Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5782–5791
Marseille, 11–16 May 2020

Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity

Venelin Kovatchev[1,3], Darina Gold[4], M. Antònia Martí[1,3], Maria Salamó[2,3], Torsten Zesch[4]
[1]Facultat de Filología, Universitat de Barcelona, [2]Facultat de Matemàtiques i Informàtica, Universitat de Barcelona,
[3]Universitat de Barcelona Institute of Complex Systems, [4]Language Technology Lab, University of Duisburg-Essen
[1,2,3] Gran Vía de les Corts Catalanes, 585, 08007 Barcelona, Spain, [4] Forsthausweg LE, 47057 Duisburg, Germany
{vkovatchev, amarti, maria.salamo}@ub.edu, {darina.gold,torsten.zesch}@uni-due.de

Abstract

In this paper, we present a methodology for decomposing and comparing multiple meaning relations (paraphrasing, textual entailment, contradiction, and specificity). The methodology includes SHARel - a new typology that consists of 26 linguistic and 8 reason-based categories. We use the typology to annotate a corpus of 520 sentence pairs in English and we demonstrate that unlike previous typologies, SHARel can be applied to all relations of interest with a high inter-annotator agreement. We analyze and compare the frequency and distribution of the linguistic and reason-based phenomena involved in paraphrasing, textual entailment, contradiction, and specificity. This comparison allows for a much more in-depth analysis of the workings of the individual relations and the way they interact and compare with each other. We release all resources (typology, annotation guidelines, and annotated corpus) to the community.

Keywords: Paraphrasing, Textual Entailment, Specificity

1. Introduction

This paper proposes a new approach for the decomposition of textual meaning relations. Instead of focusing on a single meaning relation we demonstrate that Paraphrasing, Textual Entailment, Contradiction, and Specificity can all be decomposed to a set of simpler and easier-to-define linguistic and reason-based phenomena. The set of "atomic" phenomena is shared across all relations.

In this paper, we adopt the definitions of meaning relations used by Gold et al. (2019). **Paraphrasing** is a symmetrical relation between two differently worded texts with approximately the same content (1a and 1b). **Textual Entailment** is a directional relation between two texts in which the information of the *Premise* (2a) entails the information of the *Hypothesis* (2b). **Contradiction** is a symmetrical relation between two texts that cannot be true at the same time (3a and 3b)[1]. **Specificity** is a directional relation between two texts in which one text is more precise (4a) and the other is more vague (4b).

1 **a)** *Education is equal for all children.*
b) *All children get the same education.*

2 **a)** *All children get the same education.*
b) *Education exists.*

3 **a)** *All children get the same education.*
b) *Some children get better education.*

4 **a)** *Girls do not get good education.*
b) *Some children do not get good education.*

The detection, extraction, and generation of pairs of texts with a particular meaning relation are popular and non-trivial tasks within Computational Linguistics (CL) and Natural Language Processing (NLP). Multiple datasets exist for each of these tasks (Dolan et al., 2004; Dagan et al., 2006; Agirre et al., 2012; Ganitkevitch et al., 2013; Bowman et al., 2015; Iyer et al., 2017; Lan et al., 2017; Kovatchev et al., 2018a). These tasks are also related to the more general problem of Natural Language Understanding (NLU) and are part of the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018).

Recently, several researchers have argued that a single label such as "paraphrasing", "textual entailment", or "similarity" is not enough to characterize and understand the meaning relation (Sammons et al., 2010; Bhagat and Hovy, 2013; Vila et al., 2014; Cabrio and Magnini, 2013; Lopez-Gazpio et al., 2016; Benikova and Zesch, 2017; Kovatchev et al., 2018a). These authors demonstrate that the different instances of meaning relations require different capabilities and linguistic knowledge. For example, the pairs 5 and 6 are both examples of a "paraphrasing" relation. However determining the relation in 5a–5b only requires lexical knowledge, while syntactic knowledge is also needed for correctly predicting the relation in 6a–6b. This distinction cannot be captured by a single "paraphrasing" label. The lack of distinction between such examples can be a problem in error analysis and in downstream applications.

5 **a)** *Education is equal for all <u>children</u>.*
b) *Education is equal for all <u>kids</u>.*

6 **a)** *<u>All children receive</u> the same education.*
b) *The same education <u>is provided to all children</u>.*

A richer set of labels is needed to better characterize the complexity of meaning relations. We believe that a typology of "paraphrasing", "textual entailment", and "semantic similarity" would capture the distinctions between the different instances of each relation. Kovatchev et al. (2019) empirically demonstrate that in the case of Paraphrase Identification (PI), the different "paraphrase types" are processed in a different way by automated PI systems.

[1] In the Recognizing Textual Entailment (RTE) literature, contradiction is often understood as the lack of entailment. However we adopt a more strict definition of the phenomenon.

In this paper, we demonstrate that multiple meaning relations can be decomposed using a shared typology. This is the first step towards building a single framework for analyzing, comparing, and evaluating multiple meaning relations. Such a framework has not only theoretical importance, but also clear practical implications. Representing every meaning relation with the same set of linguistic and reason-based phenomena allows for a better understanding of the nature of the relations and facilitates the transfer of knowledge (resources, features, and systems) between them.

For the purpose of decomposing the meaning relations we propose **S**ingle **H**uman-Interpretable Typology for **A**nnotating Meaning **Rel**ations (SHARel). With the goal of showing the applicability of the new typology, we also perform an annotation experiment using the SHARel typology. We annotate a corpus of 520 text pairs in English, containing paraphrasing, textual entailment, contradiction, and textual specificity. The quality of the typology and of the annotation is evident from the high inter-annotator agreement.

Finally, we present a novel, quantitative comparison between the different meaning relations in terms of the types involved in each of them.

The rest of this article is organized as follows. Section 2. lists the Related Work. Section 3. presents the typology, the objectives behind it and the process of selection of the types. Section 4. describes the annotation process - the corpus, the annotation guidelines, and the annotation interface. Section 5. shows the results of the annotation. Section 6. discusses the implications of the findings and the way our results relate to our objectives and research questions. Finally, Section 7. concludes the paper and addresses the future work.

2. Related Work

The last several years have seen an increasing interest towards the decomposition of paraphrasing (Bhagat and Hovy, 2013; Vila et al., 2014; Benikova and Zesch, 2017; Kovatchev et al., 2018a), textual entailment (Sammons et al., 2010; LoBue and Yates, 2011; Cabrio and Magnini, 2013), and textual similarity (Lopez-Gazpio et al., 2016).

Sammons et al. (2010) argue that in order to process a complex meaning relation such as textual entailment a competent speaker has to take several "inference steps". This means that a meta-relation such as paraphrasing, textual entailment, or semantic similarity can be "decomposed" or broken down into such "inference steps". These "inference steps", traditionally called "types" can be either linguistic or reason-based in their nature. The linguistic types require certain linguistic capabilities from the speaker, while the reason-based types require common-sense reasoning and world knowledge.

The different authors working on decomposing meaning relations all follow a similar approach. First, they propose a typology - a set of "atomic" linguistic and/or reasoning types involved in the inference process of the particular meta-relation (paraphrasing, entailment, or similarity), Then, they use the "atomic" types in a corpus annotation and finally, they analyze the distribution and correlation of the types. The corpus based studies have demonstrated that different atomic types can be found in various corpora for paraphrasing, textual entailment, and semantic similarity research.

Kovatchev et al. (2019) empirically demonstrated that the performance of a Paraphrase Identification (PI) system on each candidate-paraphrase pair depends on the "atomic types" involved in that pair. That is, they showed that state-of-the-art automatic PI systems process "atomic paraphrases" in a different manner and with a statistically significant difference in quantitative performance (Accuracy and F1). They show that more frequent and relatively simple types like "lexical substitution", "punctuation changes" and "modal verb changes" are easier across multiple automated PI systems, while other types like "negation switching", "ellipsis" and "named entity reasoning" are much more challenging.

Similar observations have been made in the field of Textual Entailment. Gururangan et al. (2018) discovered the presence of annotation artifacts that enable models that take into account only one of the texts (the hypothesis) to achieve performance substantially higher than the majority baselines in SNLI and MNLI. Glockner et al. (2018) showed that models trained with SNLI fail to resolve new pairs that require simple lexical substitution. Naik et al. (2018) create label-preserving adversarial examples and conclude that automated NLI models are not robust. Wallace et al. (2019) introduce universal triggers, that is, sequences of tokens that fool models when concatenated to any input. All these authors identify different problems and biases in the datasets and the systems trained on them. However they focus on a single phenomenon and/or a specific linguistic construction. A typology-based approach can evaluate the performance and robustness of automated systems on a large variety of tasks.

One limitation of the different decompositional approaches is that there exist many different typologies and each typology is created considering only one meaning relation (paraphrasing, textual entailment, textual similarity). This follows the traditional approach in the research on meaning relations: each relation is studied in isolation, with its own theoretical concepts, datasets, and practical tasks.

In recent years, the "single relation" approach has been questioned by several authors. Androutsopoulos and Malakasiotis (2010) analyze the relations between paraphrasing and textual entailment. Marelli et al. (2014) present SICK: a corpus that studies entailment, contradiction, and semantic similarity. Lan and Xu (2018) and Aldarmaki and Diab (2018) explore the transfer learning capabilities between paraphrasing and textual entailment. Gold et al. (2019) present a corpus that is annotated for paraphrasing, textual entailment, contradiction, specificity, and textual similarity. These works demonstrate that the different meaning relations can be studied together and can benefit from one another.

However, to date, the joint research of meaning relations is limited only to the binary textual labels. There has been no work on comparing the different typologies and the way different relations can be decomposed. None of the existing typologies is fully compatible with multiple meaning relations, which further restricts the research in this area. We aim to address this research gap in this paper.

3. Shared Typology for Meaning Relations

This section is organized as follows. Section 3.1. presents the problem of decomposing meaning relations. Section 3.2. describes our proposed typology and the rationale behind it. Section 3.3. formulates our research questions.

3.1. Decomposing Meaning Relations

The goal behind the **S**ingle **H**uman-Interpretable Typology for **A**nnotating Meaning **Rel**ations (SHARel) is to come up with a unified list of linguistic and reason-based phenomena that are required in order to determine the meaning relations that hold between two texts. The list of types should not be limited to texts that hold a specific single textual relation, such as paraphrasing, textual entailment, contradiction, and textual specificity. Rather, the types should be applicable to texts holding multiple different relations.

7 **a** All children *receive* the same education.
b The same education *is received* by all kids.

8 **a** All children *receive* the same education.
b The same education *is* **not** *received* by all kids.

In 7a and 7b, the meaning relation at a textual level is paraphrasing, while in 8a and 8b, the textual relation is contradiction. In order to determine the meaning relation for both 7 and 8, a competent speaker or an automated system needs to make several inference steps. First, they have to determine that "kids" and "children" have the same meaning and the same syntactic and semantic role in the texts. Second, they need to account for the change in grammatical voice. In terms of typology, these inference steps involve two different types - "same polarity substitution" ("kids" - "children") and "diathesis alternation" ("receive" - "is received"). In addition, in example 8b, the human or the automated system needs to determine the presence and the function of "negation" (**not**).

By successfully performing all necessary inference steps, the human (or the automated system) is able to determine that in the pair 7a-7b there is equivalence of the expressed meaning, while in the pair 8a-8b there is a logical contradiction. The required inference steps in the two examples are not specific to the textual label (paraphrasing or contradiction). The "types" are general linguistic or reason-based phenomena.

With the goal of addressing such situations, we propose a list of types that, following the existing theoretical research, can be applied to multiple meaning relations. We justify the choice of types for SHARel in the context of existing typologies.

3.2. The SHARel Typology

Table 1 shows the SHARel Typology and its 34 different types, organized in 8 categories. The first 6 categories (morphology, lexicon, lexico-syntactic, syntax, discourse, other) consist of the 24 "linguistic" types. The two types in the "extremes" category ("identity" and "unrelated") are neither linguistic, nor reason-based. The last category consists of the 8 "reason-based" types.

ID	Type
Morphology-based changes	
1	Inflectional changes
2	Modal verb changes
3	Derivational changes
Lexicon-based changes	
4	Spelling changes
5	Same polarity substitution (habitual)
6	Same polarity substitution (contextual)
7	Same polarity sub. (named entity)
8	Change of format
Lexico-syntactic based changes	
9	Opposite polarity sub. (habitual)
10	Opposite polarity sub. (contextual)
11	Synthetic/analytic substitution
12	Converse substitution
Syntax-based changes	
13	Diathesis alternation
14	Negation switching
15	Ellipsis
16	Anaphora
17	Coordination changes
18	Subordination and nesting changes
Discourse-based changes	
18	Punctuation changes
20	Direct/indirect style alternations
21	Sentence modality changes
22	Syntax/discourse structure changes
Other changes	
23	Addition/Deletion
24	Change of order
Extremes	
25	Identity
26	Unrelated
Reason-based changes	
27	Cause and Effect
28	Conditions and Properties
29	Functionality and Mutual Exclusivity
30	Named Entity Reasoning
31	Numerical Reasoning
32	Temporal and Spatial Reasoning
33	Transitivity
34	Other (General Inference)

Table 1: The SHARel Typology

The distinction between linguistic and reason-based types is introduced by Sammons et al. (2010) and Cabrio and Magnini (2013) for textual entailment. The linguistic phenomena require certain linguistic capabilities from the human speaker or the automated system. The reason-based phenomena require world knowledge and common-sense

Typology	Relation	Types	Linguistic	Reasoning	Hierarchy
Sammons et al. (2010)	TE, CNT	22	13	9	No
LoBue and Yates (2011)	TE, CNT	20	0	20	No
Cabrio and Magnini (2013)	TE, CNT	36	24	12	Yes
Bhagat and Hovy (2013)	PP	25	22	3	No
Vila et al. (2014)	PP	23	19	1	Yes
Kovatchev et al. (2018a)	PP	27	23	1	Yes
SHARel	TE, CNT PP, SP, TS	34	24	8	Yes

Table 2: Comparing typologies of textual meaning relations

reasoning.

For the linguistic types, we compared the existing typologies and decided to use the Extended Paraphrase Typology (EPT) (Kovatchev et al., 2018a) as a starting point. The authors of EPT have already combined various linguistic types from the fields of Paraphrasing and Textual Entailment and have taken into account the work of Sammons et al. (2010), Vila et al. (2014), Cabrio and Magnini (2013). As such, the majority of the linguistic types that they propose are in principle applicable to both Paraphrasing and Textual Entailment.

We examined the types from EPT and made several adjustments in order to make the linguistic types fully independent of the textual relation.

- EPT contains "entailment" and "non-paraphrase" types in the category "extremes". These types were created specifically for the task of Paraphrase Identification (PI). We removed these types from the list.
- We added "unrelated" type (#26) to the category "extremes" to capture information which is not related at all to the other sentence in the pair.
- We added "anaphora" type (#16) in the syntax category. This change was suggested by our annotators during the process of corpus annotation.

For the reason-based types we studied the typologies of Sammons et al. (2010), LoBue and Yates (2011) and Cabrio and Magnini (2013). While these typologies have a lot of similarities and shared types, they are not fully compatible. We analyzed the type of common-sense reasoning and background knowledge that is required for each of the types in these three typologies. We combined similar types into more general types and reduced the original list of over 30 reason-based types to 8. For example, the "named entity reasoning" (#30) includes both reasoning about geographical entities and publicly known persons (those two were originally separated types). [2]

With respect to specificity, we propose a fine-grained token level annotation, which allows us to determine the particular elements in one sentence that are more (or less) specific than their counterpart in the other sentence. Ko et al. (2019) demonstrated that specificity needs to be more linguistically and informational theoretically based to be more semantically plausible. This could partially be solved through a more fine-grained annotation of specificity, as it is performed in this study.

Table 2 lists some properties of the existing meaning relations. All typologies before SHARel were created only for one (or two) meaning relations. SHARel contains general types that are not specific to any particular meaning relation and can be applied to pairs holding Textual Entailment, Contradiction, Paraphrasing, Textual Specificity, or Semantic Textual Similarity meaning relation. SHARel follows the good practices of typology research and organizes the types in a hierarchical structure of 8 categories and has a good balance between linguistic and reasoning types.

3.3. Research Questions

There are two main objectives that motivated this paper:
1) To demonstrate that multiple meaning relations can be decomposed using a single, shared typology;
2) To demonstrate some of the advantages of a shared typology of meaning relations.

Based on our objectives, we pose two research questions (RQs) that we want to address in this article.

> **RQ1:** Is it possible to use a single typology for the decomposition of multiple (textual) meaning relations?
>
> **RQ2:** What are the similarities and the differences between the (textual) meaning relations in terms of types?

We address these research questions in a corpus annotation study. For the first research question we evaluate the quality of the corpus annotation by measuring the inter-annotator agreement. For the second research question we measure the relative frequencies of the types in sentence pairs with each textual meaning relation.

4. Corpus Annotation

This section is organized as follows: Section 4.1. describes the corpus that we chose to use in the annotation. Section 4.2. presents the annotation setup. Finally, in Section 4.3. we report the annotation agreement.

4.1. Choice of Corpus

In order to determine the applicability of SHARel to all relations of interest, we carried out a corpus annotation. We used the publicly available corpus of Gold et al. (2019).

[2] The annotation guidelines and examples for all types can be seen at `https://github.com/venelink/sharel`

It consists of 520 text pairs and is already annotated at sentence level for paraphrasing, entailment, contradiction, specificity and semantic similarity. Gold et al. (2019) performed the annotation for each relation independently. That is, for each pair of sentences 10 annotators were asked whether a particular relation (paraphrasing, entailment, contradiction, specificity) held or not.

The corpus of Gold et al. (2019) contains 160 pairs annotated as paraphrases, 195 pairs annotated as textual entailment (in one direction or in both) and 68 pairs annotated as contradiction. As the annotation for the different relations was carried out independently, there is an overlap between the relations. For example 52% of the pairs annotated as entailment were also annotated as paraphrases. The total number of pairs annotated with at least one relation among paraphrasing, entailment, and contradiction is 256. The remaining 244 pairs were annotated as unrelated. In 381 of the pairs, one of the sentences was marked as more specific than the other.

The corpus of Gold et al. (2019) is the only corpus to date which contains all relations of interest. All text pairs are in the same domain and topic, they have similar syntactic structure and vocabulary. The lexical overlap between the two sentences in each pair is much lower than in corpora such as MRPC (Dolan et al., 2004) or SNLI (Bowman et al., 2015). This means that even though the two sentences in a pair are in a meaning relation such as paraphrasing or textual entailment, there are very few words that are directly repeated. All these properties of the corpus were taken into consideration when we chose it for our annotation.

4.2. Annotation Setup

We performed an annotation with the SHARel typology on all pairs from Gold et al. (2019) that have at least one of the following relations: paraphrasing, forward entailment, backwards entailment, and contradiction. We discarded pairs that are annotated as ”unrelated”. This is a typical approach when decomposing meaning relations. Sammons et al. (2010; Cabrio and Magnini (2013; Vila et al. (2014) only decompose pairs with a particular relation (entailment, contradiction, or paraphrasing).

After discarding the unrelated portion, the total number of pairs that we annotated with SHARel was 276. Prior to the annotation we tokenized each sentence using the NLTK python library.

During the annotation process, our annotators go through each pair in the corpus. For each linguistic and reason-based phenomenon that they encounter, they annotate the type and the scope (the specific tokens affected by the type). We used an open source web-based annotation interface, called WARP-Text (Kovatchev et al., 2018b).

We prepared extended guidelines with examples for each type. Each pair of texts was annotated independently by two trained expert annotators. In the cases where there were disagreements, the annotators discussed their differences in order to obtain the best possible annotation for the example pair [3].

4.3. Agreement

For calculating inter-annotator agreement, we use the two different versions of the IAPTA-TPO measures. The IAPTA-TPO measures was proposed by Vila et al. (2015) specifically for the task of annotating paraphrase types. They were later on refined by Kovatchev et al. (2018a). IAPTA-TPO measure the agreement on both the label (the annotated phenomenon) and the scope, which is non-trivial to capture using traditional measures such as Kappa. IAPTA-TPO (Total) measures the cases where the annotators fully agree on both label and scope. IAPTA-TPO (Partial) measures the cases where the annotators agree on the label, but the scope overlaps only partially.

The agreement of our annotation can be seen in Table 3. We calculate the agreement on all pairs (all), and we also report the agreement for the pairs with textual label paraphrases (pp), entailment (ent), and contradiction (cnt).

	TPO-Partial	**TPO-Total**
This corpus (all)	.78	.52
This corpus (pp)	.77	.51
This corpus (ent)	.77	.52
This corpus (cnt)	.75	.50
MRPC-A	.78	.51
ETPC (non-pp)	.72	.68
ETPC (pp)	.86	.68

Table 3: Inter-annotator Agreement

To put our results in perspective, we compare our agreement with the one reported in MRPC-A (Vila et al., 2015) and ETPC (Kovatchev et al., 2018a). For ETPC the authors report both the agreement on the pairs annotated as paraphrases (pp) and as non-paraphrases (non-pp). To date, MRPC-A and ETPC are the only two corpora of sufficient size annotated with a typology of meaning relations. They also use the same inter-annotation measure to report agreement, so we can compare with them directly.

The overall agreement that we obtain (.52 Total and .78 Partial) is almost identical to the agreement reported for MRPC-A (.51 Total and .78 Partial) and slightly lower than the agreement reported for ETPC (.68 Total and .86 Partial). Kovatchev et al. (2018a) detected a significant difference in the agreement between paraphrase and non-paraphrase pairs. In their annotation, the “non-paraphrase” includes mostly entailment and contradiction pairs and the lower agreement indicates that their typology is not well equipped for dealing with those cases. However in our corpus, we don’t observe such a difference. Our annotation agreement is very consistent across all pairs indicating that SHARel is successfully applied to all relations of interest.

The consistently high agreement score indicates the high quality of the annotation. Even though our task and our typology are much more complex than those of Vila et al. (2014) and Kovatchev et al. (2018a), we still obtain comparable results.

In addition to calculating the inter-annotation agreement, we also asked the annotators to mark and indicate any examples and/or phenomena not covered by the typology.

[3] The annotation guidelines and the annotated corpus are available at `https://github.com/venelink/sharel`

Based on their ongoing feedback during the annotation, we decided to introduce the “anaphora” type. We re-annotated the portion of the corpus that was already annotated at the time when we introduced the new type.

Arriving at this point, we have demonstrated that it is possible to successfully use a single typology for the decomposition of multiple (textual) meaning relations. This answers our first research question (RQ1).

5. Analysis of the Results

Before this paper, the comparison between textual meaning relations was limited to measuring the overlap and correlation between the binary label of the pairs. Gold et al. (2019) present such an analysis. They find some expected results such as the high correlation and overlap between paraphrasing and (uni-directional) entailment and the negative correlation between paraphrasing and contradiction or entailment and contradiction. They also report some interesting and unexpected results. They point that in practical setting paraphrasing does not equal bi-directional entailment. With respect to specificity they find that it does not correlate with other textual meaning relations, and does not overlap with textual entailment.

In this section, we go further than the binary labels of the textual meaning relations and compare the distribution of types across all relations. A typological comparison can be much more informative about the interactions between the different relations.

This section is organized as follows. Section 5.1. analyzes and compares the frequency distribution of the different types in pairs with the following textual relations: Paraphrasing, Textual Entailment, and Contradiction. Section 5.2. discusses the Specificity relation and the types involved in it.

5.1. Type Frequency

To determine the similarities and differences between the textual meaning relations in terms of types, we measured the relative type frequencies for pairs that have the corresponding label. Table 4 shows the relative frequencies in pairs that have paraphrasing, entailment, or contradiction relations at textual level. For the entailment relation we consider only the pairs marked as “uni-directional entailment”. That is, pairs that have entailment only in one of the directions. We discard the pairs that have bi-directional entailment to reduce the overlap with paraphrases (94 % of the bi-directional entailment pairs are also paraphrases).

For reference, we have also included the type frequencies for the paraphrase portion of the ETPC (Kovatchev et al., 2018a) corpus. ETPC is the largest corpus to date annotated with paraphrase types. The EPT typology used to annotate the ETPC also shares the majority of the linguistic types with SHARel. This allows us to put our results in perspective and to determine to what extent are they consistent with previous findings.

We can observe that the distribution of types is not balanced for any of the portions. Some types are over-represented, while others are under-represented or not represented at all. We focus our analysis on four different tendencies: 1) linguistic types that are frequent across all relations; 2) types whose frequency changes across the different relations; 3) the frequency of reason-based types; and 4) types that are infrequent or not represented at all.

Frequent linguistic types across all relations

The most frequent types across all relations are *same polarity substitution (habitual)* (#5), *same polarity substitution (contextual)* (#6), *same polarity substitution (named entity)* (#7), *addition/deletion* (#23), and *identity* (#25). These phenomena account for more than 50% of the types in the corpus. This finding is also consistent with the results reported in the ETPC. It is worth noting that in the ETPC, the distribution within the different *same polarity substitution* types (#5, #6, #7) differs from our annotation. The frequency of *same polarity substitution (habitual)* (#5) is lower, while *same polarity substitution (contextual)* (#6) and *same polarity substitution (named entity)* (#7) have a much higher frequency.

Other frequent types shared across all relations are *inflectional* (#1), *opposite polarity substitution (habitual)* (#9), *synthetic/analytic substitution* (#11), *converse substitution* (#12), *diathesis alternation* (#13), and *negation switching* (#14). For all of these types, the frequency that we obtain is substantially higher than in the ETPC corpus.

Differences in type frequencies across relations

We can observe that paraphrasing has the highest frequency of *Same Polarity Substitution*, both habitual (#5) and contextual (#6). This is a tendency that can also be observed in ETPC.

Entailment is the relation with the highest relative frequency of phenomena in the reason-based category. The reason-based phenomena (#27-#34) account for 13.1% of all phenomena within entailment, doubling the frequency of these phenomena in paraphrasing (5.65%) and contradiction (6.2%). Most of that difference comes from the ”conditions/properties” (#28) type. The entailment relation also has the lowest frequency of same polarity substitutions (#5, #6, and #7).

Contradiction has the highest frequency of opposite polarity substitution (#9 and #10) and negation switching (#14), doubling the frequency of these phenomena in paraphrasing and entailment pairs. Interestingly, contradictions have a comparable frequency of same polarity substitution (#5, #6, and #7) and identity (#25) to paraphrases. This suggests that contradictions are more similar to paraphrases than to entailment, at least in terms of the phenomena involved.

Frequency of reason-based types

We can observe that reason-based types (#27-#34) are much less frequent than linguistic types. Reasoning accounts for less than 14% of the examples across all relations. That means that in the majority of the cases, the textual relation can be determined via linguistic means and does not require reasoning or world knowledge. The most frequent reasoning type across all relations is *cause/effect*.

It is important to note that the frequency of reasoning phenomena in our annotation is much higher than the 1.5% reported in ETPC. In ETPC all reason based phenomena were annotated with a single label - *Other (General Inferences)* (#34) so the frequency of this type corresponds to the sum

ID	Type	Paraphrasing	Entailment	Contradiction	ETPC
Morphology-based changes					
1	Inflectional changes	4 %	4 %	1.9 %	2.78 %
2	Modal verb changes	0.25 %	1 %	0	0.83 %
3	Derivational changes	2 %	0	0.6 %	0.85 %
Lexicon-based changes					
4	Spelling changes	0.25 %	0.4 %	0	2.91 %
5	Same polarity substitution (habitual)	25.2 %	17 %	26 %	8.68 %
6	Same polarity substitution (contextual)	9.7 %	6.3 %	5.5 %	11.66 %
7	Same polarity sub. (named entity)	0.7 %	0.4 %	1.2 %	5.08 %
8	Change of format	0.7 %	0.9 %	0	1.1 %
Lexico-syntactic based changes					
9	Opposite polarity sub. (habitual)	2.7 %	3.5 %	7.5%	0.07 %
10	Opposite polarity sub. (contextual)	0.5 %	0.9 %	1.2 %	0.02 %
11	Synthetic/analytic substitution	6.7 %	6.8 %	3.7 %	3.80 %
12	Converse substitution	2.5 %	3.2 %	3.1 %	0.20 %
Syntax-based changes					
13	Diathesis alternation	1.5 %	2.2%	1.9 %	0.73 %
14	Negation switching	4 %	4 %	11.2 %	0.09 %
15	Ellipsis	0	0	0	0.30 %
16	Anaphora	1.7 %	2.7 %	0.6 %	0
17	Coordination changes	0	0	0	0.22 %
18	Subordination and nesting changes	0.25 %	0	0	2.14 %
Discourse-based changes					
18	Punctuation changes	0	0	0	3.77 %
20	Direct/indirect style alternations	0	0	0	0.30 %
21	Sentence modality changes	0	0	0	0
22	Syntax/discourse structure changes	0	0	0	1.39 %
Other changes					
23	Addition/Deletion	16.25 %	16.4 %	16.2 %	25.94 %
24	Change of order	0.5 %	0.9 %	0.6 %	3.89 %
Extremes					
25	Identity	12.5 %	14.5 %	11.8 %	17.5 %
26	Unrelated	0	0	0	3.81 %
Reasoning					
27	Cause and Effect	4.7 %	5.4 %	5 %	n/a
28	Conditions and Properties	2 %	6 %	0.6 %	n/a
29	Functionality and Mutual Exclusivity	0	0.4 %	0	n/a
30	Named Entity Reasoning	0	0	0	n/a
31	Numerical Reasoning	0	0	0	n/a
32	Temporal and Spatial Reasoning	0	0	0	n/a
33	Transitivity	0.25 %	0.9 %	0	n/a
34	Other (General Inference)	0.5 %	0.4 %	0.6 %	1.53 %

Table 4: Type Frequencies

of all types from #27 to #34 in our annotation. These findings indicate that the methodology of Gold et al. (2019) successfully addresses one of the problems in the ETPC corpus, already emphasized by other researchers - the lack of reason-based types.

Low frequency types and missing types

In our annotation, there are several linguistic and reason-based types that are not represented at all. Regarding the linguistic types, there are no discourse based types, no *ellipsis* (#15), no *coordination changes* (#17), and almost no *subordination and nesting changes* (#18). Regarding the reason-based types, there are no *Named Entity Reasoning* (#30), *Numerical Reasoning* (#31), and no *Temporal and Spatial Reasoning* (#32).

We argue that the absence of these types in our annotation is due to the way in which the Gold et al. (2019) corpus was created. The authors of that corpus aimed at obtaining simple, one-verb sentences. The average length of a sentence is 10.5 tokens, which is much lower than the length of sentences in other corpora (ex.: 22 average length for ETPC). The corpus contains almost no Named Entities (proper names, locations, or quantities). These characteristics of the corpus do not facilitate transformations at the syntactic and discourse levels or Named Entity Reasoning. Our intuition that the lack of these types is due to the corpus creation is further reinforced by the fact that these types are missing across all meaning relations. However, these missing types can be observed in other paraphrasing and entailment corpora, such as Sammons et al. (2010), Cabrio and Magnini (2013), and Kovatchev et al. (2018a). For these reasons we decided to keep them as part of the ShaRel typology. It would, nevertheless, require a further research and richer corpora to empirically determine the importance of these phenomena for the different meaning relations.

Summary The similarities and common tendencies between paraphrases, entailment, and contradiction clearly indicate that these relations belong within the same conceptual framework and should be studied and compared together. The results also suggest the possibility of the transfer of knowledge and technologies between these relations. The differences between the textual meaning relations in terms of the involved types can help us to understand each of the individual relations better. This information can also be useful in the automatic classification of the different relations in a practical task.

5.2. Decomposing Specificity

We define specificity as the opposite of generality or fuzziness. Yager (1992) defines specificity as the degree to which a fuzzy subset points to one element as its member. This meaning relation has not been studied extensively. It has also not been decomposed. To the best of our knowledge this is the first work to do so. Gold et al. (2019) show that there is no direct correlation between specificity and the other textual meaning relations, including textual entailment. For that reason, we took a different approach to the decomposition of specificity and treat it separately from the other relations. We added one extra step in the annotation process, focused on the specificity relation.

The corpus of Gold et al. (2019) is annotated for specificity at the textual level. That is, the crowd workers identified which of the two given sentences is more specific. In 9, the annotators would indicate that **b** is more specific than **a**.

9 **a** All children receive the same education.
b The same education is received by all girls.

In our annotation, we performed an additional annotation of the specificity and we identified the particular elements (words, phrases, clauses) in one sentence that were more specific than their counterpart. In example 9, we can identify that "girls" is more specific than "children". The difference in the specificity of "girls" and "children" is the reason why **b** is annotated as more specific than **a**. We called that "scope of specificity".

In 80% of the pairs with specificity at textual level, our annotators were able to point at one or more particular elements that are responsible for the difference in specificity. In 20% of the pairs, the specificity was not decomposable. This finding also confirms (Ko et al., 2019)'s findings, who showed that frequency-based features are well-suited for automatic specificity detection.

In our analysis on the nature of the specificity relation, we combined the annotation of "scope of specificity" and the traditional annotation of linguistic and reason-based types discussed in the previous sections. In particular, we looked for overlap between the "scope of specificity" and the scope of linguistic and reason-based types. Example 10 shows the two separate annotations side by side. In **a** and **b**, we show the annotation of the linguistic and reason-based types: *"same polarity substitution (habitual)"* of "children" and "girls", and *"diathesis alternation"* of "receive" and "is received by". In **c** and **d** we show the annotation of the specificity: "children" - "girls". When we compare the two annotations we can observe that the "scope of specificity" overlaps with the scope of *"same polarity substitution (habitual)"*.

10 **a** All children *receive* the same education.
b The same education *is received* by all girls.
c All **children** receive the same education.
d The same education is received by all **girls**.

We argue that when there is an overlap between the "scope of specificity" and a linguistic or a reason-based type, it is the linguistic or reason-based phenomenon that is responsible for the difference in specificity. In example 10 we can say that the substitution of "children" and "girls" is responsible for the difference of specificity.

ID	Type	Freq.
3	Derivational Changes	1 %
5	Same Pol. Sub. (habitual)	17 %
6	Same Pol. Sub. (contextual)	9 %
7	Same Pol. Sub. (named entity)	2 %
9	Opp. Pol. Sub (habitual)	2 %
11	Synthetic / Analytic sub.	9 %
14	Negation Switching	1 %
16	Anaphora	1%
23	Addition / Deletion	50 %
27	Cause and Effect	7 %
28	Condition / Property	1 %
33	Transitivity	1 %
34	Other (General Inferences)	1 %

Table 5: Decomposition of Specificity

Table 5 shows the overlap between "scope of specificity" and "atomic types". In 97 % of the cases where specificity was decomposable the more/less specific elements overlapped with an atomic type. In 50 % of the cases the specificity was due to additional information (#23). The other frequent cases include *same polarity substitution* (#5, #6, and #7), *synthetic/analytic substitution* (#11), and *cause and effect* (#27) reasoning. While the overall tendencies are similar to the other meaning relations, specificity also has its unique characteristics. We found almost no specificity at morphological level and the frequency of *Same polarity substitution* (#5, #6, and #7), while still high, was lower than that of paraphrasing and contradiction pairs. The relative frequency of *Synthetic/analytic substitution* (#11) was the highest of all relations and the reasoning types were almost as frequent as in entailment pairs, although the type distribution is different. We found no syntactic or discourse driven specificity changes.

6. Discussion

In Section 3., we posed two Research Questions that we wanted to address within this paper. We answered both of them in sections 4. and 5.. Our annotation demonstrated that a shared typology can be successfully applied to multiple relations. The quality of the annotation is attested by the high inter-annotator agreement. We also demonstrated that a shared typology, such as SHARel, is useful to compare different meaning relations in a quantitative and human interpretable way.

In this paper we provide a new perspective on the joint research into multiple meaning relations. Traditionally, the meaning relations have been studied in isolation. Only recently researchers have started to explore the possibility of a joint research and a transfer of knowledge. We propose a new framework for a joint research on meaning relations via a shared typology. This framework has clear advantages: it is intuitive to use and interpret; it is easy to adapt in practical setting - both in corpora creation and in empirical tasks; it is based on solid linguistic theory. We believe that our approach can lead to a better understanding of the workings of the meaning relations, but also to improvements in the performance of automated systems.

The biggest challenge in the joint study of meaning relations is the limited availability of corpora annotated with multiple relations. The corpus that we used for our study is relatively small in size. It also has restrictions in terms of sentence length and the frequency of Named Entities. However, it is the only corpus to date annotated with all relations of interest.

Despite the limitations of the chosen corpus, the obtained results are promising. We provide interesting insights into the workings of the different relations, and also outline various practical implications. Kovatchev et al. (2019) demonstrated that a corpus with a size of a few thousand sentence pairs can be successfully used as a qualitative evaluation benchmark. SHARel and the annotation methodology we used easily scale to such size of corpora. This opens up the possibility for a qualitative evaluation of multiple meaning relations as well as for easier transfer of knowledge based on the particular types involved in the relations.

7. Conclusions and Future Work

In this paper we presented the first attempt towards decomposing multiple meaning relations using a shared typology. For this purpose we used SHARel - a typology that is not restricted to a single meaning relation. We applied the SHaRel typology in an annotation study and demonstrated its applicability. We analyzed the shared tendencies and the key differences between paraphrasing, textual entailment, contradiction, and specificity at the level of linguistic and reason-based types.

Our work is the first successful step towards building a framework for studying and processing multiple meaning relations. We demonstrate that the linguistic and reasoning phenomena underlying the meaning relations are very similar and can be captured by a shared typology. A single framework for meaning relations can facilitate the analysis and comparison of the different relations and improve the transfer of knowledge between them.

As future work, we aim to use the findings and resources of this study in practical applications such as the development and evaluation of systems for automatic detection of paraphrases, entailment, contradiction, and specificity. We plan to use the SHARel typology for a general-purpose qualitative evaluation framework for meaning relations.

Bibliographical References

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 385–393, Stroudsburg, PA, USA. Association for Computational Linguistics.

Aldarmaki, H. and Diab, M. (2018). Evaluation of unsupervised compositional representations. In *Proceedings of COLING 2018*.

Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Benikova, D. and Zesch, T. (2017). Same same, but different: Compositionality of paraphrase granularity levels. In *Proceedings of RANLP 2017*.

Bhagat, R. and Hovy, E. H. (2013). What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Cabrio, E. and Magnini, B. (2013). Decomposing Semantic Inferences. *Linguistics Issues in Language Technology - LiLT. Special Issues on the Semantics of Entailment*, 9(1), August.

Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of*

Coling 2004, pages 350–356, Geneva, Switzerland, Aug 23–Aug 27. COLING.

Ganitkevitch, J., Durme, B. V., and Callison-Burch, C. (2013). Ppdb: The paraphrase database. In Lucy Vanderwende, et al., editors, *HLT-NAACL*, pages 758–764. The Association for Computational Linguistics.

Glockner, M., Shwartz, V., and Goldberg, Y. (2018). Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July. Association for Computational Linguistics.

Gold, D., Kovatchev, V., and Zesch, T. (2019). Annotating and analyzing the interactions between meaning relations. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 26–36, Florence, Italy, August. Association for Computational Linguistics.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., and Smith, N. A. (2018). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June. Association for Computational Linguistics.

Iyer, S., Dandekar, N., and Csernai, K. (2017). First quora dataset release: Question pairs.

Ko, W.-J., Durrett, G., and Li, J. J. (2019). Domain agnostic real-valued specificity prediction. In *AAAI*.

Kovatchev, V., Martí, M. A., and Salamó, M. (2018a). Etpc - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, may.

Kovatchev, V., Martí, M. A., and Salamó, M. (2018b). WARP-text: a web-based tool for annotating relationships between pairs of texts. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 132–136, Santa Fe, New Mexico, August. Association for Computational Linguistics.

Kovatchev, V., Martí, M. A., Salamó, M., and Beltrán, J. (2019). Qualitative evaluation of paraphrase identification systems. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2019*.

Lan, W. and Xu, W. (2018). Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of COLING 2018*.

Lan, W., Qiu, S., He, H., and Xu, W. (2017). A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1224–1234.

LoBue, P. and Yates, A. (2011). Types of common-sense knowledge needed for recognizing textual entailment. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 329–334, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L., and Agirre, E. (2016). Interpretable semantic textual similarity: Finding and explaining differences between sentences. *CoRR*, abs/1612.04868.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Naik, A., Ravichander, A., Sadeh, N., Rose, C., and Neubig, G. (2018). Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Sammons, M., Vydiswaran, V. G. V., and Roth, D. (2010). "Ask Not What Textual Entailment Can Do for You...". In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 1199–1208.

Vila, M., Martí, M. A., and Rodríguez, H. (2014). "Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology. ". *Open Journal of Modern Linguistic*, pages 205–218.

Vila, M., Bertran, M., Martí, M. A., and Rodríguez, H. (2015). Corpus annotation with paraphrase types: new annotation scheme and inter-annotator agreement measures. *Language Resources and Evaluation*, 49(1):77–105.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China, November. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November. Association for Computational Linguistics.

Yager, R. (1992). Default knowledge and measures of specificity. 61:1–44, 04.