the state of North Carolina [...].*
H: *Ernesto is the name given to a tropical storm.*

World knowledge is an omni-pervasive phenomenon (as discussed in Section 3.2). It has not been categorized separately.

4.2 Empirical analysis on RTE-5-SAMPLE

In order to assess the feasibility of the proposed approach, we annotated RTE-5-SAMPLE (described in Section 2), with the categories of entailment phenomena described in Section 4.1. The annotation has been carried out by two annotators with linguistic skills and inter-annotator agreement has been calculated on a subset of the annotated pairs[19] (i.e. 90 pairs, randomly extracted from the sample, and balanced with respect to *entailment*, *contradiction* and *unknown* pairs). A first measure of *complete* agreement was considered, counting when judges agree on all phenomena present in a given original T-H pair. The complete agreement on the full sample amounts to 64.4% (58/90 pairs). In order to account for partial agreement on the set of phenomena present in the T-H-pairs, we used the *Dice coefficient* (Dice, 1945).[20] The Dice coefficient is computed as follows:

[19] Same sample used to calculate the inter annotator agreement in Section 3.2.

[20] The *Dice coefficient* is a typical measure used to compare sets in IR and is also used to calculate inter-annotator agreement in a number of tasks where an

$$Dice = 2C/(A + B)$$

where C is the number of common phenomena chosen by the annotators, while A and B are respectively the number of phenomena detected by the first and the second annotator. Inter-annotator agreement on the whole sample amounts to 0.78. Overall, we consider this value high enough to demonstrate the stability of the (micro and macro) phenomena categories, thus validating their classification model. Table 2 shows inter-annotator agreement rates grouped according to the type of the original pairs, i.e. *entailment*, *contradiction* and *unknown* pairs.

The highest percentage of *complete* agreement is obtained on *unknown* pairs. This is due to the fact that since the H in *unknown* pairs typically contains information which is not present in (or inferable from) T, for 19 pairs out of 30 both the annotators agreed that no linguistic phenomena relating T to H could be detected.

	Complete	**Partial (Dice)**
entailment	60%	0.86
contradiction	57%	0.75
unknown	76%	0.68

TABLE 2 Agreement measures on linguistic phenomena per entailment type.

With respect to the Dice coefficient, the highest inter-annotator agreement can be seen for the *entailment* pairs, whereas the agreement rates are lower for *contradiction* and *unknown* pairs. This is due to the fact that for the *entailment* pairs, all the single phenomena are directly involved in the entailment relation, making their detection straightforward. On the contrary, in the original *contradiction* and *unknown* pairs not only the phenomena directly involved in the contradiction/unknown relation are to be detected, but also those preserving the entailment, which do not play a direct role on the relation under consideration (contradiction/unknown) and are thus more difficult to identify. To clarify this aspect, let's consider Example 1.14 (pair 125, marked as *contradiction*).

assessor is allowed to select a set of labels to apply to each observation. In fact, in these cases, and in ours as well, measures such as the widely used K are not good to calculate agreement. This is because K only offers a dichotomous distinction between agreement and disagreement, whereas what is needed is a coefficient that also allows for partial disagreement between judgments.

(1.14) T: *Mexico's new president, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers. He's appointed new people to key military [...]*
H: *Felipe Calderon is the outgoing President of Mexico.*

The phenomena that should be detected in order to correctly judge the pair are: *argument realization*, *apposition* and *semantic opposition*. While the phenomenon that triggers the contradiction is the semantic opposition, (*new* $\not\Rightarrow$ *ongoing*) the other two phenomena contribute to the inference process, and should be taken into consideration to reach a decision about the entailment label. Contrary to the semantic opposition, in this example both the argument realization (*Mexico's new president* $\Rightarrow$ *new president of Mexico*) and the apposition (*Mexico's new president Felipe Calderon* $\Rightarrow$ *Felipe Calderon is Mexico's new president*) would support the entailment.

The distribution of the phenomena present in RTE-5-SAMPLE, as resulting after a reconciliation phase carried out by the annotators, is shown in Table 3. The total number of occurrences of each specific phenomenon is given in the Column *TOT*, while in the next columns we report the number of occurrences of each specific phenomenon in *entailment* pairs (Column *E*), and in negative examples, i.e. *contradiction* and *unknown* pairs (Columns *C* and *U*, respectively).

A number of remarks can be made on the data presented in Table 3. Both macro categories and fine-grained phenomena are well represented but show a different absolute frequency: some have a high number of occurrences, whereas some others occur very rarely. To highlight the main features and the points of strengths of our annotation strategy, we compare it with two relevant works in the literature, i.e. Garoufi (2007) and Sammons et al. (2010).

In Garoufi (2007), a scheme for manual annotation of textual entailment data sets (ARTE) is proposed, with the aim of highlighting a wide variety of entailment phenomena in the data. ARTE views the entailment task in relation to three levels, i.e. *Alignment*, *Context* and *Coreference*, according to which 23 different features for positive entailment annotation are extracted. Each level is explored in depth for the positive entailment cases, while for the negative pairs a more basic and elementary scheme is conceived. The ARTE scheme has been applied to the complete positive entailment RTE-2 test set (400 pairs, i.e. 100 pair of each task), and to a random 25% portion of the negative entailment test set, equally distributed among the four tasks (100 pairs, i.e. 25 pairs of each task). *Reasoning* is the most frequent feature appearing altogether in 65.75% of the annotated pairs: this indicates

Phenomena	**RTE Pairs**			
	TOT	**E**	**C**	**U**
Lexical:	**60**	**38**	**18**	**4**
Identity/mismatch	8	2	6	0
Format	2	0	2	0
Acronymy	7	6	1	0
Demonymy	4	4	0	0
Synonymy	18	14	3	1
Semantic opposition	4	0	4	0
Hypernymy	13	9	1	3
Geographical knowledge	4	3	1	0
Lexical-syntactic:	**38**	**29**	**5**	**4**
Transparent head	4	2	1	1
Nominalization/verbalization	11	7	3	1
Causative	1	0	1	0
Paraphrase	22	20	0	2
Syntactic:	**133**	**98**	**28**	**7**
Negation	1	0	1	0
Modifier	31	24	3	4
Argument Realization	26	21	4	1
Apposition	55	40	15	0
List	1	1	0	0
Coordination	10	7	1	2
Active/Passive alternation	9	5	4	0
Discourse:	**108**	**72**	**26**	**10**
Coreference	64	43	15	6
Apposition	4	4	0	0
Anaphora Zero	26	17	5	4
Ellipsis	9	5	4	0
Statements	5	3	2	0
Reasoning:	**147**	**91**	**43**	**13**
Apposition	4	3	1	0
Modifier	4	4	0	0
Genitive	2	1	1	0
Relative Clause	2	1	1	0
Elliptic Expression	1	1	0	0
Meronymy	6	3	2	1
Metonymy	4	4	0	0
Membership/representative	2	2	0	0
Quantity	9	3	5	1
Temporal	5	2	1	2
Spatial	1	1	0	0
Common background/ general inferences	107	66	32	9
TOTAL	**486**	**328**	**120**	**38**

TABLE 3 Distribution of linguistic phenomena in T-H original pairs (RTE-5-SAMPLE).

that a significant portion of the data involves deeper inferences. The combination of the entailment features is analyzed together with the entailment types and their distribution in the data.

More recently, Sammons et al. (2010) carried out an annotation work that is very similar in spirit to the approach proposed in Bentivogli et al. (2010), and that we extend in this work. Highlighting the need of resources for solving textual inference problems in the context of RTE, the authors challenge the NLP community to contribute to a joint, long term effort in this direction, making progress both in the analysis of relevant linguistic phenomena and their interaction, and developing resources and approaches that allow more detailed assessment of RTE systems. The authors propose a linguistically-motivated analysis of entailment data, based on a step-wise procedure to resolve entailment decision, by first identifying parts of T that match parts of H, and then identifying connecting structures. Their inherent assumption is that the meanings of T and H could be represented as sets of n-ary relations, where relations could be connected to other relations (i.e. could take other relations as arguments). The authors carried out a feasibility study applying the procedure to 210 examples from RTE-5 (the same that we also included in RTE-5-SAMPLE), marking for each example the entailment phenomena that are required for the inference.[21]

Both our annotation methodology and the ones adopted in these related works attempt to align (or transform) textual snippets of T into H, highlighting all the phenomena that trigger such alignment (or transformation). We all consider levels beyond bags of words, taking syntactic structure into account (depending on the granularity of the phenomena). The direction of the alignment is from H to T, so that H is covered exhaustively while T may contain irrelevant parts that are not aligned. Differently from Sammons et al. (2010), both the annotation we and Garoufi (2007) provide consists in marking the phenomena in the text allowing an easy individuation and their isolation. With respect to the choice of the categories to cluster the phenomena, our work is more similar to Garoufi (2007), since we both rely on more "standard" linguistic categories, even if our classification is more fine-grained (they cluster their categories according to three upper levels, i.e. *Alignment*, *Context* and *Coreference*).

Sammons et al. (2010) propose instead an ontology of phenomena that is iteratively hypothesized and refined while proceeding in the annotation phase, with the goal of identifying: *i)* the roles for background knowledge in terms of domains and general inference steps, *ii)* the lin-

[21] `https://agora.cs.illinois.edu/display/rtedata/Annotation+Resources`

guistic phenomena involved in representing the same information in different ways, or *iii)* detecting the key differences in two similar fragments. The resulting set of labels have less strict definitions with respect to well-established linguistic categories, and are often not very intuitive to understand. More recently, their Entailment Phenomena Ontology has been revised, and the new proposed annotation adopts more standard labels.[22] Since their categories are not mutually exclusive (and some levels of annotation are transversal with respect to the others, e.g. *domain*), their classification of the phenomena turns out to be more fuzzy, and complex to map on ours for a comparison. Another difference with respect to our approach lies in the fact that we annotate only the differences between T and H (i.e. if two fragments are equal in T and H we do not consider them), while they annotate also the cases of equal Named Entities (NE) in the two sentences.

For instance, given Example 1.15 (pair 6), we annotate it with one linguistic phenomenon, i.e. *syntax:modifier* (*respected traditional healer* ⇒ *healer*), while Sammons et al. (2010) annotate it as *hyp_has_NE* and *work* (to identify the domain). According to our intuition, in this case their annotation fails to circumscribe the phenomenon that should actually be tackled by a TE system to solve the entailment and provide the correct label to the pair.

(1.15) T: *Rain is pelting down on Doa Porcela's treatment room in Puerto Cabezas, the main town on Nicaragua's Northern Caribbean coast. [...] Doa Porcela is a respected traditional healer here and the bottles are filled with her secret medicinal potions. [...]*
H: *Doa Porcela is a healer.*

Differently from our approach, both Garoufi (2007) and Sammons et al. (2010) add a list of phenomena that are peculiar to negative cases. The former classifies the negative entailment cases into three major categories, according to the most prominent and direct reason why the entailment cannot be established. In particular, they focus on the single phenomenon that they consider as the most obvious "trap" for systems (and humans) judging the entailment. In those negative examples, they do not consider all the other phenomena that are part of the inference process (as we do), omitting some steps that are required while reasoning on such pairs. Also Sammons et al. (2010) define an apriori polarity of the phenomena, adding a set of categories for the negative entailment phenomena, or for missing relations between T

[22] `https://wiki.engr.illinois.edu/display/rtedata/Revised+Entailment+Phenomena+Ontology`

and H (e.g. *missing modifier*, or *missing argument*).

In our approach the linguistic categories are neutral (except *semantic opposition*), and we detect the polarity of the phenomena from their occurrences in the data, depending on whether the phenomenon supports the entailment or the contradiction judgment in a certain pair. For instance, in example 1.15 the phenomenon *syntax:modifier* supports the entailment relation (*respected traditional healer* $\Rightarrow$ *healer*), but if T and H were inverted, it would have triggered a negative judgment (i.e. *healer* $\not\Rightarrow$ *respected traditional healer*).

As in Garoufi (2007), our study confirms that a huge amount of background knowledge and reasoning is required to face the RTE task, given the fact that phenomena belonging to the category *reasoning* are the most frequent. LoBue and Yates (2011) have attempted to characterize them proposing 20 categories of common-sense knowledge that are prevalent in TE. Their categories can be loosely organized into *form-based* categories (e.g. *cause and effect*, *simultaneous conditions*) and *content-based* categories (e.g. *arithmetic*, *has parts)*. While some of their fine-grained categories can be mapped to ours (e.g. *arithmetic=quantity* and *has parts= meronymy*), we plan to extend our annotation of the *reasoning* phenomena adopting some of the labels they propose, to subcategorize the phenomena we annotated as *reasoning:general_inference*.

5 Analyzing semantic inference by decomposition

Basing ourselves on the classification of the phenomena previously described, in this section we go a step further, and decompose the complexity of TE focusing on single phenomena involved in the inference process. Our goal is to better understand the relations between the entailment judgments supported by each linguistic phenomenon in isolation and the overall judgment of the pair in which it occurs.

5.1 Towards total evidence

The underlying idea is to create *atomic pairs*, i.e. T-H pairs where a phenomenon relevant to the inference task is highlighted and isolated,[23] on the basis of the phenomena which are actually present in the RTE T-H pairs. As claimed before, one of the advantages of testing the proposed methodology on RTE data consists of the fact that the actual distribution of the linguistic phenomena involved in the entailment relation emerges. In Section 4.1 we proposed a classification of the phenomena we detected while analyzing a sample of RTE pairs,

[23] In Bentivogli et al. (2010), atomic T-H pairs are referred as *monothematic* pairs. In this work we decided to switch the terminology to be compliant with the theoretical framework we propose.

and we decided to group them using both fine-grained categories and broader categories. Grouping specific phenomena into macro categories would allow us to create specialized data sets of atomic pairs representing those phenomena, containing enough pairs to train and test TE systems. Macro categories are defined referring to widely accepted linguistic categories in the literature (Garoufi, 2007), and to the inference types typically addressed in RTE systems: lexical, syntactic, lexical-syntactic, discourse and reasoning.

Moreover, we assume that humans have knowledge about the linguistic phenomena relevant to TE, and that such knowledge can be expressed through *entailment rules* (Szpektor et al., 2007). An entailment rule is either a directional or bidirectional relation between two sides of a pattern, corresponding to text fragments with variables (typically phrases or parse sub-trees, according to the granularity of the phenomenon they formalize). The left-hand side of the pattern (LHS) entails the rights-hand side (RHS) of the same pattern under the same variable instantiation. In addition, a rule may be defined by a set of constraints, representing variable typing (e.g. PoS, NE type) and relations between variables, which have to be satisfied for the rule to be correctly applied. For instance, the entailment rule for demonyms can be expressed as:

Pattern: *X Y* $\Leftarrow / \Rightarrow$ *X (is) from Y*

Constraint: *DEMONYMY(X,Z)*
TYPE(X)= ADJ_NATIONALITY; *TYPE(Z)=GEO*

meaning that *X Y* entails *Y is from Z* if there is a *entailment* relation of demonymy between x and y, where x is an adjective expressing a nationality and z is a geographical entity (e.g. *A team of European astronomers* $\Leftarrow / \Rightarrow$ *A team of astronomers from Europe*, pair 205). The entailment rules for a certain phenomenon aim to be as general as possible, but for the cases in which the semantics of the words is essential (e.g. general inference), text snippets extracted from the data are used. Different rules can be needed in order to formalize the variants in which the same phenomenon occurs in the pairs. For example, the following entailment rules both formalize the phenomenon of apposition (syntax):

a) Pattern: *X Y* $\Leftrightarrow$ *Y X*
Constraint: *APPOSITION(Y,X)*

b) Pattern: *X, Y* $\Leftrightarrow$ *Y is X*
Constraint: *APPOSITION(Y,X)*

Given such basic concepts, the procedure for the creation of atomic pairs we propose consists of a number of steps carried out manually. We start from a T-H pair taken from the RTE data sets and we decompose T-H in a number of atomic pairs T-H_i, where T is the original Text and H_i are Hypotheses created for each linguistic phenomenon relevant for judging the entailment relation in T-H. The procedure is schematized in the following steps:

1. individuate the linguistic phenomena which contribute to the entailment in T-H
2. For each phenomenon i:
   (a) individuate a general entailment rule r_i for the phenomenon i, and instantiate the rule using the portion of T which expresses i as the LHS of the rule, and information from H on i as the RHS of the rule.
   (b) substitute the portion of T that matches the LHS of r_i with the RHS of r_i.
   (c) consider the result of the previous step as H_i, and compose the atomic pair $T - H_i$. Mark the pair with phenomenon i.
3. Assign an entailment judgment to each atomic pair.

After applying this procedure to the original pairs, all the atomic $T - H_i$ pairs relative to the same phenomenon i should be grouped together in a data set specialized for phenomenon i.

In the following, some examples of the application of the procedure to RTE pairs, namely entailment, contradiction and unknowns pairs are illustrated.

Decomposing entailment pairs.

Table 4 shows the decomposition of an original entailment pair (pair 199) into atomic pairs. In step 1 of the method, the phenomena (i.e. modifier, coreference, transparent head and general inference) are considered relevant to the entailment between T and H. In the following, we apply the procedure step by step to the phenomenon we define as modifier. In step 2a the general rule:

Entailment rule:	**modifier**
Pattern:	$X\ Y \Leftrightarrow Y$
Constraint:	*MODIFIER(X,Y)*
Probability:	1

is instantiated (*The tiny Swiss canton* $\Rightarrow$ *The Swiss canton*), while

in step 2b the substitution in T is carried out (*The Swiss canton of Appenzell Innerrhoden has voted to prohibit [...]*).

In step 2c the atomic pair $T-H_1$ is composed and marked as *modifier* (macro-category *syntactic*). Finally, in step 3, this pair is judged as *entailment*. Step 2 (a, b, c) is then repeated for all the phenomena individuated in that pair in step 1.

The same token can be an instance of several different phenomena. In such cases, in order to create an atomic H for each phenomenon, the method is applied recursively. It means that after applying it once to the first phenomenon of the chain (thereby creating the pair $T - H_i$), it is applied again to H_i (that becomes T') to solve the second phenomenon of the chain (creating the pair $T' - H_j$).

Decomposing contradiction pairs.

Table 5 shows the decomposition of an original contradiction pair (pair 125) into atomic pairs. In step 1 both the phenomena that preserve the entailment and the phenomena that break the entailment rules causing a contradiction in the pair should be detected. In the example reported in Table 5, the phenomena that should be recognized in order to correctly judge the pair are: argument realization, apposition and semantic opposition. While the atomic pairs created basing on the first two phenomena preserve the entailment, the semantic opposition generates a contradiction. In the following, we apply the procedure step by step to the phenomenon of semantic opposition.
In step 2a the general rule:

Contradiction rule:	**semantic_opposition**
Pattern:	$X \nLeftrightarrow Y$
Constraint:	*SEMANTIC_OPPOSITION(Y,X)*
Probability:	1

is instantiated (*new* $\nLeftrightarrow$ *outgoing*), and in step 2b the substitution in T is carried out (*Mexico's outgoing president, Felipe Calderon [...]*). In step 2c a negative atomic pair $T - H_1$ is composed and marked as semantic opposition (macro-category *lexical*), and the pair is judged as *contradiction*. We noticed that negative atomic T-H pairs (i.e. both contradiction and unknown) may originate either from the application of contradiction rules (e.g. semantic opposition or negation, as in pair $T - H_1$, in Table 5) or as a wrong instantiation of a positive entailment rule. For instance, the positive rule for active/passive alternation:

Text (pair 199 RTE-5 test set)			Rule	Phenomena	J.
T	**The tiny Swiss canton of Appenzell Innerrhoden has voted prohibit the phenomenon of naked hiking.** [...]				
H	The Swiss canton of Appenzell has prohibited naked hiking.			**synt:modifier, disc:coref, lsynt:tr_head, reas:gen_infer**	**E**
	H_1	**The Swiss canton** of Appenzell Innerrhoden has voted to prohibit the phenomenon of naked hiking.	**x y ⇒ y modif(x,y)**	**synt:modifier**	**E**
	H_2	The tiny Swiss canton of **Appenzell** has voted to prohibit the phenomenon of naked hiking.	**x⇔y coref(x,y)**	**disc:coref**	**E**
	H_3	The tiny Swiss canton of Appenzell Innerrhoden has voted to prohibit **naked**. voted to prohibit **hiking**.	**x of y ⇒y tr_head(x,y)**	**lsynt:tr_head**	**E**
	H_4	The tiny Swiss canton of Appenzell Innerrhoden **prohibited** the phenomenon of naked hiking.	**vote to prohibit (+ will now be fined) ⇒ prohibit**	**reas:gen_infer**	**E**

TABLE 4 Decomposition method applied to an *entailment* pair.

	Text (pair 408 RTE-5 test set)		Rule	Phenomena	J.
T	**Mexico's new president, Felipe Calderon,** seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]				**C**
H	Felipe Calderon is the outgoing President of Mexico.			**lex:sem_opp synt:arg_real synt:apposit**	
	H_1	Mexico's **outgoing** president, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's drug traffickers. [...]	$x \not\Leftrightarrow y$	**sem_opp(x,y)**	**C**
	H_2	The new president **of Mexico**, Felipe Calderon, seems to be doing all the right things in cracking down on Mexico's	**x's y⇒y of x**	**synt:arg_real**	**E**
		drug traffickers. [...]			
	H_3	**Felipe Calderon is** Mexico's **new president**.	**x,y ⇒ y is x apposit(y,x)**	**synt:apposit**	**E**

TABLE 5 Decomposition method applied to a *contradiction* pair.

Entailment rule:	**active/passive_alternation**
Pattern:	*X Y Z* ⇔ *Z W X*
Constraint:	*SAME_STEM(X,W)*
	TYPE(X)=V_ACT; *TYPE(W)=V_PASS*
Probability:	1

when wrongly instantiated, as in *Russell Dunham killed nine German soldiers* ⇏ *Russell Dunham was killed by nine German soldiers* (*X Y Z* ⇔ *Z W X*), generates a negative atomic pair.

Decomposing unknown pairs.

Table 6 shows the decomposition of an original unknown pair (pair 82) into atomic pairs. As in the previous cases, in step 1 all the relevant phenomena are detected: coreference, general inference, and modifier.

	Text (pair 82 RTE-5 test set)			**Rule**	**Phenomena**	**J.**
T	Currently, there is no specific treatment available against **dengue fever, which is the most widespread tropical disease after malaria.** [...] "Controlling **the mosquitos that transmit dengue** is necessary [...]"					
H	Malaria is the most widespread disease transmitted by mosquitos.				**disc:coref, r:gen_infer, synt:modif,**	**U**
	H_1 $\rightarrow T'$	**Dengue fever** is the most widespread tropical disease after malaria.		**x⇔y coref(x,y)**	**disc:coref**	**E**
		H_2	**Malaria is the most widespread** tropical disease.	**x is after y⇒ y is the first**	**r:gen_infer**	**E**
		H_3	Dengue fever is the most widespread disease **transmitted by mosquitos** after malaria.	**x =? ⇒ x y (restr. relat. clause)**	**synt:modif**	**U**

TABLE 6 Decomposition method applied to an *unknown* pair.

While the first two preserve the entailment relation, the atomic pair resulting from the third phenomenon is judged as unknown. As discussed in Section 3.1, the last atomic pair is an argument with a very low inductive probability (i.e. the fact that a certain disease is the most widespread among the ones transmitted by a certain cause, does not allow us to infer that it is the most widespread ever). If we try to apply the procedure step by step to the phenomenon of modifier, in step 2a

the generic rule:

Entailment rule:	**modifier**
Pattern:	$X \Rightarrow X\ Y$
Constraint:	*MODIFIER(Y,X)*
Probability:	0.1

is instantiated (*disease* ⇒ *disease transmitted by mosquitoes*) (this rule has a very low probability), and in step 2b the substitution in T is carried out. In step 2c the atomic pair T'-H_3 is composed and marked as *modifier* (restrictive relative clause, macro-category *lexical*), and the pair is judged as *unknown*. However, there is no reason to collect such rules for computational purposes, since it would mean to collect almost all the relations among all the words and the expressions of a language. These rules can be obtained in a complementary way with respect to high-probability rules, i.e. if a certain rule is not present among the highly probable ones, it means that it has a low probability, and therefore it is not strong enough to support the related inferential step.

5.2 Applying pair decomposition to RTE-5-SAMPLE

To assess the feasibility of the decomposition strategy, we applied the method described in Section 5.1 to RTE-5-SAMPLE. Table 7 reports both the distribution of the phenomena present in the original RTE-5 pairs (column *RTE pairs*, equal to Table 3), together with their distribution according to the entailment judgment they support (i.e. independently of the overall judgment of the pair, column *Atomic pairs*). Again, the total number of occurrences of each specific phenomenon is given (Column *TOT*), corresponding to the number of atomic pairs created for that phenomenon. The number of atomic pairs is then divided into positive examples, i.e. *entailment* atomic pairs (Column *E*), and negative examples, i.e. *contradiction* and *unknown* atomic pairs (Columns *C* and *U*, respectively).

Comparing the two distributions of the phenomena among E/C/U pairs, we can see that some phenomena appear more frequently or only among the positive examples (e.g. *apposition* or *coreference*) and others among the negative ones (e.g. *quantitative reasoning*). In general, the total number of positive examples is much higher than that of the negative ones and, for some macro-categories no negative examples are found. As can be seen when comparing the two main columns of Table 7, applying our decomposition strategy brings to light the fact that, for instance, all the *lexical-syntactic* phenomena occurring in the RTE

Phenomena		**RTE Pairs**			**Atomic Pairs**		
	TOT	**E**	**C**	**U**	**E**	**C**	**U**
Lexical:	**60**	**38**	**18**	**4**	**46**	**11**	**3**
Identity/mismatch	8	2	6	0	2	6	0
Format	2	0	2	0	2	0	0
Acronymy	7	6	1	0	7	0	0
Demonymy	4	4	0	0	4	0	0
Synonymy	18	14	3	1	18	0	0
Semantic opposition	4	0	4	0	0	4	0
Hypernymy	13	9	1	3	10	0	3
Geographical knowledge	4	3	1	0	3	1	0
Lexical-syntactic:	**38**	**29**	**5**	**4**	**38**	**0**	**0**
Transparent head	4	2	1	1	4	0	0
Nominalization/verbaliz.	11	7	3	1	11	0	0
Causative	1	0	1	0	1	0	0
Paraphrase	22	20	0	2	22	0	0
Syntactic:	**133**	**98**	**28**	**7**	**116**	**13**	**4**
Negation	1	0	1	0	0	1	0
Modifier	31	24	3	4	26	2	3
Argument Realization	26	21	4	1	26	0	0
Apposition	55	40	15	0	47	8	0
List	1	1	0	0	1	0	0
Coordination	10	7	1	2	9	0	1
Active/Passive alternation	9	5	4	0	7	2	0
Discourse:	**108**	**72**	**26**	**10**	**107**	**1**	**0**
Coreference	64	43	15	6	63	1	0
Apposition	4	4	0	0	4	0	0
Anaphora Zero	26	17	5	4	26	0	0
Ellipsis	9	5	4	0	9	0	0
Statements	5	3	2	0	5	0	0
Reasoning:	**147**	**91**	**43**	**13**	**112**	**29**	**6**
Apposition	4	3	1	0	3	1	0
Modifier	4	4	0	0	4	0	0
Genitive	2	1	1	0	2	0	0
Relative Clause	2	1	1	0	2	0	0
Elliptic Expression	1	1	0	0	1	0	0
Meronymy	6	3	2	1	5	1	0
Metonymy	4	4	0	0	4	0	0
Membership/represent.	2	2	0	0	2	0	0
Quantity	9	3	5	1	3	5	1
Temporal	5	2	1	2	4	0	1
Spatial	1	1	0	0	1	0	0
Common background/ general inferences	107	66	32	9	81	22	4
TOTAL (# atomic pairs)	**486**	**328**	**120**	**38**	**419**	**54**	**13**

TABLE 7 Distribution of linguistic phenomena in T-H original and atomic pairs (RTE-5-SAMPLE).

pairs we analyzed support the entailment judgment, even if they are present in contradiction or unknown pairs (it means that in those pairs other phenomena trigger the negative judgment). Also from a qualitative standpoint, we notice that compared to the positive pairs the variability of phenomena in negative examples is reduced.

The differences in the distributions of the phenomena when occurring in RTE pairs and with respect to the judgment they independently support, provide also an explanation about the non optimal results obtained by the ablation tests, introduced as a requirement for systems participating in RTE-5 and RTE-6 main tasks. Such ablation tests consist in removing one resource at a time from a TE system, and rerunning the system on the test set with the other modules, except the one tested. The results obtained from ablation tests turned out not to be straightforward in determining the actual impact of the resources, since the different uses made by the systems of the same resources, make it difficult to compare the results. Moreover, basing on our observations we can now demonstrate that evaluating for instance the impact of WordNet (Fellbaum, 1998) on original RTE pairs would be misleading, since lexical phenomena (as *synonymy*) can be found in both positive and negative pairs, but the phenomenon in itself always supports entailment (even when it is present in a contradiction pair).

To provide a stronger basis for our assumptions, we measured the correlation (linear dependence) between the two observed phenomena distribution. We applied the Pearson product-moment correlation coefficient[24] between the distribution of phenomena on original RTE pairs and in relation to the supported judgment. The Pearson correlation is +1 in the case of a perfect positive (increasing) linear relationship (correlation), -1 in the case of a perfect decreasing (negative) linear relationship (anticorrelation), and some value between -1 and 1 in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship (i.e. it is closer to uncorrelated). In our framework, obtaining a low correlation between the two distributions of a certain category of phenomena has to be interpreted as a proof of concept of our decomposition approach, since it would mean that training a TE system only on original pairs is misleading (i.e. the occurrence of a certain phenomenon is not always an indication of the judgment it bears). On the contrary, a high correlation between the two distributions would mean that the mere oc-

[24] `http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`. We calculated it on the normalized occurrences of phenomena, and using the open source software Wessa.net (Wessa, 2012)

currence of the phenomena in the original pairs is a sufficient condition to learn their judgment (i.e. atomic pairs are not necessary, TE systems would learn the same model when trained on both distributions).

Table 8 shows the correlation indexes we obtained per each macro-category of phenomena and per entailment judgment. The significance (P-value) for the Pearson's correlation is also reported.

Phenomena	**Ent**		**Contr**		**Unk**	
	corr.	*p<0.05*	**corr.**	*p<0.05*	**corr.**	*p<0.05*
Lexical	0.62	x	0.66	x	0.97	
Lex-synt	0	-	0	-	0	-
Syntactic	0.96	x	0.97	x	0.47	
Discourse	0.07		-0.06		0	-
Reasoning	0.62	x	0.55	x	0.34	

TABLE 8 Correlations per macro-categories of phenomena.

With the exception of the distributions of the syntactic phenomena that correlate well with the entailment and the contradiction judgment, the correlation values are pretty low, meaning that the linear dependence between the two distributions is not very strong. In several cases, it approaches 0 (e.g. for *lexical-syntactic* or for *discourse* phenomena), meaning that training a TE system on the occurrences of the linguistic phenomena in original RTE pairs only is not always reliable. In most of the cases, such correlation is statistically significant (the non-significance for unknown pairs is probably due to the low number of observations). Even for categories of phenomena with a strong correlation between the distributions, for some finer-grained phenomena belonging to those categories the difference between their occurrences in positive and negative pairs is particularly strong. For instance, the correlation index for syntactic phenomena approaches 1, but in Table 7 we can see that for *active passive alternation* the distribution in the two tables is very different, and a TE system trained on the first table would learn that 50% of the times this phenomenon triggers a contradiction, while it is not the case (it supports contradiction only in 20% of the pairs in which it occurs).

Cases of low correlation (e.g. *lexical-syntactic* phenomena) should not be interpreted, however, as absolute evidence that such phenomena are not useful at all as discriminators for textual entailment judgments. Rather, such correlations are always relative to the complexity of the pair: intuitively, the more the phenomena connecting T and H in the pair, the less relevant is a single low-correlated phenomenon. As a con-

sequence, the results presented in Table 8, hold for a data set whose complexity is similar to the RTE data we have analyzed, and could change in case of pairs with a different complexity.

With respect to the approaches proposed by Garoufi (2007) and Sammons et al. (2010), our methodology goes a step further suggesting to decompose the pairs to highlight and isolate the linguistic and knowledge phenomena relevant to semantic inference. Carrying out such decomposition allows for a level of analysis not possible following current methodologies. In particular, the approach of Garoufi (2007) allows for the identification of the phenomena in the text, but, on contradiction and unknown pairs, all the phenomena not triggering these judgments are ignored, so it is not possible to have a clear view of their distributions in the pairs. Sammons et al. (2010) assign an apriori polarity to the phenomena to compensate for the need for a clear distinction between the occurrences of the phenomena in positive or in negative pairs. Instead our approach is grounded in a clearer and standard classification of the phenomena, where their polarity emerges from their occurrences in the data and is not apriori defined. Moreover, beside the annotation of the phenomena on real data, the decomposition method results in the creation of atomic pairs, allowing evaluations of TE systems on specific phenomena both when isolated and when interacting with the others.

As introduced before, due to the natural distribution of phenomena in RTE data, we found that applying the decomposition methodology we generate a higher number of atomic positive pairs (76.7%) than negative ones (23.3%, divided into 17% *contradiction* and 6.3% *unknown*, as shown in Table 7). We analyzed the three subsets composing the RTE-5 sample separately, (i.e. 107 *entailment* pairs, 37 *contradiction* pairs, and 66 *unknown*) in order to verify the productivity of each subset with respect to the atomic pairs created from them. Table 9 shows the absolute distribution of the atomic pairs among the three RTE-5 classes.

RTE-5 pairs	**Generated atomic pairs**			
	E	**C**	**U**	**Total**
E (117)	328	–	–	328/117 (2.8)
C (51)	66	54	–	120/51 (2.35)
U (75)	25	–	13	38/21 (1.8)

TABLE 9 Distribution of the atomic pairs wrt original E/C/U pairs.

When the methodology is applied to RTE-5 *entailment* examples, averagely 2.8 all positive atomic pairs are derived from the original pairs. When the methodology is applied to RTE-5 *contradiction* examples, we create an average of 2.35 atomic pairs, among which 1.29 are entailment pairs and 1.05 are contradiction pairs. This means that the methodology is productive for both positive and negative examples.

As introduced before, in 54 out of 75 *unknown* examples no atomic pairs can be created, due to the lack of specific phenomena relating T and H (typically the H contains information which is neither present in T nor inferable from it). For the 11 pairs that have been decomposed into atomic pairs, we created an average of 1.8 atomic pairs, among which 1.19 are entailment and 0.61 are unknown pairs. This analysis shows that the only source of negative atomic pairs are the *contradiction* pairs, which actually correspond to 20% of RTE-5 data set.

Overall, the study showed that the decomposition methodology we propose can be applied on RTE-5 data. As for the quality of the atomic pairs, the high inter-annotator agreement rate obtained (reported in Section 4.2) shows that the methodology is stable enough to be applied on a large scale.

6 Related work

This section presents a number of studies that analyze RTE data sets from the point of view of linguistic phenomena.

An attempt to isolate the set of T-H pairs whose categorization can be accurately predicted based solely on syntactic cues has been carried out in Vanderwende et al. (2005). The aim of this work is to understand what proportion of the entailment pairs in the RTE-1 test set could be solved using a robust parser. Two human annotators evaluated each T-H pair of the test set, deciding whether the entailment was: *true by syntax*; *false by syntax*; *not syntax*; *can't decide*. Additionally, annotators were allowed to indicate whether the recourse to information in a general purpose thesaurus entry would allow a pair to be judged true or false. Their results show that 37% of the test items can be handled by syntax, broadly defined (including phenomena such as argument assignment, intra-sentential pronoun anaphora resolution); 49% of the test items can be handled by syntax plus a general purpose thesaurus. Even if we carried out our analysis on RTE-5 data, the results we reported in Table 3 are in line with those proposed in Vanderwende et al. (2005). According to their annotators, it is easier to decide when syntax can be expected to return *true*, and it is uncertain when to assign *false*. Basing on their own observations, their system

(Vanderwende et al., 2006) predicts entailment using syntactic features and a general purpose thesaurus, in addition to an overall alignment score. The syntactic heuristics used to recognize false entailment rely on the correct alignment of words and multiwords units between T and H logical forms.

Bar-Haim et al. (2005) define two intermediate models of TE, which correspond to lexical and lexical-syntactic levels of representation. Their lexical level captures knowledge about lexical-semantic and morphological relations, as well as lexical world knowledge. The lexical-syntactic level additionally captures syntactic relationships and transformations, lexical-syntactic inference patterns (rules) and co-reference. They manually annotated a sample from the RTE-1 data set according to each model, compared the outcomes for the two models as a whole as well as for their individual components, and explored how well they approximate the notion of entailment. It was shown that the lexical-syntactic model outperforms the lexical one, mainly because of a much lower rate of false-positives, but both models fail to achieve high recall. The analysis also showed that lexical-syntactic inference patterns stand out as a dominant contributor to the entailment task.

Clark et al. (2007) agree that only a few entailments can be recognized using simple syntactic matching, and that the majority rely on a significant amount of "common human understanding" of lexical and world knowledge. We also agree on the same conclusions (see Table 3). The authors present an analysis of 100 (25%) of the RTE-3 positive entailment pairs, to identify where and what kind of world knowledge are needed to fully identify and justify entailment. They discuss several existing resources and their capacity for supplying that knowledge. After showing the frequency of the different entailment phenomena from the sample they analyzed, they state that very few entailments depend purely on syntactic manipulation and a simple lexical knowledge (synonyms, hypernyms), and that the vast majority of entailments require significant world knowledge.

Dagan et al. (2008) present a framework for semantic inference at the lexical-syntactic level. The authors show that the inference module can be also exploited to improve unsupervised acquisition of entailment rules through canonization (i.e. the transformation of lexical-syntactic template variations that occur in a text into their canonical form - this form is chosen to be the active verb form with direct modifier). The canonization rule collection is composed by two kinds of rules: *i)* syntactic-based rules (e.g. passive/active forms, removal of conjunctions, removal of appositions), *ii)* nominalization rules, trying to capture the relations between verbs and their nominalizations. The authors propose to solve

the learning problems using this entailment module at learning time as well.

A definition of contradiction for TE task is provided by Marneffe et al. (2008), together with a collection of contradiction corpora. Detecting contradiction appears to be a harder task than detecting entailment, since it requires deeper inferences, assessing event coreference and model building. Contradiction is said to occur when two sentences are extremely unlikely to be true simultaneously; furthermore, they must involve the same event. The first empirical results for contradiction detection are presented in Harabagiu et al. (2006) (they focused only on contradictions involving negation and formed by paraphrases).

Kirk (2009) describes his work of building an inference corpus for spatial inference about motion, while Wang and Zhang (2008) focus on recognizing TE involving temporal expressions. Akhmatova and Dras (2009) experiment current approaches on hypernymy acquisition to improve entailment classification.

Basing on the intuition that frame-semantic information is a useful resource for modeling TE, Burchardt et al. (2009) provide a manual frame-semantic annotation for the test set used in RTE-2 (i.e. the FATE corpus) and discuss experiments conducted on this basis.

Bentivogli et al. (2009a) focus on some problematic issues related to resolving coreferences to entities, space, time and events at the corpus level, as emerged during the annotation of the data set for the RTE Search Pilot. Again at the discourse level, Mirkin et al. (2010b), and Mirkin et al. (2010a) analyze various discourse references in entailment inference (manual analysis on RTE-5 data set) and show that while the majority of them are nominal coreference relations, another substantial part is made up by verbal terms and bridging relations.

7 Conclusion

In this paper we have presented an investigation aiming at highlighting the relations between the logical dimension of textual semantic inferences, i.e. the capacity of the inference to prove the conclusion from its premises, and their linguistic dimension, i.e. the linguistic devices that are used to accomplish the goal of the inference. We think that the relation between the two dimensions has not received enough attention in the current stream of research on textual inferences in Computational Linguistics, and we believe that more empirical data and analysis are actually crucial to the progress of the many supervised systems that have been proposed in recent years in the area.

We have proposed a decomposition approach, where single linguistic

phenomena are isolated in what we have called *atomic inference pairs*. It is at this level of granularity that the actual correlation between the linguistic and the logical dimensions of semantic inferences emerges and can be empirically observed. For each of the two dimensions (i.e. logical and linguistic) we have proposed a number of features, mostly derived from previous literature, which help in the analysis. In order to support our thesis we have conducted an empirical analysis over a manually annotated data set of Textual Entailment pairs, derived from the recent RTE-5 evaluation campaign (the data we annotated are available online[25]). The results of the investigation show that the correlation between linguistic phenomena and logical judgments (i.e. entailment, contradiction, unknown) is quite poor, meaning that most of the linguistic phenomena we have observed and that occur in T-H pairs do not have an a priori polarity with respect to the logical relation holding in that pair. A relevant consequence of this fact is that the polarity of most of the phenomena is not predictable from the logical judgments, with an evident impact on the possibility to learn it from the available annotated RTE data sets. On the base of these findings we suggest that future developments should exploit the decomposition approach on specialized data sets, composed of atomic pairs.

In several respects the work we have presented in this paper is incomplete. It opens the way to further research in this direction. Particularly, we think that much more investigation and empirical experiments would be necessary in order to better determine the relations between linguistic phenomena and logical judgments in semantic inferences. Our hope is that these future data oriented studies will support computational approaches by e.g. driving search heuristics in transformation-based approaches, or optimizing feature selection in machine learning systems.

Acknowledgments

The work of the second author has been partially supported by the EXCITEMENT project (Exploring Customer Interactions through Textual Entailment), under the EU grant FP7 ICT-287923. The authors wish to thank Dr. Sara Tonelli for her help and availability in the annotation phase.

[25] http://www-sop.inria.fr/members/Elena.Cabrio/resources.html

References

Akhmatova, E. and M. Dras. 2009. Using hypernymy acquisition to tackle (part of) textual entailment. In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer 2009)*. Singapore.

Bar-Haim, R., I. Szpektor, and O. Glickman. 2005. Definition and analysis of intermediate entailment levels. In *Proceedings of the ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor, MI.

Bentivogli, L., E. Cabrio, I. Dagan, D. Giampiccolo, M. Lo Leggio, and B. Magnini. 2010. Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*. Valletta, Malta.

Bentivogli, L., I. Dagan, H.T. Dang, D. Giampiccolo, M. Lo Leggio, and B. Magnini. 2009a. Considering discourse references in textual entailment annotation. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*. Pisa, Italy.

Bentivogli, L., B. Magnini, I. Dagan, H.T. Dang, and D. Giampiccolo. 2009b. The fifth pascal recognizing textual entailment challenge. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*. Gaithersburg, Maryland.

Bos, J. and K. Markert. 2006. When logical inference helps determining textual entailment (and when it doesn't). In *Proceedings of the second PASCAL Challenge Workshop on Recognizing Textual Entailment*. Venice, Italy.

Burchardt, A., M. Pennacchiotti, S. Thater, and M. Pinkal. 2009. Measures of the amount of ecologic association between species. *Natural Language Engineering (JNLE)* 15(Special Issue 04).

Carletta, Jean. 1996. Assessing agreement on classification tasks: the kappa statistic. *Comput. Linguist.* 22(2):249–254.

Clark, P., P. Harrison, J. Thompson, W. Murray, J. Hobbs, and C. Fellbaum. 2007. On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-07 Workshop on Textual Entailment and Paraphrasing*. Prague, Czech Republic.

Cooper, R., D. Crouch, J. van Eijck, C. Fox, J. van Genabith, J. Jaspars, H. Kamp, D. Milward, M. Pinkal, M. Poesio, and S. Pulman. 1996. Using the framework. In *Technical Report LRE 62-051 D-16, The FraCaS Consortium*. Prague, Czech Republic.

Dagan, I., R. Bar-Haim, I. Szpektor, I. Greental, and E. Shnarch. 2008. Natural language as the basis for meaning representation and inference. In *Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing08)*. Haifa, Israel.

Dagan, I., B. Dolan, B. Magnini, and D. Roth. 2009. Recognizing textual entailment: Rational, evaluation and approaches. *Natural Language Engineering (JNLE)* 15(Special Issue 04):i–xvii.

Dagan, I., O. Glickman, and B. Magnini. 2005. The pascal recognizing textual entailment challenge. In *Proceedings of the First PASCAL Challenges Workshop on RTE*. Southampton, U.K.

Dagan, I., O. Glickman, and B. Magnini. 2006. The pascal recognizing textual entailment challenge. In *MLCW 2005, LNAI Volume 3944*. Springer-Verlag.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology* 26(3):297–302.

Fellbaum, C. 1998. Wordnet: An electronic lexical database. In *Language, Speech and Communication*. MIT Press.

Garoufi, K. 2007. Towards a better understanding of applied textual entailment. In *Master Thesis*. Saarland University. Saarbrücken, Germany.

Harabagiu, S., A. Hickl, and F. Lacatusu. 2006. Negation, contrast, and contradiction in text processing. In *Proceedings of the Twenty-First National Conference on Artificial Intellingence (AAAI-06)*. Boston, Massachusetts.

Hobbs, J. R. 2008. Abduction in natural language understanding. In L. R. Horn and G. Ward, eds., *The Handbook of Pragmatics*. Blackwell Publishing Ltd, Oxford.

Kirk, R. 2009. Building an annotated textual inference corpus for motion and space. In *Proceedings of the 2009 Workshop on Applied Textual Inference (TextInfer 2009)*. Singapore.

LoBue, P. and A. Yates. 2011. Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics*, pages 329–334. Portland, Oregon, USA.

Manning, C.D. 2006. Local textual inference: its hard to circumscribe, but you know it when you see it - and nlp needs it. In *Proceedings of the Eighth International Conference on Computational Semantics (IWCS-8)*. Unpublished manuscript.

Marneffe, M.C. De, A.N. Rafferty, and C.D. Manning. 2008. Finding contradictions in text. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL-08)*. Columbus, OH.

Mirkin, S., J. Berant, I. Dagan, and Eyal Shnarch. 2010a. Recognising entailment within discourse. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Mirkin, S., I. Dagan, and Sebastian Padò. 2010b. Assessing the role of discourse references in entailment inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*. Uppsala, Sweden.

Nenkova, A., R. Passonneau, and K. McKeown. 2007. The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Computational Logic* V, No. N, February:1–23.

Nolt, J., D. Rohatyn, and A. VArzi. 1998. *Schaum's outline of Theory and Problems of Logic 2nd ed.*. McGraw-Hill.

Peñas, Anselmo, Álvaro Rodrigo, Valentín Sama, and Felisa Verdejo. 2008. Testing the reasoning for question answering validation. *J. Log. and Comput.* 18(3):459–474.

Sammons, M., V.G.V Vydiswaran, and D. Roth. 2010. Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-10)*. Uppsala, Sweden.

Szpektor, I., E. Shnarch, and I Dagan I. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-07)*. Prague, Czech Republic.

Vanderwende, L., D. Coughlin, and B. Dolan. 2005. What syntax can contribute in entailment task. In *Proceedings of the First PASCAL Challenges Workshop on RTE*. Southampton, U.K.

Vanderwende, L., A. Menezes, and R. Snow. 2006. Microsoft research at rte-2: Syntactic contributions in the entailment task: an implementation. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*. Venice, Italy.

Wang, R. and Y. Zhang. 2008. Recognizing textual entailment with temporal expressions in natural language texts. In *Proceedings of the IEEE International Workshop on Semantic Computing and Applications (IWSCA-2008)*. Incheon, South Korea.

Wessa, P. 2012. Free statistics software. In *Office for Research Development and Education, version 1.1.23-r7*.

Zaenen, A., L. Karttunen, and R. Crouch. 2005. Local textual inference: can it be dened or circumscribed? In *Proceedings of the Workshop on the Empirical Modeling of Semantic Equivalence and Entailment*. Ann Arbor, MI.