

PARAPHRASING, TEXTUAL ENTAILMENT, AND SEMANTIC SIMILARITY ABOVE WORD LEVEL

VENELIN ORLINOV KOVATCHEV

Tesis presentada para optar
al grado de **Doctor en Lingüística** con mención europea
en el programa de doctorado *Ciencia Cognitiva y Lenguaje*,
Departament de Filologia Catalana i Lingüística General,
Universitat de Barcelona,

bajo la supervisión de
Dra. M. Antònia Martí
Universitat de Barcelona
Dra. Maria Salamó
Universitat de Barcelona



UNIVERSITAT DE
BARCELONA

Mayo de 2020

To Mila, who supported me every step of the way.

To Maya and Orlin, for always encouraging my curiosity.

Abstract

This dissertation explores the linguistic and computational aspects of the meaning relations that can hold between two or more complex linguistic expressions (phrases, clauses, sentences, paragraphs). In particular, it focuses on Paraphrasing, Textual Entailment, Contradiction, and Semantic Similarity. This thesis is composed of seven different articles and is divided into three thematic Parts.

In *Part I: “Similarity at the Level of Words and Phrases”*, I study the Distributional Hypothesis (DH). DH is central for most contemporary approaches for automatic processing of meaning and meaning relations within Computational Linguistics (CL) and Natural Language Processing (NLP). Part I of this thesis explores different methodologies for quantifying semantic similarity at the levels of words and short phrases. I measure the importance of the corpus size and the role of linguistic preprocessing. I also show that (lexical) semantic similarity can interact with syntactic-based compositional rules and result in productive patterns at the phrase level. The research in Part I resulted in the publication of two articles.

In *Part II: “Paraphrase Typology and Paraphrase Identification”*, I focus on the meaning relation of paraphrasing and the empirical task of automated Paraphrase Identification (PI). Paraphrasing is one of the most widely studied meaning relation both in theoretical and practical research. PI is among the most popular tasks in CL and NLP. In Part II of this thesis I present: 1) EPT: a new typology of the linguistic and reason-based phenomena involved in paraphrasing; 2) WARP-Text: a new web-based annotation interface capable of annotating paraphrase types; 3) ETPC: the largest corpus to date to be annotated with paraphrase types; and 4) a qualitative evaluation framework for automated PI systems. The findings presented in Part II provide in-depth knowledge on the nature of the paraphrasing relation and improve the evaluation, interpretation, and error analysis in the task of PI. The research in Part II resulted in the publication of three articles.

In *Part III: “Paraphrasing, Textual Entailment, and Semantic Similarity”*, I present a novel direction in the research on textual meaning relations, resulting from joint research carried out on paraphrasing, textual entailment, contradiction, and semantic similarity. Traditionally, these meaning relations are studied in isolation and the transfer of knowledge and resources between them is limited.

In Part III of this thesis I present: 1) a methodology for the creation and annotation of corpora containing multiple textual meaning relations; 2) the first corpus annotated independently with Paraphrasing, Textual Entailment, Contradiction, Textual Specificity, and Semantic Similarity; 3) a statistical corpus-based analysis of the interactions, correlations, and overlap between the different meaning relations; 4) SHARel - a shared typology of textual meaning relations; 5) a corpus of paraphrasing, textual entailment, and contradiction annotated with SHARel. Part III of the thesis gives a new perspective on the research of textual meaning relations. I show that a joint study of multiple meaning relations is both possible and beneficial for processing and analyzing each individual relation. I provide the first empirical data on the interactions between paraphrasing, textual entailment, contradiction, and semantic similarity. The research in Part III resulted in the publication of two articles.

This thesis has advanced our understanding of important issues associated with the empirical analysis, corpus annotation, and computational treatment of textual meaning relations. I have addressed existing gaps in the research field, posed new research questions, and explored novel research directions. The findings and resources presented in this dissertation have been released to the community to facilitate further research and knowledge transfer.

Resumen

En esta tesis se exploran los aspectos lingüísticos y computacionales de las relaciones semánticas que puede haber entre dos o más expresiones lingüísticas complejas (sintagmas, cláusulas, oraciones, párrafos). En particular, se centra en la paráfrasis, la implicación, la contradicción y la similitud semántica. La tesis se compone de siete artículos y se estructura en tres partes.

En la *Parte I: “Similitud de palabras y sintagmas”*, realizo un estudio sobre la Hipótesis distribucional (HD). La HD es relevante en muchos de los trabajos actuales sobre el procesamiento del significado y de las relaciones de significado en el área de la Lingüística Computacional (LC) y el Procesamiento del Lenguaje Natural (PLN). En esta parte se exploran diferentes métodos para la cuantificación de la similitud semántica de palabras y de sintagmas. He calculado la importancia del tamaño del corpus y el papel que juega el preprocesado lingüístico. También muestro que la similitud semántica léxica puede interactuar con reglas de composición sintáctica lo que da como resultado patrones productivos al nivel de sintagma. La investigación de esta parte de mi tesis ha dado lugar a la publicación de dos artículos.

En la *Parte II: “Tipología de paráfrasis e identificación de paráfrasis”* me centro en la relación semántica de paráfrasis y en la tarea empírica de la identificación automática de paráfrasis (IP). La paráfrasis es una de las relaciones de significado más estudiadas, tanto a nivel teórico como aplicado. La IP es una de las tareas más populares en LC y en el PLN. En la Parte II de esta tesis presento: 1) EPT, una nueva tipología de fenómenos lingüísticos y de fenómenos basados en el razonamiento implicados en la paráfrasis; 2) WARP-Text, una nueva interfaz web para la anotación de diferentes tipos de paráfrasis; 3) ETPC: hasta el momento, el corpus de mayor tamaño anotado con tipos de paráfrasis; y 4) un entorno de evaluación cualitativa de sistemas automáticos de IP; Los resultados de esta segunda parte proporcionan un conocimiento más a fondo sobre la naturaleza de la relación de paráfrasis y mejoran la evaluación, interpretación y análisis de errores referentes a la tarea de IP. La investigación de esta segunda parte ha dado lugar a tres publicaciones.

En la *Parte III: “Paráfrasis, Implicación textual y Similitud semántica”*, pre-

sento una nueva línea en la investigación sobre las relaciones de significado. Llevo a cabo una investigación conjunta sobre paráfrasis, implicación textual, contradicción y similitud semántica. Tradicionalmente, estas relaciones se han estudiado separadamente y la transferencia de conocimiento entre ellas ha sido muy limitado. En esta tercera parte de la tesis presento: 1) una metodología para la creación y anotación de corpus que contienen diversas relaciones de significado; 2) el primer corpus anotado independientemente con Paráfrasis, Implicación textual, Contradicción, Especificidad y Similitud semántica; 3) un análisis estadístico de las interacciones, correlaciones y coincidencias entre las diferentes relaciones de significado; 4) SHARel, una tipología compartida para las relaciones semánticas textuales; 5) un corpus de paráfrasis, implicación textual y contradicción anotado con SHARel. Esta tercera parte de la tesis da una nueva perspectiva sobre la investigación en las relaciones de significado a nivel textual. Pongo de manifiesto que es posible el estudio conjunto de diversas relaciones de significado y también que repercute positivamente para cada una de las relaciones en particular. Proporciono por primera vez un conjunto de datos empíricos sobre la integración de paráfrasis, implicación textual, contradicción y similitud semántica. La investigación de esta tercera parte ha dado lugar a dos artículos.

Esta tesis ha permitido avanzar en la comprensión de aspectos importantes relacionados con el análisis empírico, la anotación de corpus, y el tratamiento computacional de las relaciones de significado a nivel textual. He tratado diversas áreas de conocimiento poco atendidas hasta ahora, he planteado nuevas preguntas para la investigación posterior y he explorado en nuevas directrices. Los resultados y recursos presentados en esta tesis son de libre disposición para el colectivo que investiga en LC y PLN con el fin de facilitar la investigación futura y la transferencia de conocimiento.

Acknowledgments

First of all, I'd like to thank my partner, Mila. She has been with me every step of the way, through deadlines, submissions, acceptances, and rejections. She encouraged me and supported me throughout the whole process, she changed countries and jobs and never lost faith in me. She has also endured countless hours of talks about language, cognition, and machine learning.

I would also like to thank my supervisors, Toni Martí and Maria Salamó. It has been a great privilege for me to work with them over the past five years. They have helped me and guided me through my Master and PhD and have taught me everything I know about Computational Linguistics and Natural Language Processing. They have given me so much of their time, energy, and knowledge and have been incredibly patient with my stubbornness. I could not have asked for better supervisors.

I would also want to thank Torsten Zesch for hosting me for a research stay at the University of Duisburg-Essen. His feedback on my work has given me a different perspective on the problem of processing textual meaning relations and has helped me improve as a researcher.

Horacio Rodríguez Hontoria proposed the topic of this thesis and has been very helpful with ideas, references and feedback during my PhD. It was a pleasure to work with him and I admire his productivity and competence.

I have been very lucky to meet and collaborate with fantastic colleagues at the University of Barcelona and the University of Duisburg-Essen. I'm very grateful to Mariona Taulé, Darina Gold, Javier Beltran, Eloi Puertas, Montse Nofre and David Bridgewater.

Many colleagues and friends have played an important role in the success of my PhD. Irina Temnikova was one of the first people that I met during my research. She has been very kind and helpful to me and introduced me to the CL and NLP community and has always been a good friend. I would also like to thank Ahmed AbuRa'ed, Amir Hazem, Sanja Štajner, Tobias Horsmann, Michael Wojatzki, Alejandro Ariza Casabona, Jeremy Barnes, Alexander Popov, Thomas O'Rourke, Kristen Schroeder, Elisabet Vila Borrellas, Ruslan Mitkov, Galia Angelova, Nadezhda Stoyanova, Nikolay Metev, Maria Marinova-Panova, Penko

Kirov, Stefan Lekov, Irina Ivanova, Anna Ignatova, Chris Childress, Theresia Scholten, Andhi Tang, Gabriel Sevilla, Flavia Felletti, and the whole team behind LxMLS.

Last but not least I am grateful to my family. My parents Maya and Orlin, my grandparents Anka, Emilia, and Venelin, my sisters Alissa and Emily, and my aunt Anissia. They have all been very supportive and have always been excited to learn more about my research.

This work has been funded by the APIF grant awarded to me by the University of Barcelona; by the Spanish Ministry of Economy Project TIN2015-71147-C2-2; by the Spanish Ministry of Science, Innovation, and Universities Project PGC2018-096212-B-C33; and by the CLiC research group (2017 SGR 341).

Contents

Abstract	i
Resumen	iii
Acknowledgments	v
List of Figures	xiii
List of Tables	xvi
1 Introduction	1
1.1 Related Work	2
1.1.1 Textual Meaning Relations. Empirical Tasks	3
1.1.2 Typologies of Textual Meaning Relations	9
1.1.3 Joint Research on Textual Meaning Relations	12
1.1.4 Other Related Work	15
1.2 Motivation and Objectives of the Thesis	16
1.2.1 Paraphrase Typology and Paraphrase Identification	16
1.2.2 Joint Study on Meaning Relations	17
1.3 Thesis Development	17
1.4 Thesis Outline	20
 I Similarity at the Level of Words and Phrases	 23
2 Comparing Distributional Semantics Models for Identifying Groups of Semantically Related Words	25
2.1 Introduction	26
2.2 Related Work	26
2.3 Data and Tools	28
2.3.1 The Corpus	28
2.3.2 Grouping with CLUTO	28

2.3.3	Grouping with Word2Vec	31
2.4	Experiments	32
2.4.1	Setup	32
2.4.2	Evaluation	33
2.4.3	Results	34
2.5	Conclusions and Future Work	36
3	DISCOVer: DIStributional Approach Based on Syntactic Dependencies for Discovering CONstructions	39
3.1	Introduction	40
3.2	Related Work	42
3.3	Methodology for Discovering Constructions	45
3.3.1	Description of the Task	46
3.3.2	The Corpus	46
3.3.3	Matrix	47
3.3.4	Clustering	50
3.3.5	Generalization: Linking and Filtering Clusters	53
3.3.6	Pattern Generation	55
3.4	Evaluation	58
3.4.1	Clustering Evaluation	58
3.4.2	Pattern Evaluation	58
3.5	Conclusions and Future Work	65
II	Paraphrase Typology and Paraphrase Identification	67
4	WARP-Text: a Web-Based Tool for Annotating Relationships between Pairs of Texts	69
4.1	Introduction	70
4.2	Related Work	70
4.3	WARP-Text	71
4.3.1	Annotation Scheme	72
4.3.2	Administrator Interface	72
4.3.3	Annotator Interface	73
4.4	Conclusions and Future Work	75
5	ETPC - a Paraphrase Identification Corpus Annotated with Extended Paraphrase Typology and Negation	77
5.1	Introduction	78
5.2	Related Work	79

5.3	Extended Paraphrase Typology	80
5.3.1	Basic Terminology	81
5.3.2	From Atomic to Textual Paraphrases	81
5.3.3	Objectives of EPT and Research Questions.	83
5.3.4	The Extended Paraphrase Typology	83
5.4	Annotation Scheme and Guidelines	86
5.4.1	Non-Sense Preserving Atomic Phenomena	87
5.4.2	Sense Preserving Atomic Phenomena	88
5.4.3	Inter-Annotator Agreement	89
5.4.4	Annotation of Negation	91
5.5	The ETPC Corpus	91
5.5.1	Non-Sense Preserving Atomic Phenomena	91
5.5.2	Sense Preserving Atomic Phenomena	92
5.5.3	Discussion	93
5.5.4	Applications of ETPC	94
5.6	Conclusions and Future Work	95
6	A Qualitative Evaluation Framework for Paraphrase Identification	97
6.1	Introduction	98
6.2	Related Work	99
6.3	Qualitative Evaluation Framework	100
6.3.1	The ETPC Corpus	100
6.3.2	Evaluation Methodology	100
6.4	PI Systems	101
6.5	Results	103
6.5.1	Overall Performance	103
6.5.2	Full Performance Profile	104
6.5.3	Comparing Performance Profiles	107
6.5.4	Comparing Performance by Phenomena	108
6.6	Discussion	109
6.7	Conclusions and Future Work	110
III	Paraphrasing, Textual Entailment, and Semantic Similarity	113
7	Annotating and Analyzing the Interactions between Meaning Relations	115
7.1	Introduction	116
7.2	Related Work	117

7.2.1	Interactions between Relations	117
7.2.2	Corpora with Multiple Semantic Layers	118
7.3	Corpus Creation	119
7.3.1	Sentence Pool	119
7.3.2	Pair Generation	120
7.3.3	Relation Annotation	121
7.3.4	Final Corpus	123
7.4	Interactions between Relations	126
7.4.1	Correlations between Relations	126
7.4.2	Overlap of Relation Labels	128
7.4.3	Discussion	130
7.5	Conclusion and Further Work	131
8	Decomposing and Comparing Meaning Relations:	
	Paraphrasing, Textual Entailment, Contradiction, and Specificity	133
8.1	Introduction	134
8.2	Related Work	136
8.3	Shared Typology for Meaning Relations	137
8.3.1	Decomposing Meaning Relations	138
8.3.2	The SHARel Typology	138
8.3.3	Research Questions	141
8.4	Corpus Annotation	142
8.4.1	Choice of Corpus	142
8.4.2	Annotation Setup	143
8.4.3	Agreement	143
8.5	Analysis of the Results	145
8.5.1	Type Frequency	145
8.5.2	Decomposing Specificity	149
8.6	Discussion	151
8.7	Conclusions and Future Work	152
9	Conclusions	155
9.1	Contributions and Discussion of the Results	155
9.2	Resources	160
9.3	Future Research Directions	160
A	Annotation Guidelines for ETPC	183
A.1	Presentation	183
A.1.1	Credits	183
A.2	The task	184

A.2.1	Is This a Paraphrase Pair	184
A.2.2	The Tagset	185
A.2.3	The Scope	187
A.3	Tagset Definition	191
A.3.1	Morphology based changes	191
A.3.2	Lexicon based changes	192
A.3.3	Lexico-syntactic based changes	197
A.3.4	Syntax based changes	201
A.3.5	Discourse based changes	204
A.3.6	Other changes	206
A.3.7	Extremes	207
A.4	Annotating non-paraphrases	209
A.5	Annotating negation	211
B	Annotation Guidelines for Gold et al. [2019]	213
B.1	Paraphrasing	213
B.2	Textual Entailment	214
B.3	Contradiction	215
B.4	Similarity	216
B.5	Specificity	217
C	Annotation Guidelines for Kovatchev et al. [2020]	219
C.1	Presentation	219
C.2	Annotating reason-based Phenomena	219
C.3	List of reason-based phenomena	220

List of Figures

3.1	Main steps in DISCOVer methodology	45
3.2	Dependency parsed sentence: <i>El barbero afeita la larga barba de Jaime</i> ('The barber shaves off James's long beard')	49
4.1	Annotating relationships at textual level.	73
4.2	Annotating relationships at token level.	74
4.3	Scope selection page.	75
6.1	Critical Difference diagram of the average ranks by phenomena . .	109
7.1	Similarity scores of sentences annotated with different relations . .	127

List of Tables

1.1	Typologies of textual meaning relations	11
1.2	Popular corpora for textual meaning relations	13
2.1	Diana-Araknion Format	29
2.2	PoS tagset modifications	30
2.3	Syntactic Dependencies	31
2.4	Wordnet Similarity	34
2.5	Expert evaluation	35
2.6	PoS coherence	36
3.1	Example of a real cluster (421_n) in the Diana-Araknion corpus in Spanish	52
3.2	Some examples of cluster linking process in cluster $i=421_n$ (de- scribed in Table 3.1).	54
3.3	Distribution of the number of related and unrelated clusters and their percentage	55
3.4	Distribution of the number of related clusters and their percentage by POS	56
3.5	Distribution of the generated patterns	56
3.6	Association score of Attested-Patterns compared with statistical chance	61
3.7	Average association score of Attested-Patterns and BI-patterns . .	62
3.8	Occurrence of Unttested-Patterns and FL-Patterns	62
3.9	Association scores of Unttested-Patterns	63
3.10	Interannotator agreement test	64
3.11	Expert evaluation	64
5.1	Extended Paraphrase Typology	84
5.2	Non-sense preserving phenomena	88
5.3	Sense preserving phenomenon	89
5.4	Inter-annotator Agreement	90

5.5	Distribution of non-sense preserving phenomena	92
5.6	Distribution of Sense preserving phenomena in textual paraphrases and textual non-paraphrases	93
6.1	Overall Performance of the Evaluated Systems	103
6.2	Performance profile of Wang et al. [2016]	105
6.3	Performance profiles of all systems	106
6.4	Difference in phenomena performance between S3 [Wang et al., 2016] and S4 [He and Lin, 2016]	107
6.5	Difference in phenomena performance: S3 [Wang et al., 2016] and S5 [Lan and Xu, 2018b]	108
7.1	List of given source sentences	119
7.2	Inter-annotator agreement for binary relations ✓denotes a relation being there ✗denotes a relation not being there	123
7.3	Distribution of Inter-annotator agreement	124
7.4	Distribution of meaning relations within different pair generation patterns	125
7.5	Comparison of BLEU scores between the sentence pairs in differ- ent corpora	126
7.6	Correlation between all relations	127
7.7	Distribution of overlap within relations	128
7.8	Annotations of sentence pairs on all meaning relations taken from our corpus	129
8.1	The SHARel Typology	139
8.2	Comparing typologies of textual meaning relations	141
8.3	Inter-annotator Agreement	144
8.4	Type Frequencies	146
8.5	Decomposition of Specificity	151
A.1	Extended Paraphrase Typology	186

Chapter 1

Introduction

This thesis is about the meaning relations that can hold between language expressions (words, phrases, clauses, and sentences). In particular, it focuses on the meaning relations of paraphrasing, textual entailment, and semantic similarity. The automatic processing of these meaning relations is an unsolved problem in Computational Linguistics (CL) and Natural Language Processing (NLP) and has attracted the attention of many researchers. This thesis explores two different directions within the research on meaning relations:

1. Incorporating linguistic knowledge in the empirical tasks of processing meaning relations. In particular, I focus on the paraphrasing meaning relation and the empirical task of Paraphrase Identification (PI). By combining PI with Paraphrase Typology (PT) I aim:
 - a) to improve the evaluation and interpretation of automated PI systems.
 - b) to empirically validate PT.
2. Analyzing and processing multiple meaning relations together. I contrast previous work and propose a novel research approach that does not focus on a single meaning relation. I present a joint study on Paraphrasing, Textual Entailment, Contradiction, and Semantic Similarity:
 - a) to compare the different meaning relations empirically.
 - b) to create a shared typology for textual meaning relations.

My work offers a valuable insight into the nature and interactions of the different meaning relations and also aims to improve the automated systems for processing meaning relations. I also release to the community three new corpora, two new typologies of meaning relations, a new web-based annotation tool, and a new

software program for a qualitative evaluation of automated paraphrase identification systems.

The structure of this thesis is intentionally chronological¹ in order to capture the four year development of the ideas and arguments behind the thesis. The thesis consists of nine Chapters, organized as follows:

- Chapter 1 is the Introduction.
- Chapters 2 to 8 correspond to seven published articles. They are grouped in three thematically organized parts.
- Chapter 9 presents the contributions, the discussion of the results, the conclusions, and the directions for future work.

The rest of this Introduction chapter is organized as follows. In Section 1.1, I familiarize readers with the related work in the research on meaning relations. In Section 1.2, I present my main objectives and justify them in the context of the preexisting research. In Section 1.3, I describe the development of this thesis and the connecting thread that runs between the individual articles. Finally in Section 1.4, I present the outline and structure of the whole dissertation.

1.1 Related Work

This section is meant to provide the reader with a compact overview of the previous and latest research related to this thesis in order to supplement and bind together the “background” sections in each paper. From a thematic perspective, the subject matter can be broken down into the following research areas:

- (i) Textual Meaning Relations. Empirical Tasks. (**Section 1.1.1**)
- (ii) Typologies of Textual Meaning Relations (**Section 1.1.2**)
- (iii) Joint Research on Textual Meaning Relations (**Section 1.1.3**)
- (iv) Other Related Work (**Section 1.1.4**)

I will deal with each of the areas in turn, highlighting the main trends and milestones. The reader is referred to the original papers for details.

¹Articles are presented in the order in which they were written, which does not necessarily correspond to the order in which they were published.

1.1.1 Textual Meaning Relations. Empirical Tasks

Meaning relations between complex language expressions (e.g.: clauses, sentences, paragraphs), henceforth “textual meaning relations” are the object of study of this thesis. Research on textual meaning relations has to account not only for the meaning of a single word or a phrase, but also for the compositionality of meaning. In this thesis, I focus on the textual meaning relations of Paraphrasing, Textual Entailment, Contradiction², and Semantic Similarity. It is important to note that the interactions between the different relations are non-trivial. In some cases they can overlap (e.g.: two texts that are paraphrases often also hold an entailment relation) and in some cases the negative examples for one relation can be positive examples for another (e.g.: two texts that are not paraphrases can sometimes hold an entailment relation or a contradiction relation).

Empirical Tasks on Textual Meaning Relations

Androutsopoulos and Malakasiotis [2010] distinguish three types of empirical tasks that are focused on processing meaning relations: recognition, generation, and extraction. Their definitions for these paraphrasing and textual entailment tasks are as follows:

Recognition: *“The main input to a paraphrase or textual entailment recognizer is a pair of language expressions (or templates), possibly in particular context. The output is a judgment, possibly probabilistic, indicating whether or not the members of the input pair are paraphrases or a correct textual entailment pair; the judgments must agree as much as possible with those of humans.”*

Generation: *“The main input to a paraphrase or textual entailment generator is a single language expression (or template) at a time, possibly in a particular context. The output is a set of paraphrases of the input or a set of language expressions that entail or are entailed by the input; the output set must be as large as possible, but including as few errors as possible.”*

Extraction: *“The main input to a paraphrase or textual entailment extractor is a corpus, for example a monolingual corpus of parallel or comparable texts. The system outputs pairs of paraphrases (possibly templates) or pairs of language expressions (or templates) that constitute correct textual entailment pairs, based on the evidence of the corpus; the goal is again to produce as many output pairs as possible, with as few errors as possible.”*

²Contradiction is typically studied jointly with Textual Entailment.

The empirical tasks focused on the automatic processing of meaning relations are inspired by human capabilities. We, as competent language users, can quickly and unconsciously determine the meaning relation that holds between two simple or complex language expressions. We can successfully recognize, generate, and extract paraphrases, entailment pairs, and contradiction pairs. The empirical tasks focused on textual meaning relations in CL and NLP aim to produce automated systems that can achieve on-task performance comparable with that of humans. Human judgments are typically taken as a gold standard for evaluation.

In this thesis, I focus on recognition tasks. In particular, I study **Paraphrase Identification**, **Recognizing Textual Entailment**, and **Semantic Textual Similarity**. In the rest of this section I present the definition, corpora, and state-of-the-art for each of these three tasks.

Paraphrase Identification

Task format and definition: Paraphrase Identification (PI) is framed as a binary classification task. In PI, a human or an automated system needs to determine whether or not a paraphrasing relation holds between two given texts. The definition of “paraphrasing” provided by Dolan and Brockett [2005] is “*whether two sentences at the high level “mean the same thing” /.../ despite obvious differences in information content.*”.

- (1) **Sentence 1:** The genome of the fungal pathogen that causes Sudden Oak Death has been sequenced by US scientists.

Sentence 2: Researchers announced Thursday they’ve completed the genetic blueprint of the blight-causing culprit responsible for sudden oak death.

Two sentences that are connected with a paraphrasing relation can be seen in Example 1³. While the two sentences are not completely equivalent, in the context of PI they are considered paraphrases. Dolan and Brockett [2005] argue that if human annotators are required to only mark full equivalence of meaning, only identical sentences are considered paraphrases. Therefore, in the practical setting of PI, they propose a less strict definition of paraphrasing and allow for some difference in the information content.

PI Corpora: The task of (PI) was first popularized with the creation of the Microsoft Research Paraphrase Corpus (MRPC), presented in Dolan et al. [2004] and Dolan and Brockett [2005]. The MRPC corpus is semi-automatically created from the articles in the news domain and consists of 5,801 text pairs, annotated as

³The example is taken from Dolan and Brockett [2005].

“paraphrase” or “non-paraphrase”. To date, MRPC is still used for the evaluation of automated PI systems despite, its relatively small size.

The Paraphrase Database (PPDB) [Ganitkevitch et al., 2013] (and later on its second version PPDB2 [Pavlick et al., 2015]) was the first large scale paraphrase corpus. It is an automatically constructed collection of over 100 million paraphrases at different granularity. While the MRPC only contains sentences and longer chunks of text, the PPDB also contains “paraphrases” of words and short phrases. The second version of PPDB also includes the entailment relation. PPDB and PPDB2 are collections of paraphrases, rather than corpora specifically created for the task of PI. However, they can be adapted for use in PI tasks.

The Quora Question Pair Dataset [Iyer et al., 2017] is a semi-automatically collected corpus of 400,000 question pairs marked as “duplicate” or “non-duplicate” by Quora users. The corpus was used in an online competition⁴ and facilitated the use of Deep Learning based systems for the task of PI. Due to its size, the Quora corpus is very popular for training state-of-the-art PI systems.

The Language-Net corpus [Lan et al., 2017] is the largest PI dataset to date. It was extracted from Twitter and contains over 51,000 human-annotated sentence pairs and over 2.8 million automatically extracted candidate paraphrases.

MRPC, PPDB, Quora, and Language-net are all created for the English language. The work on PI for languages other than English is very limited. We can mention the work of Creutz [2018] on the creation of paraphrase corpus in six languages using open subtitles dataset.

State-of-the-art in PI: The first automated PI systems were based on manually engineered features [Finch et al., 2005, Kozareva and Montoyo, 2006] or on a combination of lexical similarity metrics and cosine similarity [Mihalcea et al., 2006]. Word2Vec [Mikolov et al., 2013b] and Glove [Pennington et al., 2014] introduced a new paradigm in PI, but also in CL and NLP in general. The systems based on Word2Vec and Glove outperformed previous unsupervised systems and pushed the state-of-the-art further. Deep Learning based systems using autoencoders [Socher et al., 2011], Long Short Term Memory Networks (LSTM) [He and Lin, 2016], and Convolutional Neural Networks (CNN) [He et al., 2015] set the new state-of-the-art for the Supervised PI systems. More recently, Transformer based architectures [Devlin et al., 2019] have made a considerable improvement to automated PI systems, approaching human level performance on the datasets⁵.

⁴<https://www.kaggle.com/c/quora-question-pairs>

⁵The official ACL page for PI ([https://aclweb.org/aclwiki/Paraphrase_Identification_\(State_of_the_art\)](https://aclweb.org/aclwiki/Paraphrase_Identification_(State_of_the_art))) and the GLUE benchmark page (<https://gluebenchmark.com/leaderboard>) contain the full leaderboard of PI systems for a variety of corpora.

Recognizing Textual Entailment

Task format and definition: Recognizing Textual Entailment (RTE), also known as Natural Language Inference (NLI), has two different formats. The original RTE was framed as a binary classification task. In RTE, a human or an automated system needs to determine whether or not a paraphrasing relation holds between two given texts. The practical definition of Textual Entailment in RTE is *“a directional relationship between pairs of text expressions, denoted by T - the entailing “Text”, and H - the entailed “Hypothesis”. We say that T entails H if the meaning of H can be inferred from the meaning of T , as would typically be interpreted by people.”*. An example of textual entailment relation can be seen in 2. In the example given the Text entails Hyp 1, but not Hyp 2, or Hyp 3.

- (2) **Text:** The purchase of Houston-based LexCorp by BMI for \$2Bn prompted widespread sell-offs by traders as they sought to minimize exposure. LexCorp had been an employee-owned concern since 2008.

Hyp 1: BMI acquired an American company.

Hyp 2: BMI bought employee-owned LexCorp for \$3.4Bn.

Hyp 3: BMI is an employee-owned concern.

The second format of the RTE was introduced in [Giampiccolo et al., 2008] and the task was reformulated as a three class classification between “entailment”, “contradiction”, and “neutral” text pairs. In example 2, the Text entails Hyp 1, contradicts Hyp 2, and is neutral with respect to Hyp 3.

RTE Corpora: The task of RTE was popularized with the introduction of the yearly Recognizing Textual Entailment challenge in Dagan et al. [2006]. The first three editions of the RTE challenge were called the Pascal RTE challenge [Dagan et al., 2006, Bar-Haim et al., 2006, Giampiccolo et al., 2007] and were framed as a binary classification between “entailment” and “non entailment” text pairs. In the fourth edition of the challenge [Giampiccolo et al., 2008], the Pascal RTE challenge became the Text Analysis Conference (TAC) RTE challenge. The task was reformulated as a three class classification between “entailment”, “contradiction”, and “neutral” text pairs. The TAC RTE challenge ran for four years: Giampiccolo et al. [2008], Bentivogli et al. [2009], Bentivogli et al. [2010], and Bentivogli et al. [2011]. Like the MRPC corpus, the RTE datasets are not very large in size, however due to the high quality of the annotation they are still used as an evaluation benchmark for state-of-the-art systems.

The increasing popularity of Deep Learning systems and the need for more training data led to the creation of the Stanford Natural Language Inference corpus (SNLI) [Bowman et al., 2015] and later on the Multi-Genre Natural Language

Inference corpus (MultiNLI) [Williams et al., 2018]. The SNLI contains 570,000 human-written English sentences, while the MultiNLI contains 433,000 sentences but covers a more diverse range of texts. SNLI and MultiNLI are currently the most popular corpora for training automated RTE/NLI systems. Both SNLI and MultiNLI use the three-way classification format of the task.

As with PI, the work on RTE and NLI is mostly for English. Notable exceptions are the XNLI corpus [Conneau et al., 2018], a machine-translated portion of MultiNLI and the SPARTE corpus [Peñas et al., 2006] for RTE in Spanish, created from question-answering corpora.

State-of-the-art in RTE: The development of the automated RTE/NLI systems follows a similar trend as the development of the automated PI systems. The first RTE systems used manually engineered features and simple similarity metrics. Then, there was a paradigm shift towards various Deep Learning architectures, such as autoencoders, LSTMs, and CNNs. And finally, the current state-of-the-art are Transformer based architectures ⁶.

With the state-of-the-art systems approaching human level performance on the datasets, many researchers have tried to **analyze the workings** of the different RTE and NLI systems. 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 67% (SNLI) and 52.3-53.9% (MultiNLI) accuracy, which is substantially higher than the majority baselines of 34-35%. Glockner et al. [2018] showed that models trained with SNLI fail to resolve new pairs that require simple lexical substitution. For example the models have problems determining that “*holding a saxophone*” contradicts “*holding an electric guitar*”. The human annotators indicate a contradiction in this example, as the annotation guidelines instruct them to assume that the same event is referred to by both texts. Naik et al. [2018] created label-preserving adversarial examples and concluded that automated NLI models are not robust. Wallace et al. [2019] introduced universal triggers, that is, sequences of tokens that fool models when concatenated to any input.

All of these findings indicate that the existing RTE and NLI datasets are much simpler than what native speakers are capable of. Furthermore, the datasets contain many annotation artifacts and the systems trained on them are not robust to adversarial examples. Therefore, despite the high performance achieved on the datasets, the general problem of RTE and NLI is far from resolved.

⁶The official ACL page for the RTE Challenge (https://aclweb.org/aclwiki/Recognizing_Textual_Entailment), the official SNLI corpus page (<https://nlp.stanford.edu/projects/snli/>), the official MultiNLI corpus page (<https://www.nyu.edu/projects/bowman/multinli/>), and the GLUE benchmark page (<https://gluebenchmark.com/leaderboard>) contain the full leaderboard of RTE systems for a variety of corpora.

Semantic Textual Similarity

Task format and definition: Semantic Textual Similarity (STS) is framed as a regression task. In STS, a human or an automated system needs to determine the degree of similarity between two given texts on a continuous scale from 0 to 5. The practical definition for Semantic Similarity in STS is “*how similar two sentences are to each other according to the following scale:*

[5] Completely equivalent, as they mean the same thing.

[4] Mostly equivalent, but some unimportant details differ.

[3] Roughly equivalent, but some important information differs/missing.

[2] Not equivalent, but share some details.

[1] Not equivalent, but are on the same topic

[0] On different topics.

Examples for each semantic similarity from 0 to 5 can be seen in 3.

(3) **Similarity 5:**

The bird is bathing in the sink.

Birdie is washing itself in the water basin.

Similarity 4:

In May 2010, the troops attempted to invade Kabul.

The US army invaded Kabul on May 7th last year, 2010.

Similarity 3:

John said he is considered a witness but not a suspect.

“He is not a suspect anymore.” John said.

Similarity 2:

They flew out of the nest in groups.

They flew into the nest together.

Similarity 1:

The woman is playing the violin.

The young lady enjoys listening to the guitar.

Similarity 0:

John went horse back riding at dawn with a whole group of friends.

Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

STS Corpora: The most popular corpora for the STS task are the datasets from the yearly STS competition [Agirre et al., 2012]. While the STS corpora are not large in size, they come from a variety of domains and their coverage is extended every year. Unlike the tasks of PI and RTE, the competition in STS includes non-English texts (Arabic, Spanish, Turkish). Also unlike PI and RTE,

at the time this dissertation was begun there were no large scale corpora explicitly designed for STS.

State-of-the-art in STS: The development of the automated STS systems is similar to that of the automated systems for PI and RTE. The system architecture transitions from feature based through Deep Learning based systems, and finally to the current state of the art, which are transformer based systems⁷.

1.1.2 Typologies of Textual Meaning Relations

In the context of the empirical tasks of Paraphrase Identification (PI), Recognizing Textual Entailment (RTE), and Semantic Textual Similarity (STS), the corresponding meaning relations are typically considered atomic. That is, the researchers in these areas make several assumptions about the data and the task:

- Each pair of texts has a single label corresponding to it. The label is one of a pre-defined set.
- The label applies to the whole text pair and cannot be expressed (decomposed) as a combination of more simple phenomena.
- Each pair of texts is processed the same way by the human annotators and the automated systems. It has the same complexity as any other pair in the dataset and it contributes the same weight to the evaluation of the model.

These assumptions are made to facilitate the definition and evaluation of the empirical tasks. However, several researchers working on Paraphrasing, Textual Entailment, and Semantic Similarity have questioned the applicability of these simplifications and have provided counter examples, such as Examples 4 and 5:

- (4) **Sentence 1:** All **kids** receive the same education .
Sentence 2: All **children** receive the same education .
- (5) **Sentence 1:** All **kids** receive the same education .
Sentence 2: The same education is provided to all **children** .

In both Examples 4 and 5, the two texts have approximately the same meaning and they can be labeled as “paraphrases”. In the context of PI, these two examples

⁷The official page for the STS challenge⁸ and the GLUE benchmark page (<https://gluebenchmark.com/leaderboard>) contain the full leaderboard of STS systems for a variety of corpora.

have the same label, the same degree of complexity, and the same weight in the final evaluation of the system. However, when looking at the examples, the human intuition would suggest that:

- Processing Examples 4 and 5 requires different (linguistic) capabilities and follows different (linguistic) strategies.
- Example 5 is arguably harder than Example 4.

These intuitions contradict the empirical assumptions concerning the atomic nature of the data. If the meaning relations are indeed atomic and non-decomposable, then Examples 4 and 5 should have approximately the same degree of complexity and determining the correct label should require similar linguistic capacities and strategies.

Starting from linguistic theory and from examples like 4 and 5, several researchers have questioned the atomic nature of the Paraphrasing, Textual Entailment, and Semantic Similarity meaning relations. The "non-atomic" approach of studying meaning relations historically began in the field of Textual Entailment with the works of Garoufi [2007] and Sammons et al. [2010]. Later on Cabrio and Magnini [2014] carried out a large theoretical and empirical study on the nature of the phenomena involved in entailment. Independently from the research on textual entailment, Vila et al. [2014] and Bhagat and Hovy [2013] proposed different ways to decompose and characterize the paraphrasing relation. In the area of semantic similarity, Agirre et al. [2016] proposed a new task of "interpretable semantic textual similarity".

In the context of this thesis, there are two important hypotheses, shared by the majority of the authors working on decomposing meaning relations.

The first hypothesis argues that in order to determine the meaning relation that holds between two texts, a human or an automated system needs to make one (or more) simple "inference steps". In Example 4, such inference steps would be:

- 1) determining that "kids" in Example 4.1 means the same as "children" in Example 4.2 within the given context.
- 2) determining that all of the linguistic units in the two sentences in Example 4 are the same, except for "kids" - "children".

Based on 1) and 2), a human or an automated system can determine that in Example 4, the two texts have approximately the same meaning and therefore the correct label is "paraphrases". The hypothesis argues that to correctly predict the textual meaning relation in Example 4, a human or an automated system needs to have the capabilities and the background knowledge to process each individual "inference step".

The second hypothesis argues out that a single example can contain various numbers of “inference steps”. Example 4 has two inference steps. Example 5 has one additional step: the substitution of “receive” with “is provided to” and the corresponding change in the syntactic structure of the two sentences. Following from this hypothesis, the different number and nature of inference steps would result in different strategies for processing the examples and different degrees of complexity.

All of the authors working on decomposing meaning relations propose a list of linguistic phenomena that can be considered to be inference steps. In the rest of this dissertation these lists are called “typologies”. In Table 1.1 I compare the different typologies. I also include the data for the two typologies proposed in this thesis: EPT and SHARel, presented in Chapters 5 and 8.

Table 1.1 Typologies of textual meaning relations

Typology	Relation	Types	Lvls	Neg-Ex	Corpus
Garoufi [2007]	TE	28	Yes	Yes	500 pairs
Sammons et al. [2010]	TE, CNT	22	No	Yes	210 pairs
Cabrio and Magnini [2014]	TE, CNT	36	Yes	Yes	500 pairs
Bhagat and Hovy [2013]	PP	25	No	No	355 pairs
Vila et al. [2014]	PP	23	Yes	No	3900 pairs
Agirre et al. [2016]	STS	9	No	Yes	3000 pairs
<i>EPT (Chapter 5)</i>	PP	27	Yes	Yes	5801 pairs
<i>SHARel (Chapter 8)</i>	PP, STS, TE, CNT	34	Yes	Yes	520 pairs

Table 1.1 compares typologies of textual meaning relations in terms of:

Relation: The textual meaning relation (or relations) that can be decomposed using the typology. TE - “Textual Entailment”; CNT - “Contradiction”; PP - “Paraphrasing”; STS - “Semantic Textual Similarity”.

Types: The number of phenomena in the typology.

Lvls: Whether or not the typology is organized in hierarchical levels. For example, some typologies distinguish between morphological, lexical, syntactic, etc. phenomena, while others have no explicit structure.

Neg-Ex: Whether the typology can be used to decompose and analyze negative examples (i.e.: “non-paraphrases”, “non-entailment”, “0 semantic similarity”) or if it is only applicable to positive examples.

Corpus: The size of the available corpora annotated with the typology.

With respect to the **relation**, each typology is built around a single empirical task. The typologies of Garoufi [2007], Bhagat and Hovy [2013], Vila et al. [2014], and Agirre et al. [2016] are all built around a single textual meaning relation. The typologies of Sammons et al. [2010] and Cabrio and Magnini [2014] can be applied to two textual meaning relations: Textual Entailment and Contradiction.

Considering the number of **types**, most of the typologies contain between 23 and 28 phenomena. The majority of these phenomena are in fact shared across the typologies of paraphrasing and textual entailment. The typology for semantic textual similarity is much more simple and task specific.

Taking into account the **levels** of hierarchical structure, three of the typologies [Garoufi, 2007, Cabrio and Magnini, 2014, Vila et al., 2014] organize the types in terms of the linguistic level of the phenomena (morphological, lexical, lexico-syntactic, syntactic, discourse, reasoning). The remaining typologies propose a list of phenomena without trying to organize them.

Looking at the decomposition of **negative examples**, the typologies for textual entailment and semantic similarity can be applied to both positive and negative examples. The typologies for paraphrasing [Bhagat and Hovy, 2013, Vila et al., 2014] can only decompose pairs of text that hold a “paraphrasing” relation. They cannot be applied to “non-paraphrases”.

Finally, with respect to the size of the available **corpora**, most typologies have been used to annotate only a small corpus (200-500 text pairs). Vila et al. [2014] and Agirre et al. [2016] are the only authors that provide corpora of a size sufficient for machine learning experiments.

Table 1.1 demonstrates some clear tendencies across the different typologies. It also illustrates some important gaps in the research field. First, at the time of beginning this dissertation each of the typologies was built around a single task and focused on one (or two) textual meaning relations. There was no typology that could be applied to multiple textual meaning relations without adaptation. Second, at the time of beginning this dissertation there was no corpus of paraphrasing or textual entailment, annotated with a typology and suitable for “recognition” machine learning experiments. The corpora of Garoufi [2007], Sammons et al. [2010], Bhagat and Hovy [2013], and Cabrio and Magnini [2014] are too small in size and the corpus of Vila et al. [2014] contains only “paraphrases”, without negative examples. With the creation of EPT and SHARel, I aimed to address these gaps in the field, as shown in the last two rows of Table 1.1.

1.1.3 Joint Research on Textual Meaning Relations

Despite the obvious similarities and interactions between the textual meaning relations, the joint research on them has been very limited, both in theoretical and

in empirical aspects. Table 1.2 shows some of the most popular corpora explicitly annotated with textual meaning relations. Most of the corpora comes from the empirical tasks of PI, RTE, and STS. I also include the data for the corpus I present in Chapter 7 of this thesis.

Table 1.2 Popular corpora for textual meaning relations

Corpus	Paraph.	Entailment	Contradiction	Similarity
MRPC	Yes	No	No	No
Quora	Yes	No	No	No
Language-Net	Yes	No	No	No
RTE (1-3)	No	Yes	No	No
RTE (4-6)	No	Yes	Yes	No
SNLI	No	Yes	Yes	No
MultiNLI	No	Yes	Yes	No
STS (all)	No	No	No	Yes
SICK	No	Yes	Yes	Yes
Sukhareva et al. [2016]	Yes *	Yes	No	No
Chapter 7 (this thesis)	Yes	Yes	Yes	Yes

Table 1.2 clearly demonstrates the separation between the different meaning relations in existing corpora. Each corpus is typically built around one single relation, or two in the case of textual entailment. At the time of beginning this thesis the only corpora that contained multiple textual meaning relations were:

- the SICK corpus [Marelli et al., 2014], which is annotated for textual entailment, contradiction, and semantic similarity.
- the corpus of Sukhareva et al. [2016] who annotate paraphrasing as a specific sub-class of entailment.

The corpus presented in Chapter 7 addresses this gap in the existing resources and is the first corpus annotated with the four most popular textual meaning relations: Paraphrasing, Textual Entailment, Contradiction, and Semantic Similarity.

In a more theoretical setting, Madnani and Dorr [2010] and Androutsopoulos and Malakasiotis [2010] discuss and compare different aspects of paraphrasing and textual entailment. They argue that paraphrasing is typically a bi-directional entailment. Cabrio and Magnini [2014] and Sukhareva et al. [2016] also suggest that paraphrasing is a sub-class of textual entailment.

However, Dolan and Brockett [2005] point out that if they enforced a strict bi-directional entailment and full equivalence of the information content, the annotators would only mark identical texts as paraphrases, which would make the

Paraphrase Identification task trivial. Therefore in their annotation setup they also allow for a limited difference in the information content in the two texts. As a result, the equivalence between bi-directional entailment and paraphrasing does not hold in their corpus (MRPC). A similar approach to annotating the paraphrasing relation has also been adopted in the rest of the PI corpora. Therefore, the relation between entailment and paraphrasing is non-trivial to define in an empirical setting. However, the lack of corpora annotated for multiple textual meaning relations has limited the possibilities for empirical data-driven research on the interactions between paraphrasing, textual entailment, and contradiction.

There has also been some research on using one textual meaning relation to predict another and for the transfer of knowledge across tasks. Cer et al. [2017] argue that to find paraphrases or entailment, some level of semantic similarity must be given. Bosma and Callison-Burch [2006] use techniques from Paraphrase Identification in order to solve textual entailment. Castillo and Cardenas [2010] and Yokote et al. [2011] use semantic similarity to solve entailment.

The recent work by several authors is indicative of an increasing interest towards the joint study of meaning relations. In particular, the topic is interesting within the context of transfer learning in NLP and CL. Lan and Xu [2018a] and Aldarmaki and Diab [2018] demonstrate the transfer learning capabilities of different systems in the tasks of PI and RTE. They cover a wide range of supervised and unsupervised machine learning architectures and demonstrate promising results.

The interest and success of the transfer learning techniques have also resulted in the creation of the GLUE [Wang et al., 2018] and SuperGLUE [Wang et al., 2019] benchmarks. GLUE and Super GLUE are a collection of multiple datasets for several tasks, including PI, RTE and STS. The authors of those benchmarks argue that systems working on Natural Language Understanding (NLU) should be able to perform well on all of the tasks, and not just on one. The GLUE and SuperGLUE are now the most popular benchmarks for evaluating NLU systems and general purpose meaning representation models.

However, I would argue that a benchmark of multiple datasets is not a replacement for a single dataset annotated with multiple textual meaning relations. Similarly, a transfer learning experiment is not a replacement for a single task of multi-class classification. At the time of beginning this thesis there was an apparent gap in the field - a lack of resources (annotation guidelines and corpora) that would enable the joint theoretical and empirical research of multiple textual meaning relations.

1.1.4 Other Related Work

Distributional Semantics (DS) is the predominant framework for representing and comparing the meaning of linguistic units in contemporary Computational Linguistics (CL) and Natural Language Processing (NLP). DS has an important role both in theoretical research and in developing practical applications. The core hypothesis in DS is the Distributional Hypothesis (DH), as formulated by different authors:

“Difference in meaning correlated with difference in distribution”
[Harris, 1954]

“You shall know a word by the company it keeps”
[Firth, 1957]

“The meaning of a word is its use in the language”
[Wittgenstein, 1953]

While these authors formulate DH in slightly different ways, the central assumption remains the same and can be stated as follows: *“similar (or semantically related) linguistic units appear in similar contexts”*. This assumption allows for a radical empirical approach towards formalizing the meaning of linguistic units. There exist many Distributional Semantic Models (DSM) for representing the meaning of words or complex language expressions. Baroni and Lenci [2010], Turney and Pantel [2010], and Lapesa and Evert [2014] compare different DSMs. More recently, the popular DSMs are based on neural network architectures (Word2Vec [Mikolov et al., 2013b], Glove [Pennington et al., 2014], Skip-Thought [Kiros et al., 2015], InferSent [Conneau et al., 2017], and ELMO [Peters et al., 2018]). DSMs are used in many practical applications. They are also very popular for empirical tasks focused on textual meaning relations. Paraphrasing, Textual Entailment, and Semantic Textual Similarity are often considered evaluation benchmarks for the quality of DSMs.

Within CL and NLP there are many empirical tasks focused on meaning relations at the level of tokens (i.e.: words and multi-word expressions), henceforth **“lexical meaning relations”**. Hill et al. [2015] and Bruni et al. [2014] propose datasets for out-of-context lexical similarity, while Huang et al. [2012] and Levy et al. [2015] propose datasets for context-sensitive lexical similarity. Kremer et al. [2014] present a dataset for the “lexical substitution” task. Hendrickx et al. [2010] propose the task of “relation classification” at the lexical level.

There are many manually created resources for studying and processing lexical meaning relations. These resources include, for example, lists of words with a particular relation, morphological rules for creating a particular relation (e.g.:

“happy” - “unhappy”, “agree”- “disagree”) and knowledge bases such as WordNet [Miller, 1995], WikiData [Vrandečić, 2012], DBPedia [Auer et al., 2008] and ConceptNet [Speer and Havasi, 2012]. The tasks and resources for lexical meaning relations are also relevant for the research on textual meaning relations.

Textual Meaning Relations also have an impact on **Other Areas of CL and NLP** Systems that can successfully process meaning relations can also be used in other tasks in CL and NLP, such as text summarization [Lloret et al., 2008, Harabagiu and Lacatusu, 2010], text simplification [Yimam and Biemann, 2018], plagiarism detection [Barrón-Cedeño et al., 2013], question answering [Harabagiu and Hickl, 2006], and machine translation evaluation [Padó et al., 2009], among others.

1.2 Motivation and Objectives of the Thesis

This thesis arose from an interest in applying linguistic knowledge to the empirical studies of textual meaning relations. My research was motivated by two gaps in the research field:

- a lack of large-scale corpora for research on decomposing textual meaning relations and, as a consequence, a lack of machine learning experiments.
- insufficient resources (annotation guidelines and corpora) and a lack of empirical studies on multiple textual meaning relations.

I address both these gaps in turn by combining theoretical knowledge and empirical, data-driven approaches (human judgments, statistical analysis, and machine learning experiments). First, I bring together paraphrase typology and the task of Paraphrase Identification. Second, I present a joint study on the textual meaning relations of Paraphrasing, Textual Entailment, and Semantic Similarity. In the rest of this section I present in more detail the objectives behind each of the two research directions of my thesis.

1.2.1 Paraphrase Typology and Paraphrase Identification

The work on Paraphrase Typology (PT) uses knowledge from theoretical linguistics to understand the Paraphrasing phenomenon. Paraphrase Identification (PI) is an empirical task that aims to produce systems capable of recognizing paraphrasing in an automatic manner. However, at the time of beginning this dissertation, there had been almost no interaction or intersection between these two areas of

Paraphrasing research. PT research, prior to this dissertation, was mostly theoretical, with very limited practical implications and applications. PI research in the era of deep learning is radically empirical, focused on quantitative performance, with little to no interpretability and theoretical justification. My intuition was that these two research areas are not mutually exclusive, however there was no previous work trying to combine them. My **objectives** in combining PT and PI were twofold:

- Obj1 To use linguistic knowledge and paraphrase typology in order to improve the evaluation and interpretation of automated PI systems.
- Obj2 To empirically validate and quantify the difference between the various linguistic and reason-based phenomena involved in paraphrasing.

1.2.2 Joint Study on Meaning Relations

Meaning relations, such as Paraphrasing, Textual Entailment, and Semantic Similarity, have attracted a lot of attention from the researchers in Computational Linguistics (CL) and Natural Language Processing (NLP). There is a substantial amount of theoretical and empirical research on these meaning relations and many resources, datasets, and automated systems. Traditionally, these relations have been studied in isolation and the transfer of knowledge and resources between them has been very limited. My intuition was that these textual meaning relations can be brought together in a single corpus and compared empirically. My **objectives** in this part were twofold:

- Obj3 To empirically determine the interactions between Paraphrasing, Textual Entailment, Contradiction, and Semantic Similarity in a corpus of multiple textual meaning relations.
- Obj4 To propose and evaluate a novel shared typology of meaning relations. The shared typology would then be used as a conceptual framework for joint research on meaning relations.

1.3 Thesis Development

My research has three separate phases, described in parts I, II, and III of this thesis. First, I explore the basic concepts of Distributional Semantics and the notion of Semantic Similarity at the level of words and short phrases in Part I. Second, I present my empirical research on bringing together Paraphrase Typology and

Paraphrase Identification in Part II. Finally, I describe the setup and results of my joint study on multiple textual meaning relations in Part III.

The order of the chapters follows the chronological order in which the articles were written. At the same time, the order of the chapters follows the logical progression of my dissertation. Each of the articles is self sufficient: it poses its own research questions, presents related work, proposes a methodology, and describes the experimental results. However, there is also a clear thread that connects all the articles. When brought together, the articles tell a coherent story about the linguistic phenomena involved in textual meaning relations and how these phenomena can be used to improve the evaluation and interpretation of automated systems and bring together multiple textual meaning relations.

In the rest of this section I briefly present the main motivation, research questions and findings for each of the three parts and the logical progression of the thesis. I also discuss how each article fits within the more general objectives and how the the different articles interact with each other.

Part I: Lexical Relations and Distributional Semantics

The two articles presented in this part of the thesis serve as an introduction to the research on meaning relations and aim to familiarize the reader with the core concepts and theories used in the whole thesis.

In the article “*Comparing Distributional Semantics Models for identifying groups of semantically related words*” (Chapter 2), I explore the theoretical concepts and the empirical tools within the framework of Distributional Semantics (DS). DS is the most popular framework in contemporary Natural Language Processing (NLP) and Computational Linguistics (CL) and in the research on lexical and textual meaning relations. I experiment with different methodologies for representing the meaning of individual words, different ways to quantitatively compare meaning representations, and different approaches to measuring semantic similarity at the level of words. Lexical similarity is the most “atomic” form of semantic similarity. Many aspects of lexical similarity are also important for semantic similarity at the level of longer pieces of texts.

In the article “*DISCOVer: DIStributIonal approach based on syntactic dependencies for discovering CONstructions*” (Chapter 3), I present a successful data-driven methodology that can compose individual words into short phrases. The methodology is based on Distributional Semantics, lexical semantic similarity, and syntactic similarity between words. Many of the resulting short phrases are novel and have never been observed in the training data, indicating that the system is composing as opposed to memorizing. This article demonstrates the importance of lexical similarity in the context of complex language expressions

and the compositionality of meaning.

Part II: Paraphrase Typology and Paraphrase Identification

The three articles presented in this part of the thesis tell a coherent story of how the theoretical concepts of Paraphrase Typology (PT) research can be validated empirically and, at the same time, can be used to improve the evaluation, interpretation, and, indirectly, the performance of the automated Paraphrase Identification (PI) systems.

In the article “*WARP-Text: a Web-Based Tool for Annotating Relationships between Pairs of Texts.*” (Chapter 4), I describe the workings of a novel web-based annotation interface. The annotation interfaces that existed at the beginning of this thesis were not capable of performing a simultaneous annotation of multiple texts with fine-grained linguistic phenomena. WARP-Text fills this gap in the CL and NLP toolbox and creates new opportunities for researchers.

In the article “*ETPC - a paraphrase identification corpus annotated with extended paraphrase typology and negation*” (Chapter 5), I present the first PI corpus annotated with paraphrase types. I also propose a new extended typology for the paraphrasing relation, enriching the existing work in the area. I analyze the distribution of different linguistic phenomena in the corpus, and I identify general tendencies and potential biases in the data. This corpus-based study is the first large-scale empirical research on Paraphrase Typology within Paraphrase Identification. It contrasts with the pre-existing work in the field, in which researchers typically annotate a small number of examples. This article is also the first work on paraphrase typology that analyzes both positive examples (paraphrases) and negative examples (non-paraphrases). It contrasts with the pre-existing work in the field, which focuses only on the positive examples. The ETPC corpus makes further machine learning based studies on Paraphrase Typology possible.

In the article “*A Qualitative Evaluation Framework for Paraphrase Identification*” (Chapter 6), I perform multiple machine learning experiments on the ETPC corpus. I re-implement 11 different machine learning systems and create an “evaluation framework” - a software package that can quantify and compare the paraphrase types involved in the correct and incorrect prediction of each PI system. I empirically demonstrate that 1) the different paraphrase types are processed differently by the different state-of-the-art automated PI systems; and 2) some paraphrase types are easier or harder for all evaluated systems. Furthermore, I demonstrate that the “qualitative evaluation framework” provides much more information when comparing automated systems and facilitates error analysis.

Part III: A Joint Study of Meaning Relations

The two articles presented in this part of the thesis demonstrate that multiple textual meaning relations can co-exist in the same corpus and can be expressed using the same “atomic” linguistic phenomena.

In the article “*Annotating and analyzing the interactions between meaning relations*”, I present the first corpus to explicitly annotate the meaning relations of Paraphrasing, Textual Entailment, Contradiction, Semantic Similarity, and Textual Specificity. I propose a methodology for corpus creation and annotation that guarantees that all relations are presented with a sufficient frequency. I compare the reliability of the annotation and the inter-annotator agreement across all relations. Finally, I perform an empirical analysis of the frequency, correlation, and overlap between the different meaning relations.

In the article “*Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity*” I propose SHAREl - a shared typology for Paraphrasing, Textual Entailment, Contradiction, Semantic Similarity, and Textual Specificity. I demonstrate that a single typology can successfully be applied to all textual meaning relations. I analyze the distribution of the types across all relations and I outline common tendencies and differences.

1.4 Thesis Outline

This thesis consists of a collection of seven papers, complemented by an introductory and a concluding chapter that provide the necessary context to make the thesis a coherent story. The seven papers are the following:

Part I: Similarity at the Level of Words and Phrases

1. Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2016. Comparing Distributional Semantics Models for identifying groups of semantically related words. *Procesamiento del Lenguaje Natural* vol. 57, pp.: 109-116
2. M. Antònia Martí, Mariona Taulé, Venelin Kovatchev, and Maria Salamó. 2019. DISCOVer: DIStributional approach based on syntactic dependencies for discovering CONstructions. *Corpus Linguistics and Linguistic Theory*

Part II: Paraphrase Typology and Paraphrase Identification

3. Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018. WARP-Text: a Web-Based Tool for Annotating Relationships between Pairs of Texts. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp.: 132-136

4. Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018. ETPC - a paraphrase identification corpus annotated with extended paraphrase typology and negation. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp.: 1384-1392
5. Venelin Kovatchev, M. Antònia Martí, Maria Salamó, and Javier Beltran. 2019. A Qualitative Evaluation Framework for Paraphrase Identification. *Proceedings of the Twelfth Recent Advances in Natural Language Processing Conference*, pp.: 569-579

Part III: Paraphrasing, Textual Entailment, and Semantic Similarity

6. Darina Gold, Venelin Kovatchev, Torsten Zesch. 2019. Annotating and analyzing the interactions between meaning relations *Proceedings of the Thirteenth Language Annotation Workshop*, pp.: 26-36
7. Venelin Kovatchev, Darina Gold, M. Antònia Martí, Maria Salamó, and Torsten Zesch. 2020. Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity. To appear in *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, 2020

All seven papers have been accepted and published in peer reviewed journals or conference proceedings. They are co-authored by both my advisors, with one exception: Paper 6 was written during my research stay at the University of Duisburg-Essen and is co-authored with the Language Technology Group at that university. In all papers, except paper 2, I am listed as the first author. In paper 2, I was responsible for the evaluation section of the article and part of the experimental setup. In paper 6, the first two authors (Darina Gold and myself) contributed equally to the article and the names are in alphabetical order.

The papers reprinted here have been reformatted to make the typography of the thesis consistent, and all of the references and appendices have been integrated in a single bibliography and appendix section at the end. The thesis also includes some additional material, such as annotation guidelines, which were included in the original papers as external web links.

Part I

Similarity at the Level of Words and Phrases

Chapter 2

Comparing Distributional Semantics Models for Identifying Groups of Semantically Related Words

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó
University of Barcelona

Published at
Procesamiento del Lenguaje Natural, 2016
vol. 57, pp.: 109-116

Abstract Distributional Semantic Models (DSM) are growing in popularity in Computational Linguistics. DSM use corpora of language use to automatically induce formal representations of word meaning. This article focuses on one of the applications of DSM: identifying groups of semantically related words. We compare two models for obtaining formal representations: a well known approach (CLUTO) and a more recently introduced one (Word2Vec). We compare the two models with respect to the PoS coherence and the semantic relatedness of the words within the obtained groups. We also proposed a way to improve the results obtained by Word2Vec through corpus preprocessing. The results show that: a) CLUTO outperforms Word2Vec in both criteria for corpora of medium size; b) The preprocessing largely improves the results for Word2Vec with respect to both criteria.

Keywords DSM, Word2Vec, CLUTO, semantic grouping

2.1 Introduction

In recent years, the availability of large corpora and the constantly increasing computational power of the modern computers have led to a growing interest in linguistic approaches that are automated and data-driven [Arppe et al., 2010]. Distributional semantic models (DSM) [Turney and Pantel, 2010, Baroni and Lenci, 2010] and the vector representations (VR) they generate fit very well within this framework: the process of extracting vector representations is mostly automated and the content of the representations is data-driven.

The format of the vector is suitable for carrying out different mathematical manipulations. Vectors can be compared directly through an objective mathematical function. They can also be used as a dataset for various Machine Learning algorithms. VR are more often used on tasks related to lexical similarity and relational similarity [Turney and Pantel, 2010]. In such tasks, the emphasis is on pairwise comparisons between vectors.

This article focuses on another use of the Vector Representations: the grouping of vectors, based on their similarity in the Distributional space. This grouping can be used, among other things, as a methodology for identifying groups of semantically related words. High quality groupings can serve for many purposes: they are a semantic resource on their own, but can also be applied for syntactic disambiguation or pattern identification and generation [Martí et al., 2019], for example.

We compare two different methodologies for obtaining groupings of semantically related words in English - a well known approach (CLUTO) and a more recently introduced one (Word2Vec). The two methodologies are evaluated in terms of the quality of the obtained groups. We consider two criteria: 1) the semantic relatedness between the words in the group; and 2) the PoS coherence of the group. We evaluate the role of the corpus size with both methodologies and in the case of Word2Vec, the role of the linguistic preprocessing (lemmatization and PoS tagging).

The rest of this paper is organized as follows: Section 2.2 presents the general framework and related work. Section 2.3 describes the available data and tools. Section 2.4 presents the experiments and the results obtained. Finally Section 2.5 gives conclusions and identifies directions for future work.

2.2 Related Work

Distributional Semantics Models (DSM) are based on the Distributional Hypothesis, which states that the meaning of a word can be represented in terms of the contexts in which it appears [Harris, 1954, Firth, 1957]. As opposed to seman-

tic approaches based on primitives [Boleda and Erk, 2015], approaches based on distributional semantics can obtain formal representations of word meaning from actual linguistic productions. Additionally, this data-driven process for semantic representation can mostly be automated.

Within the framework of DSM, one of the most common ways to formalize the word meaning is a vector in a multi-dimensional distributional space [Lenci, 2008]. For this purpose, a matrix with size \mathbf{m} by \mathbf{n} is extracted from the corpus, representing the distribution of \mathbf{m} words over \mathbf{n} contexts. The format of a vector allows for direct quantitative comparison between words using the apparatus of linear algebra. At the same time it is a format preferred by many Machine Learning algorithms.

The choice of the matrix is central for the implementation of a particular DSM. Turney and Pantel [2010] suggest a classification of the DSM based on the matrix used. They analyze three different matrices: term-document, word-context, and pair-pattern. The different matrices represent different types of relations in the corpus and the choice of the matrix depends on the goals of the particular research.

Baroni and Lenci [2010] present a different, sophisticated approach for extracting information from the corpus. They organize the information as a third order tensor, with the dimensions representing <‘word’, ‘link’, ‘word’ >. This third order tensor can then be used to generate different matrices, without the need of going back to the original corpus.

In this paper we focus on one of the classical vector representations - the one based on word-context relation. It measures what Turney and Pantel [2010] call “attributonal similarity”. In particular, we are interested in the possibility to group vectors together, based on their relations in the distributional space.

Erk [2012] offers a survey of possible applications of different DSM. She lists clustering as an approach that can be used with vectors, for word sense disambiguation. Moisl [2015] presents a theoretical analysis on the usage of clustering in computational linguistics and identifies key aspects of the mathematical and linguistic argumentation behind it.

Here we analyze and compare two approaches that induce vector representations from a corpus and apply algorithms to identify sets of semantically related words. We are interested in the quality of the obtained groups, as we believe that they can be a useful, empirical, linguistic resource.

Martí et al. [2019] present a methodology named DISCOveR for identifying candidates to be constructions from a corpus. As part of this methodology they use CLUTO [Karypis, 2002] for clustering words based on their vector representations. Their approach uses a word-context matrix where the context is defined by combining a syntactic dependency with a lemma. After all the vectors are extracted, CLUTO is used in order to obtain clusters of semantically related words. Later on these clusters are used to generate a list of the candidates to be construc-

tions.

Mikolov et al. [2013a] suggest a different approach towards extracting vector representations and grouping. Their methodology is based on deep learning and is intended for quick processing of very large corpora. Word2Vec¹, the tool they present, includes an integrated algorithm for grouping words based on proximity in space. The context they use for vector extraction is simple co-occurrence within a specified window of tokens. Originally, they make no use of linguistic preprocessing such as lemmatization, part of speech tagging or syntactic tagging. As part of this paper we evaluate the effect of linguistic preprocessing on the obtained vectors and groups.

2.3 Data and Tools

In this section we present the corpus that we use in the evaluation (Section 2.3.1) and the two methodologies (Section 2.3.2 and Section 2.3.3).

2.3.1 The Corpus

For all of the experiments described in this paper, we use PukWaC [Baroni et al., 2009]². It is a 2 billion word corpus of English, built up from sites in the .uk domain. It is available online and is already preprocessed: XML tags and other non-linguistic information have been removed, it is lemmatized, PoS tagged and syntactically parsed. The PoS tagset is an extended version of the Penn Treebank tagset. The syntactic dependencies follow the CONLL-2008 shared task format.

2.3.2 Grouping with CLUTO

DISCOVer [Martí et al., 2019] is a methodology for identifying candidates to be construction from a corpus. It uses vector representations, extracted from a corpus. CLUTO [Karypis, 2002] is used on these representations in order to obtain clusters of semantically related words. CLUTO is a software package for clustering low and high dimensional data sets and for analysis of the characteristics of the various clusters. CLUTO provides three different classes of clustering algorithms, based on partitional, agglomerative and graph-partitioning paradigms. It computes clustering solution based on one of the different approaches.

For this article, we are interested only in the first three steps of the DISCOVer process. Step 1 is the linguistic preprocessing of the corpus. The raw text is cleared from non-linguistic data, it is PoS tagged and syntactically parsed. In

¹Available at: <https://code.google.com/archive/p/word2vec/>

²Available at: <http://wacky.sslmit.unibo.it>

Step 2, the DSM matrix is constructed. The rows of the matrix correspond to lemmas and the columns correspond to contexts. Contexts in this approach are defined as a triple of syntactic relation, direction of the relation and lemma in [direction:relation:lemma] format³. This matrix is used to generate vector representations for the 10,000 most frequent words in the corpus. Next, Step 3 uses CLUTO to create clusters of semantically related lemmas from the DSM matrix and the corresponding vectors. The clusters are created based on shared contexts.

Martí et al. [2019] start from a raw, unprocessed corpus and in Step 1 they clear the corpus and tag it with the linguistic data relevant to the matrix extraction. The format they use is shown in Table 2.1.

Table 2.1 Diana-Araknion Format

Token	sanitarios
Lemma	sanitario
PoS	NCMP
Short PoS	n
Sent ID	000
Token ID	0
Dep ID	2
Dep Type	subj

The original DISCOVeR experiment is done with the Diana-Araknion corpus of Spanish. For the purpose of this article, we replicated the process for English, using the PukWaC corpus. For step 1 we had to make sure that our preprocessing is equivalent to the one of Diana-Araknion. The corpus PukWaC is already preprocessed and the format is similar to the one of Diana-Araknion. However, in order to make it fully compatible, we had to make several modifications of the format and linguistic decisions. Regarding the format, we removed any remaining XML tags, enumerated the sentences in the corpus, and generated “short PoS”⁴. From the linguistic side, we had to decide whether all PoS and Dependencies were relevant for the vector generation or some of them could be merged together or even discarded in order to optimize and speed up the process.

The process of generating vectors and clusters is based on analyzing the contexts where each word appears in. A word is identified by its lemma and its PoS

³For example, from the sentence “El barbero afeita la larga barba de Jaime”, three different contexts of the noun lemma barba are generated: [<:doj:afeitar_v] , [>:mod:largo_a] and [>:de_sp:pn_n]. The example is from Martí et al. [2019]

⁴short PoS is a one letter tag representing the generic PoS tag of the lemma. In this experiment, short PoS is the first letter of the full PoS

tag. However, in the PukWac tagset there are many PoS tags which specify not only the PoS of the token, but also contain information about other grammatical features, such as person, number, and tense. If these tags are kept unchanged, a separate vector will be generated for different forms of the same word, based on different PoS tag. To avoid this problem and to generate only one vector for all of the different word forms, we have decided to merge certain PoS tags under one category.

We decided to simplify the POS tagset further. It is a common practice in DSM to focus the experiment on the relations between content words. Function words and punctuation are usually not considered relevant contexts. Because of that, we have put them under the common tag “other”. All of the changes on the PoS tagset are summarized in Table 2.2.

Table 2.2 PoS tagset modifications

Tag	Original tag	Description
J	JJ JJR JJS	Adjective
M	MD	Modal verb
N	NN NNS	Noun (common)
NP	NP NPS	Noun (personal)
R	RB RBR RBS RP	Adverb
S	IN	Preposition
V	VB* VH* VV*	Verb (all)
O	CC CD DT PDT EX FW LS POS PP* SYM TO UH W* punctuation	Rest

The list of syntactic dependencies in PukWaC is also not fully relevant to the task of vector generation. While the unnecessary PoS tags may lead to multiple vectors for the same word, unnecessary dependencies generate additional contexts, increasing the dimensionality of the vectors and leading to a more complicated computational process. Therefore the modification of the dependencies is mostly related to the optimization of the computational process. After analyzing the tagset, we have decided to merge the **OBJ** and **IOBJ** tags due to some inconsistencies of their usage. We have also decided to discard the following relations: **CC** (conjunction), **CLF** (be/have in a complex tense), **COORD** (coordination), **DEP** (unclassified relation), **EXP** (experiencer in few very specific cases),

P (punctuation), **PRN** (parenthetical), **PRT** (particle), **ROOT** (root clause). The final list of dependencies is shown in Table 2.3.

Table 2.3 Syntactic Dependencies

Dependency	Description
ADV	Unclassified adv
AMOD	Modifier of adj or adv
LGS	Logical subj
NMOD	Modifier of nom
OBJ	Direct or indirect obj
PMOD	Preposition
PRD	Predicative compl
SBJ	Subject
VC	Verb chain
VMOD	Modifier of verb
empty	No dependency

Once the corpus is preprocessed, the process of matrix extraction is mostly automated. For the matrix, we have only generated vectors for words that appear at least 5 times in the corpus. Out of them we have used only the vectors of the 10,000 most frequent words for the clustering process.

For the clustering process, we configure CLUTO to use direct clustering, based on the H2 criterion function, with 25 features per cluster. We have ran the clusterization multiple times, ranging from 100 to 1,000 clusters. We then used CLUTO's H2 metric to determine the optimal number of clusters, which has been 800 for all of the experiments.

2.3.3 Grouping with Word2Vec

Word2Vec is based on the methodology proposed by Mikolov et al. [2013a]. It takes a raw corpus and a set of parameters and generates vectors and groups. The algorithm of Word2Vec is based on a two layer neural network that are trained to reconstruct linguistic context of words. Word2Vec includes two different algorithms - Continuous Bag-of-Words (CBOW) and Skip-Gram. CBOW learns representations based on the context as a whole - all of the words that co-occur with the target word in a specific window. Skip-Gram learns representation based on each single other word within a specified window. When using Word2Vec usually the emphasis is put on the choice of the parameters for the algorithm, and not on the specifications of corpus. However, we consider that the specifications of

the corpus (size and linguistic preprocessing) can largely affect the quality of the obtained results.

By default Word2Vec works with a raw corpus. Neither of the two models makes explicit use of morpho-syntactic information. However, by modifying the corpus, some morphological information can be used implicitly. If the token is replaced by its corresponding lemma or by the lemma and part of speech tag in a “lemma_pos” format, the resulting vectors would be different: using the lemma would generate only one vector for the word as opposed to separate vector for every word form; using PoS can make a distinction between homonyms with same spelling and different PoS. As part of our work we wanted to examine how linguistic preprocessing can affect the quality of the vectors. For that reason we created three separate corpus samples - one raw corpus, one where each token was replaced by its lemma, and one where each token was replaced by “lemma_pos”. We generated vectors separately for each of the corpora. Unfortunately, there was no trivial way to introduce syntactic information implicitly in the models of Word2Vec.

2.4 Experiments

In this section we present the setup for the different experiments (Section 2.4.1), the evaluation criteria (Section 2.4.2), and the obtained results (Section 2.4.3).

2.4.1 Setup

We carried out a total of 15 experiments - 3 experiments using CLUTO and 12 experiments using Word2Vec. For the experiments with CLUTO, the only variation between the experiments was the size of the corpus: 4M tokens, 20M tokens, and 40M tokens⁵. In all the experiments we used the preprocessing described at Section 2.3.2, we generated vectors for the 10,000 most frequent words and we split them into 800 clusters. For the experiments with Word2Vec, we changed three parameters of the experiments: (1) the algorithm (CBOW and Skip-Gram), (2) the linguistic preprocessing of the corpus (raw, lemma, lemma and PoS), and (3) the size of the corpus (4M, 20M, and 40M). We carried out 9 experiments with CBOW (all size and preprocessing combinations) and 3 experiments with Skip-Gram (the three variants of the 40M corpus). Mikolov et al. [2013a] identify two important parameters to be set up when using Word2Vec: the vector size and the window size. For the window size, we used 8, which is the recommended

⁵The 40M corpus contains in itself the 20M corpus. The 20M corpus contains in itself the 4M corpus. The same corpora has been used for the experiments with both CLUTO and with Word2Vec.

value. For the vector size, Mikolov et al. [2013a] show that increasing vector size from 100 to 300 leads to significant improvement of the results, however further increase does not have big impact. For that reason we have chosen vector size of 400, which is above the recommended minimum. For the number of groups we used 800: the same number that was determined optimal for CLUTO. For the number of lemmas, we used the 10,000 most frequent ones, the same setup as with CLUTO.

2.4.2 Evaluation

The two methodologies and all of the different setups are evaluated based on the quality of the obtained groups. We consider two criteria: 1) The semantic relatedness between the words in each group; and 2) The PoS coherence of the groups. The PoS coherence is a secondary criterion which should be considered in addition to the semantic relatedness. Our intuition is that groups that are semantically related and PoS coherent are a better resource than groups that are only semantically related. For evaluating the semantic relations of the words in the groups, we present two methodologies - an automated method based on WordNet distances and a manual evaluation done by experts on a subset of the groups in each experiment. The PoS coherence is calculated automatically.

There is no universal widely accepted criteria for determining the semantic relations between two words. Two of the most common approaches are calculating WordNet distances and expert intuitions. We used both when evaluating the quality of the obtained groups.

For the WordNet similarity evaluation, we use the WordNet interface built in NLTK [Bird et al., 2009]. We calculate the Leacock-Chodorow Similarity⁶ between each two words⁷ in every group. We then sum all the obtained scores and divide them by the number of pairs to obtain average WordNet similarity for each method.

For the expert evaluation, we selected a subset of groups, generated in each experiment⁸. Three experts were asked to rate each group on a scale from 1 (unrelated) to 4 (strongly related)⁹. We calculate the average between all of the scores

⁶It calculates word similarity, based on the shortest path that connects the senses and the maximum depth of the taxonomy in which the senses occur.

⁷The calculation is based on the first sense of every word

⁸We selected the groups based on a word they contain - three verb groups (the ones that contain “say”, “see”, “want”), 3 noun groups (“person”, “year”, “hand”), 1 adjective group (“good”), 1 adverb group (“well”). All of the selected words are among the 100 most commonly used words of English.)

⁹In the detailed description of the scale given to the experts: 1 corresponds to “no semantic relation”; 2 corresponds to “semantic relation between some words (less than 50% of the group); 3 corresponds to “semantic relation between most of the words in the corpus (more than 50%), but

they gave on the groups of each experiment.

We define PoS coherence as the percent of words that belong to the most common PoS tag in each group. In order to calculate it, all obtained groups are automatically PoS tagged¹⁰. Then for each group, we count the percent of words that belong to each PoS and identify the most common tag.

2.4.3 Results

Table 2.4 shows the WordNet similarity evaluation. The average similarity score obtained by CLUTO is higher than the score obtained by Word2Vec (0.81-0.96 against 0.67-0.81). This indicates that the distances between the words in the CLUTO groups are shorter and the semantic relations are stronger. Increasing the corpus size improves the results for both CLUTO and Word2Vec. Preprocessing (specifically PoS tagging) improves the obtained results for all of the Word2Vec experiments. The groups obtained using Skip-Gram get lower scores in the evaluation compared with the groups obtained using CBOW.

Table 2.4 Wordnet Similarity

Methodology	Corpus	Similarity
W2V-CBOW	4M (raw)	0.67
W2V-CBOW	4M (lemma)	0.67
W2V-CBOW	4M (pos)	0.72
W2V-CBOW	20M (raw)	0.74
W2V-CBOW	20M (lemma)	0.75
W2V-CBOW	20M (pos)	0.77
W2V-CBOW	40M (raw)	0.77
W2V-CBOW	40M (lemma)	0.78
W2V-CBOW	40M (pos)	0.81
W2V-SG	40M (raw)	0.69
W2V-SG	40M (lemma)	0.73
W2V-SG	40M (pos)	0.74
CLUTO	4M	0.81
CLUTO	20M	0.92
CLUTO	40M	0.96

with multiple unrelated words”; 4 corresponds to “semantic relation between most of the words in the corpus, without many unrelated words”

¹⁰We use only the short PoS tag for this evaluation

Table 2.5 shows the results from the expert evaluation of the semantic relations in the groups. The data is similar to the results with WordNet distances. The groups obtained by CLUTO show higher degree of semantic relatedness (2.8-3.4) compared to the groups obtained by Word2Vec (1.6-2.7). The CLUTO groups at 20M and 40M obtain average above 3, meaning that the experts consider all of the groups to be strongly related. For the experiments with Word2Vec, linguistic preprocessing improves the results, especially at bigger corpus size (2.5 against 1.8 for 20M and 2.7 against 2 for 40M). The groups obtained using Skip-Gram algorithm are rated lower than the groups obtained using CBOW. The preprocessed corpus obtains better groups, but the difference is smaller than the one observed with CBOW.

Table 2.5 Expert evaluation

Methodology	Corpus	Score
W2V-CBOW	4M (raw)	1.6
W2V-CBOW	4M (lemma)	1.4
W2V-CBOW	4M (pos)	1.8
W2V-CBOW	20M (raw)	1.8
W2V-CBOW	20M (lemma)	2.4
W2V-CBOW	20M (pos)	2.5
W2V-CBOW	40M (raw)	2
W2V-CBOW	40M (lemma)	2.1
W2V-CBOW	40M (pos)	2.7
W2V-SG	40M (raw)	1.7
W2V-SG	40M (lemma)	1.8
W2V-SG	40M (pos)	2
CLUTO	4M	2.8
CLUTO	20M	3.2
CLUTO	40M	3.4

Table 2.6 shows the results for the PoS coherence evaluation. The data shows that the groups obtained from CLUTO are more PoS coherent, compared with the groups obtained by Word2Vec (90-98% against 69-81%). For the corpora of size 20M and above, the groups obtained by CLUTO have almost 100% PoS coherence, meaning that all of the lemmas belong to the same PoS. Both CLUTO and Word2Vec show improved results with the increase of corpus size. The results with Word2Vec indicate that corpus preprocessing largely improves the obtained results (69%-73% against 75%-81%). In fact, for this experiment the corpus preprocessing have bigger impact than the corpus size: a preprocessed corpus with

a size of 4M generates more PoS coherent groups than raw 40M corpus (74-75% against 73%). The experiments with Skip-Gram obtain similar results for raw corpus. For Skip-Gram the preprocessed corpus also obtains better overall results, however lemmatized corpus obtains better results than the PoS tagged corpus.

Table 2.6 PoS coherence

Methodology	Corpus	PoS
W2V-CBOW	4M (raw)	69%
W2V-CBOW	4M (lemma)	74%
W2V-CBOW	4M (pos)	75%
W2V-CBOW	20M (raw)	72%
W2V-CBOW	20M (lemma)	77%
W2V-CBOW	20M (pos)	80%
W2V-CBOW	40M (raw)	73%
W2V-CBOW	40M (lemma)	78%
W2V-CBOW	40M (pos)	81%
W2V-SG	40M (raw)	73%
W2V-SG	40M (lemma)	80 %
W2V-SG	40M (pos)	77%
CLUTO	4M	90%
CLUTO	20M	97%
CLUTO	40M	98%

Overall, all three evaluations identify similar patterns in the obtained clusters: (1) the groups obtained by CLUTO perform better than the groups obtained by Word2Vec; (2) Increasing the corpus size improves the quality of the results for both methodologies. This is true for semantic relatedness as well as for PoS coherence. The tendency to obtain more PoS coherent groups justifies the usage of PoS coherence as evaluation criteria; (3) Linguistic preprocessing improves the quality of the groups obtained by Word2Vec (with both algorithms).

2.5 Conclusions and Future Work

This article compares two methodologies for identifying groups of semantically related words based on Distributional Semantic Models and vector representations. We applied the methodologies to a corpus of English and compared the quality of the obtained groups in terms of semantic relatedness and PoS coherence. We also analyzed the role of different factors, such as corpus size and linguistic preprocessing.

In the comparison of the two methodologies, the results show that CLUTO outperforms Word2Vec with respect to grouping, using corpora of medium size (20M - 40M). However, the quality of the results does depend on the size of the corpus. At 40M CLUTO already obtains very high quality results (98% PoS coherence and 3.4/4 strength of semantic relationships in the evaluation of the experts) so further increase of the corpus is not likely to show large improvement. On the contrary at 40M Word2Vec still has room for improvement and we expect to narrow the difference between the two methodologies using much larger corpora (1B and above).

In the comparison of the different preprocessing corpora (i.e., raw, lemma, and PoS) in Word2Vec, the results show that lemmatization and PoS tagging largely improve the quality of the groups in both CBOW and Skip-Gram algorithms. This observation is consistent throughout all of the experiments and with respect to all of the evaluation criteria.

The presented comparison opens several lines of future research. First, the evaluation can be extended to bigger corpora, bigger number of vectors, and other languages. Second, the information provided and the suggested criteria for evaluation can be applied to other approaches to DSM and grouping. Finally, the different methodologies and preprocessing options can be evaluated in as part of more complex systems.

Chapter 3

DISCOVer: DIStributional Approach Based on Syntactic Dependencies for Discovering COstructions

M. Antònia Martí, Mariona Taulé, Venelin Kovatchev, and Maria Salamó
University of Barcelona

Published at
Corpus Linguistics and Linguistic Theory, 2019

Abstract One of the goals in Cognitive Linguistics is the automatic identification and analysis of constructions, since they are fundamental linguistic units for understanding language. This article presents DISCOVer, an unsupervised methodology for the automatic discovery of lexico-syntactic patterns that can be considered as candidates for constructions. This methodology follows a distributional semantic approach. Concretely, it is based on our proposed pattern-construction hypothesis: those contexts that are relevant to the definition of a cluster of semantically related words tend to be (part of) lexico-syntactic constructions. Our proposal uses Distributional Semantic Models (DSM) for modeling the context taking into account syntactic dependencies. After a clustering process, we linked all those clusters with strong relationships and we use them as a source of information for deriving lexico-syntactic patterns, obtaining a total number of 220,732 candidates from a 100 million token corpus of Spanish. We evaluated the patterns obtained intrinsically, applying statistical association measures and they

were also evaluated qualitatively by experts. Our results were superior to the baseline in both quality and quantity in all cases. While our experiments have been carried out using a Spanish corpus, this methodology is language independent and only requires a large corpus annotated with the parts of speech and dependencies to be applied.

Keywords Constructions, Semantics, Distributional Semantic Models

3.1 Introduction

In cognitive models of language [Croft and Cruse, 2004], a construction is a conventional symbolic unit that involves a pairing of form and meaning that occurs with a certain frequency. Constructions can be of different types depending on their complexity –morphemes, words, compound words, collocates, idioms and more schematic patterns [Goldberg, 1995, 2006]. Cognitive Linguistics assumes the hypothesis that these constructions are learned from usage and stored in the human memory [Tomasello, 2000], where they are accessed during both the production and comprehension of language. Therefore, constructions are fundamental linguistic units for inferring the structure of language and their identification is crucial for understanding language.

Although a broad range of these linguistic structures have been subjected to linguistic analysis [Nunberg et al., 1994, Wray and Perkins, 2000, Fillmore et al., 2012], we assume that there exist a huge number of constructions that are as yet undiscovered. There are very different approaches to the task of identifying and discovering them, depending on the type of construction we are looking for or dealing with. This fact allows for the use of a wide range of methods and approaches aiming at the treatment of this kind of linguistic units. We distinguish between two different approaches, those that have been guided by previously gathered empirical data¹, and those approaches that apply methods oriented to discovering new constructions from scratch (see Section 3.2).

Following the latter approach, this article presents DISCOVer, an unsupervised methodology for the automatic identification and extraction of lexico-syntactic patterns that are candidates for consideration as constructions (see Section 3.3). It is based on the Harris distributional hypothesis [Harris, 1954]², which states that semantically related words (or other linguistic units) will share the same context.³

¹See Goldberg [1995].

²This idea was also developed by Firth [1957] and Wittgenstein [1953].

³Related hypotheses, such as the extended distributional hypotheses, which states that “patterns that co-occur with similar pairs tend to have similar meanings” [Lin and Pantel, 2001], and latent relation hypotheses [Turney, 2008], which states that “pairs of words that co-occur in similar

We propose the pattern-construction hypothesis, which states that those contexts that are relevant to the definition of a cluster of semantically related words tend to be (part of) lexico-syntactic constructions. What is new in our hypothesis is that we consider all the contexts that are relevant to define a cluster of semantically related words to be part of a construction. In these approaches, Distributional Space Models (DSMs) are used to represent the semantics of words on the basis of the contexts they share. This is in line with the idea proposed by Landauer et al. [2007], who states that DSMs are plausible models of some aspects of human cognition [Baroni and Lenci, 2010].

In our methodology, the DSM consists of a frequency lemma-context matrix, in which the context is modeled taking into account syntactic dependency relations. Then, we build up clusters of semantically related words that share the same context and link them using the information present in their contexts. We automatically calculate a threshold in order to determine which clusters are more strongly related. We filter out those related clusters that do not reach the determined threshold and derive lexico-syntactic patterns that are candidates to be considered as constructions. These candidates are tuples involving two lexical items (lemmas) related both by a dependency direction and a dependency label (examples in (1))⁴:

1. a. accidente_n [>:mod:mortal_a]⁵

b. aterrizar_v [>:dojb:avioneta_n]⁶

The tuples correspond to different kinds of linguistic constructions, ranging from collocates (1a) to (parts of) verbal argument structures (1b). All the lexico-syntactic patterns obtained are instances of one of the syntactic dependencies present in the source corpus. We applied this methodology to the Diana-Araknion corpus, obtaining 220,732 patterns that are good candidates to be constructions⁷.

Finally, we evaluated the quality of these patterns in two ways: applying statistical association measures and by manual revision by human experts. The results show significant improvement with respect to several baselines (see Section 3.4).

Although this method has been applied to the obtention of Spanish constructions, it is language independent and only requires a large corpus annotated with part-of-speech (POS) and syntactic dependencies.

patterns tend to have similar relations” survived in Turney and Pantel [2010] have also influenced this work.

⁴The symbols ‘<’ and ‘>’ indicate the dependency direction and *mod*, *subj* and *dojb* are dependency labels (where *mod* stands for modifier, and *subj* and *dojb* stand for subject and direct object respectively).

⁵accidente_n[>:mod:mortal_a]

⁶to_land[>:dojb:small_plane_a]

⁷All patterns obtained are available at <http://clic.ub.edu/corpus/>

The article is structured as follows. After presenting the related work in Section 3.2, the methodology applied for obtaining the constructions is described in Section 3.3. The evaluation of our methodology is presented in Section 3.4 and, finally, the conclusions and future work are drawn in Section 3.5.

3.2 Related Work

The boundaries of what a construction is are fuzzy: constructions can be lexical, syntactic, lexico-syntactic, morphological and can combine different levels of abstraction from concrete forms to abstract categories, including the possibility of using variables, so they cover a wide range of linguistic constructs. For more examples, see Goldberg [2013].

As a consequence, there is no one accepted typology of this kind of linguistic units [Wray and Perkins, 2000]. There is, therefore, a broad field of research in which to explore the characteristics, the limits and the properties of constructions. In this context, an important task is to acquire the maximum amount of empirically grounded data concerning this kind of units. Thus, when approaching the task of attempting to identify the possible constructions that constitute the core of languages, it is difficult to decide what to look at or where to start [Sag et al., 2002]. For this reason, constructions are a challenge for Linguistics and Natural Language Processing (NLP), where we find statistical and symbolic approaches to deal with them.

Several linguistic traditions converge when we are trying to define the diverse form that a construction can take. From one side, there is an (almost total) overlapping between constructions and argument structure [Goldberg, 1995] and diatheses alternations [Levin, 1993]; from another side, in the lexicographic tradition, constructions also overlap with idioms and collocates. In the field of Computational Linguistics, these linguistic units tend to be grouped under the umbrella term MultiWord Expressions (MWE). Baldwin and Kim [2010] define MWE as those lexical items that are decomposable into multiple lexemes and present idiomatic behaviour at some level of linguistic analysis, as a consequence they should be considered as a unit at some level of computational processing. Also in the Computational Linguistics field, Stefanowitsch and Gries [2003] propose the term “collostruction” to refer to the wide range of complex linguistic units as defined in theoretical proposals of Cognitive Grammar. In our approach we consider as constructions those syntactic units consisting of two or more lexical items with internal semantic coherence. These constructions are compositional and appear with a frequency higher than expected.

From the NLP perspective, most approaches for dealing with constructions tend to apply methods that use previously defined empirical knowledge to find

instances and variants of specific types of constructions in corpora. This approach allows us to obtain preidentified units and their variations at different degrees of complexity, but does not allow for the identification of as yet unidentified constructions. In order to discover new knowledge, we need an open and flexible method that give us usable and interpretable results. We organised this overview taking into consideration those approaches that try to find or discover constructions.

A frequent approach to gathering empirical data about constructions using NLP techniques is to look for well-known, highly conventionalized and previously defined constructions (see the works of Hwang et al. [2010], Muischnek and Sajkan [2009], Kesselmeier et al. [2009], O'Donnell and Ellis [2010], Duffield et al. [2010]).

Very tied to Construction Grammar theory and in the framework of the methodologies based on statistical metrics, it is worth noting the works of Stefanowitsch and Gries [2003], Stefanowitsch and Gries [2008], and Gries et al. [2005]. Their research always focuses on specific types of constructions, on the analysis of their variants and on the degree of entrenchment between their elements. Gries and Ellis [2015] summarize different statistical measures applied to the analysis of constructions and evaluate their linguistic interpretation and impact.

From the perspective of methods oriented to the discovery of new constructions, we should distinguish between those approaches that include some kind of linguistic filtering of the type of constructions to be dealt with and those that do not apply any kind of restriction. All these methods are strongly grounded on statistical measures: in Evert [2008] and Pecina [2010] there is an exhaustive summary and criticism of statistical measures that calculate the degree of association between words.⁸

Looking for ways to identify potential collocations in corpora using statistical measures, Bartsch [2004] explores certain types of collocations involving verbs of verbal communication. Her approach is semiautomatic and involves a manual revision of the results. We also highlight the work of Pecina [2010], based on fully statistical methods. However, supervised machine learning requires annotated data, which creates a bottleneck in the absence of large corpora annotated for collocation extraction. A solution to this problem is presented by Dubremetz and Nivre [2014] who propose the use of the MWEtoolkit [Ramisch et al., 2010] to automatically extract candidates that fit a certain POS pattern. See also the work of Forsberg et al. [2014], Farahmand and Martins [2014], Tutubalina [2015].

From a different perspective, based on the calculation of n -grams, we also consider the results of the StringNet project [Wible and Tsao, 2010], a knowledge

⁸The works referred to this section use the term *collocate* in a very weak sense, roughly equivalent to what is known as MWE in NLP.

base (KB) which contains candidates to be constructions. In this case, no filters are applied to the lexico-syntactic patterns obtained. As a result, StringNet is a lexicogrammatical KB automatically extracted from the British National Corpus (BNC)⁹ consisting of a massive archive of hybrid n -grams of co-occurring combinations of POS tags, lexemes and specific word forms.

We also want to highlight the approaches that use syntactic information for obtaining constructions, such as the work of Zuidema [2006], Sangati and van Cranenburgh [2015], based on the framework of Tree Substitution Grammar (TSG).

Harris distributional hypothesis has a great acceptance in the treatment of linguistic semantics to overcome traditional symbolic representations. Relying on this hypothesis, Gamallo et al. [2005] developed an unsupervised strategy to acquire syntactico-semantic restrictions for nouns, verbs and adjectives from partially parsed corpora. Although the resulting data could be used for deriving lexico-syntactic patterns their objective was to capture semantic generalizations, both for the predicates and their arguments.

Currently, there is an increasing interest in the use of distributional models for representing semantics, such as DSMs [Turney and Pantel, 2010, Baroni, 2013] or word embeddings [Mikolov et al., 2013c]. These models derive word-representations in an unsupervised way from very large corpora. All of them rely on co-occurrence patterns but differ in the way they reduce dimensionality. As pointed out in Murphy et al. [2012], the representations they derive from corpora are lacking in cognitive plausibility, with exceptions such as those defined in Baroni et al. [2010]. Our proposal shares with these authors the same semantic approach (distributional hypothesis), because we consider that these models are a good option in which to frame our methodology. In concrete, we used DSMs because they are highly linguistically interpretable and allow us to modelize the context, a key point in our methodology.

DSMs have been applied successfully in linguistic research [Shutova et al., 2010], in different NLP tasks and applications [Baroni and Lenci, 2010] and, especially, in tasks related with measuring different kinds of semantic similarity between words [Turney and Pantel, 2010]. Like us, Shutova et al. [2017] use distributional clustering techniques, though they use DSMs to investigate how to find metaphorical expressions. Recently, DSMs have been extended to phrases and sentences by means of composition operations deriving meaning representations for phrases and sentences from their parts (see Baroni [2013] and Mitchell and Lapata [2010] for an overview). Nevertheless, DSMs have rarely focused on the discovery of constructions. In this line, it is worth noting the papers presented in the shared task of the Workshop on Distributional Semantics and Compositionality [Biemann and Giesbrecht, 2011]. This workshop focused on the extraction of

⁹www.natcorp.ox.ac.uk

non-compositional phrases from large corpora by applying distributional models that assign a graded compositional score to a phrase. This score denotes the extent to which compositionality holds for a given expression. The participants applied a variety of approaches that can be classified into lexical association measures and Word Space Models. It is also worth noting that approaches based on Word Space Models performed slightly better than methods relying solely on statistical association measures.

In the next section, we describe in depth the DISCOVer methodology that we developed to discover lexico-syntactic constructions.

3.3 Methodology for Discovering Constructions

Following a distributional semantic approach, we developed an unsupervised bottom-up method for obtaining the lexico-syntactic patterns that can be considered candidates for constructions. This method uses a medium-sized corpus (100 million tokens) to obtain the distributional properties of words and to establish similarity relations among them from their contexts. The representation of the contexts is based on syntactic dependencies.

Figure 3.1 depicts the five main steps involved in obtaining the lexico-syntactic patterns, the processes involved, and the input and output of each process. Briefly, the first step is the linguistic processing of the Diana-Arakhion corpus (See Section 3.3.2). In the next step, a DSM matrix is constructed with the frequencies of the lemmas in each one of the contexts (see Section 3.3.3). Step 3 focuses on clustering semantically related lemmas, that is, those lemmas that share a set of contexts (see Section 3.3.4). In the fourth step, we applied a generalization process by linking all clusters taking into account the information contained in the contexts and then filtering only those links that maintain the strongest relationships (See Section 3.3.5). Finally, we generate the lexico-syntactic patterns to be considered as candidates to be constructions from the related clusters selected in the previous step (See Section 3.3.6).

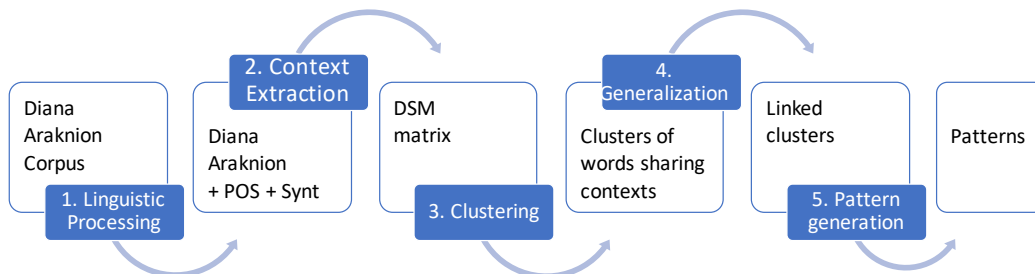


Figure 3.1: Main steps in DISCOVer methodology

3.3.1 Description of the Task

Our methodology is based on the pattern-construction hypothesis, which states that those contexts that are relevant to the definition of a cluster of semantically related words tend to be (part of) lexico-syntactic constructions. In our experiments, “lexico-syntactic constructions” are patterns in the form of [*lemma*, *dependency_direction* (*dep_dir*), *dependency_label* (*dep_lab*), *context_lemma*] (for instance, [*despeinar_v*, >: *dobj*, *cabellera_n*]¹⁰). *Dependency_label* is a type of syntactic relation between *lemma* and *context_lemma*, while *dependency_direction* is the direction of the *dependency_label*. To be considered candidates to be constructions patterns must have the following properties:

- *Syntactic-semantic coherence*: We expect the two lemmas in each pattern candidate to be syntactically and semantically related.
- *Generalizability*: The patterns can be generalized and/or derived from other patterns through generalization.

Based on these properties of constructions and the initial pattern-construction hypothesis, the main aims of the DISCOVer methodology are the following:

1. To identify the contexts that are relevant for the definition of a cluster of semantically related words. Each of these contexts is part of a pattern candidate to be construction attested in the corpus (henceforth Attested-Patterns).
2. To use the previous contexts in a generalization process in order to identify unseen, but possible candidates to be constructions (henceforth Unattested-Patterns).

As a result we obtain two sets of qualitatively different patterns that are candidates to be constructions: attested and unattested patterns. We then proceed to evaluate the internal syntactic-semantic coherence of these patterns.

3.3.2 The Corpus

As shown in Figure 3.1, corpus creation is the first step in the process of obtaining lexico-syntactic patterns. Specifically, we built the Diana-Araknion¹¹ corpus, a Spanish corpus which consists of approximately 100 million tokens¹² (corresponding to 3 million sentences) gathered mainly from the Spanish Wikipedia

¹⁰[*to_tussle_v*, >: *dobj*, *one's_hair_n*]

¹¹ All corpora are available at <http://clitc.ub.edu/corpus/> or per-request

¹²Concretely, the Diana-Araknion has 93,987,098 tokens and 1,321,174 types.

(2009), literary works and texts from Spanish parliamentary discussions, news reports, news agency documents, and Spanish Royal Family speeches.

The corpus was automatically tokenized and linguistically processed with POS and lemma tagging, and syntactic dependency parsing. We used the Spanish analyzers available in the Freeling¹³ open source language-processing library [Padró and Stanilovsky, 2012].

For the purpose of evaluation, we built Diana-Arakhion++, a new corpus gathered from web-pages in Spanish. It includes Spanish Wikipedia (2015), articles from online newspapers, speeches from the European Parliament, university articles and sites from the Spanish webspace. This corpus was automatically tokenized and POS tagged and consists of 600M tokens.

3.3.3 Matrix

To generate the frequency matrix (see Step 2 in Figure 3.1), we used only the 15,000 most frequent lemmas extracted from the Diana-Arakhion corpus including nouns (*N*), verbs (*V*), adjectives (*A*) and adverbs (*R*). We modeled the context in which the words occur giving rise to a *lemma-dep* matrix. This matrix corresponds to the type of *word-context* matrix defined in Turney and Pantel [2010] and in Baroni and Lenci [2010]. In the *lemma-dep* matrix, the context is based on parsed texts in which both dependency directions and dependency labels are taken into account. Each context is a triple of [*dependency_direction*, *dependency_label*, *context_lemma_POS*].

In what follows, we introduce how this lemma-context matrix is formally represented (see Section 3.3.3.1) and then we describe the matrix in more detail (see Section 3.3.3.2).

3.3.3.1 Formalization of the Lemma-Context Matrix

Our DSM consists of a lemma-context PPMI matrix X with n_r rows and n_c columns. Note that each row vector i corresponds to a lemma, each column j corresponds to a co-occurrence context, and each cell in X has a numerical weighted value, x_{ij} . This weighted value is the result of applying Positive Pointwise Mutual Information (PPMI) [Niwa and Nitta, 1994] to a lemma-context frequency matrix F with size $n_r \times n_c$. Each element in this matrix, f_{ij} , is computed as the number of occurrences of lemma i in context j in the whole corpus. Lapesa and Evert [2014] perform a large-scale evaluation of different co-occurrence DSM models over various tasks. They show that term weighting through association scores significantly improves the performance of the DSM model.

¹³<http://nlp.lsi.upc.edu/freeling>.

3.3.3.2 Lemma-Dep Matrix

The matrix proposed in this work is a lemma-context matrix, hereafter *lemma-dep* matrix, based on syntactic dependencies¹⁴. In this matrix, the context j of a lemma i is a context word k (*context_lemma*) directly related by a dependency direction (*dep_dir*) and a dependency label (*dep_lab*) to the lemma i . The words of the lemma i belong to the following POS: N , V , A and R . Each lemma is assigned its corresponding POS. Therefore, in the matrix, context j contains three elements as defined in 3.1:

$$context = [dep_dir : dep_lab : context_lemma] \quad (3.1)$$

where:

- *dep_dir*: has two possible values ‘<’ or ‘>’, indicating the direction of the dependency.
- *dep_lab*: indicates the dependency label of the lemma i and context_lemma k . The possible values are {*subj*, *dobj*, *iobj*, *creg*, *cpred*, *atr*, *cc*, *cag*, *spec*, *sp* and *mod*}. In the case of dependencies between a preposition and a noun, adjective or verb, the dependency label is labeled by the same preposition and its corresponding *dep_lab*, that is, *dobj*, *iobj*, *creg*, *cag*, *sp* or/and *cc*.
- *context_lemma* is the lemma of the context word k with its corresponding POS, which can be N , V , A , R , preposition(P), number(Z) and date(W). In the case of proper nouns, they are replaced by the *pn_n* (proper noun) POS.

Figure 2 shows an example of a dependency parsed sentence from which, for instance, three different contexts of the noun lemma *barba_n*¹⁵ are generated: [$<$:*dobj*:afeitar_v], [$>$:*mod*:largo_a] and [$>$:*de_sp*:pn_n]¹⁶. These contexts are represented in the *lemma-dep* matrix.

In [$<$:*dobj*:afeitar_v], ‘<’ indicates that the verb *afeitar_v*¹⁷ maintains a parent dependency relation with *barba_n*, *dobj* indicates that *barba_n* is the direct object of *afeitar_v*, and *afeitar_v* is the context word (lemma k) related to *barba_n* (lemma i). In [$>$:*mod*:largo], *mod* indicates that the adjective *largo_a*¹⁸ is a modifier of *barba_n*, and in [$>$:*de_sp*:pn_n] the proper noun (*Jaime* in Figure 2) is

¹⁴We used the Spanish syntactico-semantic analyzer Treeler to analyse the Diana-Araknion corpus: <http://devel.cpl.upc.edu/treeler>.

¹⁵‘beard’

¹⁶This context is the result of substituting the proper name “Jaime” by “pn_n”.

¹⁷‘to shave off’

¹⁸‘long’

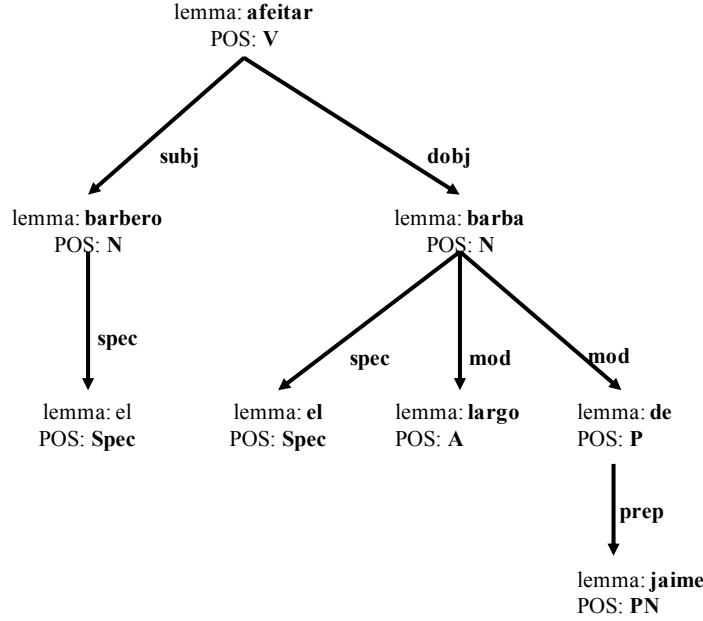


Figure 3.2: Dependency parsed sentence: *El barbero afeita la larga barba de Jaime* ('The barber shaves off James's long beard')

replaced by the *pn_n* POS tag¹⁹.

For each context obtained from the dependency structure, three different dependency contexts are generated: one that makes all the elements of the context explicit, that is, the *dep_dir*, *dep_lab* and *context_lemma* (for example, [*<:dobj:afeitar_v*]); another in which the *dep_lab* is generalized by the variable 'oth' (for example, [*<:oth:afeitar_v*])²⁰ and, finally, one context that generalizes the *context_lemma* by substituting it for the variable '*' (for example, [*<:dobj:*_v*])²¹. The three lemmas represented in example (2) do not share any context, therefore they could not be semantically related in our model. Instead, applying the generalization of contexts, we obtained a relationship between lemma₁ and lemma₂ in example (3), and between lemma₁ and lemma₃ in example (4). In example (3), the *dep_lab* is generalized, whereas in example (4) the *context_lemma* is generalized.

¹⁹Since the POS tagger does not distinguish between subclasses of proper names (person, organization, place, etc.), the grouping of all with the *pn_n* tag gives better results. We used proper nouns in the *context_lemma* configuration, but not as words in the lemma *i*. Similarly, stopwords are not included in lemma *i*.

²⁰The tag 'oth' (*other*) means that the dependency label is not specified.

²¹The symbol '*_v' means that a verb occurs in this position, but we do not specify which one it is.

2. lemma₁ [$<: subj : robar_v$ ²²]
 lemma₂ [$<: dobj : robar_v$]
 lemma₃ [$<: subj : hurtar_v$ ²³]

3. lemma₁ [$<: oth : robar_v$]
 lemma₂ [$<: oth : robar_v$]
 lemma₃ [$<: oth : hurtar_v$]

4. lemma₁ [$<: subj : *_v$]
 lemma₂ [$<: dobj : *_v$]
 lemma₃ [$<: subj : *_v$]

In this way, the generalization of contexts allows us to take into account contexts that are similar (they share two, but not all of the elements, of their context), but not identical. Therefore, we can distinguish between those lemmas that share the same or similar context, and those that have a completely different context. By adding these contexts that are similar but not identical we add new knowledge, that is, knowledge not directly present in the corpus. This new knowledge is used to generate the Unattested-Patterns.

3.3.4 Clustering

Once we described the X matrix, we proceeded to the third step detailed in Figure 3.1 that is devoted to the clustering of this matrix. The motivation of the clustering process is to find, for each lemma in the matrix, all semantically related words (lemmas). This will allow us to create new Unattested-Patterns after the linking and filtering cluster processes. To perform this clustering step, we used the CLUTO toolkit [Karypis, 2002]²⁴, which is used to cluster a collection of objects (in our case, lemmas) into a predetermined number of clusters labeled k . We applied a methodology based on Caliński and Harabasz [1974] and using cosine similarity and CLUTO's \mathcal{H}_2 metric to estimate the optimal amount of clusters.

We experimented with a number of different clustering configurations. The variables we took into account were: a) the number of most frequent lemmas,

²²'to_rob'

²³'to_steal'

²⁴We use VCLUSTER program provided in the toolkit, which computes the clustering using one of five different approaches. Four of these approaches are partitional, whereas the fifth approach is agglomerative.

with the 10,000 to 15,000 most frequent lemmas giving the best results; b) the inclusion of proper nouns or their substitution for their POS; and c) considering the lemmas with and without their POS.

We evaluated the results of these configurations manually and opted for 15,000 lemmas with proper nouns grouped according to their POS tag (*pn_n*) and with the POS tag assigned to the lemmas. This configuration gave an optimal k of 1,500 clusters applying the Caliński and Harabasz [1974] method and the \mathcal{H}_2 metric.

The inclusion of POS improves the internal consistency of the clusters. Since the POS tagger does not distinguish between subclasses of proper names (person, organization, place, etc.), grouping them according to the *pn_n* tag also gives better results. Regarding the number of lemmas, all results obtained using between 10,000 and 15,000 lemmas gave satisfactory results. The choice of the number of lemmas determines the number and the content of the clusters. In all cases, the quality of clusters obtained was acceptable. We consider a cluster as acceptable when all or almost all words contained in it share one of the following relations: synonymy, hypernymy, or hyponymy. This would allow for the use of one or more configurations for the obtention of the final lexico-syntactic patterns (see Section 3.3.6).

Using CLUTO with the selected configuration, we obtained a set of clusters $C = \{c_i : 1 \leq i \leq k\}$ from matrix X . Formally, the content of each cluster $c_i \in C$ is defined in 3.2, where le is a set of related lemmas and ctx is a set of contexts. Each lemma_pos only belongs to one cluster (i.e., it can only be defined in one le), whereas a context_lemma can be in several contexts (ctx) of different clusters.

$$c_i = \langle le, ctx \rangle \quad (3.2)$$

Formally, a context (called *context_cluster*) in ctx is described as follows:

$$context_cluster = \langle [dep_dir : dep_lab : context_lemma], score \rangle \quad (3.3)$$

where *dep_dir*, *dep_lab*, *context_lemma* corresponds to the definition of a context as shown in Section 3.3.3.2. The *score* is the sum of the different scores given by CLUTO²⁵.

For example, Table 3.1²⁶ describes the lemmas, le , and the most scored contexts, ctx , in cluster number 421_n (one of the clusters obtained in the corpus analyzed).

²⁵The sum of the twenty-five most descriptive and discriminative scores given automatically by CLUTO.

²⁶ The translation to English of Tables 1 and 2, as well as additional examples and clusters are available at <http://clic.ub.edu/corpus/>

Table 3.1 Example of a real cluster (421_n) in the Diana-Arakhion corpus in Spanish

Cluster: 421_n			
Lemmas (<i>c</i> _{421_le})	barba_n, bigote_n, cabellera_n, cabello_n, ceja_n, crin_n, melena_n, mostacho_n, patilla_n, pelaje_n, pelo_n, perilla_n, vello_n		
Contexts (<i>c</i> _{421_ctx})	[< : dobj : erizar_v],11	[< : oth : erizar_v],11	[< : oth : rizar_v],10
	[< : subj : erizar_v],10	[> : mod : espeso_a],9	[> : oth : espeso_a],9
	[> : mod : negro_a],7	[< : oth : negro_a],5	[> : mod : gris_a],8
	[< : dobj : rizar_v],8	[> : oth : gris_a],7	[< : oth : pelo_n],6
	[> : mod : rubio_a],7	[> : mod : barba_n],7	[< : oth : atusar_v],7
	[> : mod : largo_a],4	[> : oth : rubio_a],6	[< : mod : pelo_n],2
	[> : mod : rojizo_a],4	[> : oth : rojizo_a],6	[> : oth : largo_a],3
	[< : oth : bigote_n],3	[> : mod : blanco_a],3	[> : mod : cano_a],5
	[> : mod : hirsuto_a],5	[> : oth : hirsuto_a],2	[> : oth : largo_a],3
	[> : oth : negro_a],2	[> : mod : rojizo_a],2	

3.3.4.1 Results of the Clustering Process

Following our configuration, we obtained a total of 1,500 clusters in the clustering process ($k=1500$). It is worth noting that the clusters are highly morpho-syntactically and semantically cohesive.

The clusters contain lemmas belonging mostly to the same POS. It is worth mentioning that more than half of the clusters are nouns (54.20%), followed by verbs (25.80%) and adjectives (16.67%). Clusters of adverbs make up only 3.33% of the total.

Clusters contain relevant implicit information, in the sense that their lemmas belong to well-defined semantic categories, often at a very fine-grained level. For instance, we obtained clusters of adjectives with a *Positive Polarity* (5) and with a *Negative Polarity* (6)²⁶. These results encourage us to tag all the clusters with one or more semantic labels. That will enrich the obtained patterns.

5. {*c*₁₁₁, *Positive_Polarity* adjectives: admirable_a, asombroso_a, genial_a...}²⁷

6. {*c*₃₈, *Negative_Polarity* adjectives: atroz_a, aterrador_a, espantoso_a...}²⁸

²⁷‘admirable, amazing, great’

²⁸‘atrocious, scary, frightening’

3.3.5 Generalization: Linking and Filtering Clusters

The process of generalization by linking clusters (see Step 4 in Figure 3.1) is based on the set of clusters and contexts obtained using CLUTO. The processes of linking clusters and pattern generation detailed in Section 3.3.6 are the core steps of the DISCOVer methodology. The process of linking clusters uses the set of the twenty-five highest scored contexts in each cluster. According to our pattern-construction hypothesis (see Section 3.3.1), the goal of the linking of clusters is to establish the relationships between clusters using their contexts, as defined in (3.3), obtaining as a result a matrix of all possible contextual relations between clusters (see Section 3.3.5.1). Next, we apply a filtering process in order to select strongly related links taking into account different criteria (see Section 3.3.5.2).

3.3.5.1 Linking Clusters and Building the Matrix of Related Clusters

Basically, the aim of the cluster linking process is to establish the relationships between clusters and to store them in a matrix, $R_clusters$, with k rows and k columns. The k -value corresponds to the number of clusters obtained in the clustering step.

For building the matrix, for each origin cluster (x) each dep_dir and dep_lab of the $context_cluster$ (defined in Equation 3.3) are converted into a $contextual_relation$ (see Equation 3.4), while the $context_lemma$ of the $context_cluster$ is used to locate the cluster (y) in which it occurs. We obtain as a result a matrix, $R_clusters$, in which clusters are related according to a set of contextual relations stored in a $relation_set$. The sum of the scores of the $context_clusters$ in 3.3 are added together in a matrix, R_scores . The R_scores matrix is later used in the process for determining filtering thresholds.

$$contextual_relation = \langle dep_dir, dep_lab \rangle \quad (3.4)$$

For the contextual relation, defined in 3.4, dep_dir and dep_lab are the dependency direction and the dependency label defined in a context of cluster i related to cluster j . Note that the $relation_set$ of a cluster in itself is empty as $R_clusters[i][i] = \emptyset$ and $R_clusters[i][j] \neq R_clusters[j][i]$.

Following the example of cluster 421_n, described in Table 3.1, the result of the cluster linking process for this particular cluster ($i = 421_n$) is shown in Table 3.2²⁹. The first column in this table shows the related clusters, j , the second column shows the $relation_type$ that relates cluster 421_n to the related clusters j (i.e. STRONG, SEMI or WEAK, See 3.3.5.2), and finally the last column describes the lemmas in the related clusters.

²⁹For the sake of simplicity, the contexts are not included in the Table 2 and we only show a relation of each type.

Table 3.2 Some examples of cluster linking process in cluster $i=421_n$ (described in Table 3.1).

Related clusters(j)	Relation_type	Lemmas ($c_{j.le}$, where c_j refers to the related cluster, j)
1223_a	STRONG	azabache_a, bermejo_a, cano_a , canoso_a, hirsuto_a , lacio_a, lustroso_a, ondulante_a, sedoso_a...
932_v	SEMI	afeitar_v, atusar_v , cepillar_v, empolvar_v, enguantar_v, peinar_v, rasurar_v...
405_n	WEAK	contario_n, final_n, largo_n, menudo_n...

3.3.5.2 Filtering Related Clusters

In the $R_clusters$ matrix, not all contextual relationships between clusters are accepted since they have a low R_scores . For this reason, we established two criteria to automatically determine which relationships will be maintained and which ones are filtered out in the pattern generation process. For each criterion only those relations higher than a predetermined score value will be considered. The criteria are the following:

- **Criterion 1:** For each pair of clusters i and j , we take into account those relations that in each of their directions (i.e., $R_scores[i][j]$ or $R_scores[j][i]$) have a score above a minimum predetermined value, that is, $threshold_1$. This $threshold_1$ is automatically determined by finding a score value that allows for the grouping of 30% of the clusters. The relations that fulfill criterion 1 are called STRONG relations.
- **Criterion 2:** For each pair of clusters i and j , we take into account those relations in which the sum of scores in both directions (i.e., $R_scores[i][j] + R_scores[j][i]$) is higher than a predetermined value, that is, $threshold_2$, which is determined by finding a value that allows for the grouping of 50% of the clusters. The relations that fulfill criterion 2 are called SEMI relations.

Considering the example of cluster 421_n, the result of the filtering process is that, out of the three clusters linked to cluster 421_n in our example²⁶ (1223_a, 932_v, and 405_n), we will only select those with STRONG and SEMI relations, that is, 1223_a, and 932_v. Those labelled as WEAK (e.g., 405_n shown in Table 3.2) are filtered out because they do not reach the established thresholds.

3.3.6 Pattern Generation

Once the process for automatically linking and filtering clusters was carried out, we proceeded to generate the lexico-syntactic patterns to be considered as candidates for constructions (see Step 5 in Figure 3.1). Each generated pattern is defined as follows:

$$pattern = \langle lemma_i, dep_dir, dep_lab, lemma_j \rangle \quad (3.5)$$

where $lemma_i$ and $lemma_j$ are the lemmas contained in the related clusters (i and j), dep_dir and dep_lab are the dependency direction and the dependency label between the related clusters. So, there is a pattern for each $lemma_i$ and $lemma_j$ pair.

As we mentioned in Section 3.3.4, all possible configurations using between 10,000 and 15,000 lemmas gave acceptable related clusters. In order to increase the number of patterns generated we carried out the same process with a configuration using 10,000 lemmas. We combined the patterns obtained using the 10,000 and 15,000 lemmas together and removed those that were shared by both configurations. In Tables 3.3, 3.4 and 3.5, we show the number of resulting clusters and patterns, after removing the overlapping patterns, for the two configurations.

Table 3.3 Distribution of the number of related and unrelated clusters and their percentage

	10,000 lemmas	15,000 lemmas
Relation	Clusters (%)	Clusters (%)
STRONG	441 (31.50%)	461 (30.73%)
SEMI	339 (24.21%)	396 (26.40%)
Total	780 (55.71%)	857 (57.13%)
WEAK	589 (42.07%)	636 (42.40%)
Unrelated	31 (2.21%)	7 (0.47%)

As shown in Table 3.3 (second and third columns), more than 55% of the linked clusters maintain STRONG and SEMI relationships, whereas only the 2.68% of the clusters remain unrelated. Table 3.4 (second and third columns) shows the distribution of linked clusters by POS in both configurations.

The total number of lexico-syntactic patterns obtained from the two configurations of clusters (780 and 857 STRONG and SEMI related clusters) is 237,444. For the purpose of pattern generation, STRONG and SEMI clusters have been treated equally. From these patterns, we removed 16,712 patterns, those that were present in both sets of generated patterns, given as a result the total number of 220,732 patterns (See Table 3.5).

Table 3.4 Distribution of the number of related clusters and their percentage by POS

	10,000 lemmas	15,000 lemmas
POS	Clusters (%)	Clusters (%)
N	415 (53.21%)	464 (54.14%)
V	197 (25.26%)	182 (12.24%)
A	142 (18.21%)	173 (20.19%)
R	26 (3.30%)	38 (4.43%)
Total	780 (100%)	857 (100%)

Table 3.5 Distribution of the generated patterns

Lemmas	Attested-Patterns	Unattested-Patterns	Total
10,000	23,980	48,147	72,127
15,000	37,840	127,477	165,317
10,000 + 15,000	61,820	175,624	237,444
Overlapping	8,531	8,181	16,712
Sum (no overlap)	53,289	167,443	220,732

The DISCOVer methodology allows for the generation of patterns that actually occur in the corpus (Attested-Patterns), but also of lexico-syntactic patterns that are not present in the corpus but which are highly plausible in Spanish (Unattested-Patterns), since the components of the clusters are closely semantically related. As a result, we are able to enlarge the descriptive power of the source corpus. Among the patterns we generated, 61,820 were Attested-Patterns, that is, patterns that are present in the source corpus, and 175,624 were Unattested-Patterns, that is, new patterns (see Table 3.5).

Retaking the example of cluster 421_n and its related clusters we obtain patterns such as those shown in (7)³⁰:

7. <bigote_{c_421} <:dobj: cepillar_{c_932_v}>
 <melena_{c_421} <:dobj: alisar_{c_1267_v}>
 <pelaje_{c_421} >:mod: sedoso_{c_1223_a}>
 <perilla_{c_421} >:mod: gris_{c_149_a}>

All of these patterns are Unattested-Patterns, that is, they do not occur in the Diana-Araknion corpus but are generated applying our methodology and are per-

³⁰<moustache_{c_421} <:dobj: to_brush_{c_932_v}>; <mane_{c_421} <:dobj: to_smooth_{c_1267_v}>; <fur_{c_421} >:mod: silky_{c_1223_a}>; <goatee_{c_421} >:mod: grey_{c_149_a}>

fectly acceptable in Spanish. These patterns would not have been extracted using, for example, a n -gram based method or plain statistical methods.

It is worth noting the high degree of semantic cohesion between the lemmas of the same cluster and between the lemmas of the related clusters ((8)³¹, (9)³², (10)³³ and (11)³⁴).

8. <accidente_{c_470} <:dobj causar_{c_560}>
 <fuego_{c_470} <:dobj evitar_{c_560}>
 <sinistro_{c_470} <:dobj producir_{c_560}>
9. <accidente_{c_470} <:subj desencadenar_{c_560}>
 <destrozo_{c_470} <:subj producir_{c_560}>
 <incendio_{c_470} <:subj originar_{c_560}>
10. <canciller_{c_70} >:mod argentino_{c_1}>
 <embajador_{c_70} >:mod belga_{c_1}>
 <mandatario_{c_70} >:mod chileno_{c_1}>
11. <cantante_{c_155} >:mod belga_{c_1}>
 <compositor_{c_155} >:mod canadiense_{c_1}>
 <pianista_{c_155} >:mod estadounidense_{c_1}>

This strong cohesion allows for a semantic annotation of the clusters to obtain more abstract syntactico-semantic constructions that combine semantic categories (12) and (13). The semantic labels associated with each cluster have been manually added, taking into account the WordNet [Miller, 1995] upper ontologies.

12. <Event-n_{c_470} <:dobj Causative-v_{c_560}>
 <Event-n_{c_470} <:subj Causative-v_{c_560}>
13. <Person/Politician-n_{c_70} >:mod Nationality-a_{c_1}>
 <Person/Musician-n_{c_155} >:mod Nationality-a_{c_1}>

³¹<accident_{c_470} <:dobj to_cause_{c_560}>; <fire_{c_470} <:dobj to_avoid_{c_560}>; <sinister_{c_470} <:dobj to_produce_{c_560}>.

³²<accident_{c_470} <:subj to_trigger_{c_560}>, <ravage_{c_470} <:subj to_produce_{c_560}>.

³³<chancellor_{c_70} >:mod argentinian_{c_1}>; <ambassador_{c_70} >:mod belgian_{c_1}>; <representative_{c_70} >:mod chilian_{c_1}>

³⁴<singer_{c_155} >:mod belgian_{c_1}>; <song-writer_{c_155} >:mod canadian_{c_1}>; <pianist_{c_155} >:mod american_{c_1}>

In the end, we could obtain a hierarchy of candidates to be considered as different types of constructions, ranging from the most abstract syntactico-semantic constructions combining different semantic classes (12-13) to the most concrete lexico-syntactic constructions (i.e., lemma combinations) (8-11).

3.4 Evaluation

In this section we evaluate the quality of the results obtained through the DISCOVER methodology: the clusters obtained (see Section 3.4.1) and the lexico-syntactic patterns (see Section 3.4.2).

3.4.1 Clustering Evaluation

DISCOVER is a methodology for discovering lexico-syntactic patterns. The clusters of semantically related words are a by-product that we obtain as part of the process. Since the focus of this work is the methodology used and the patterns obtained, the evaluation of all possible representation and clustering algorithms is outside the scope of this article. Nevertheless, we prepared a cluster evaluation experiment in order to justify our choice and show that the quality of the obtained vectors and clusters is at least comparable with other state-of-the-art methods. As a baseline, we use standard Word2Vec [Mikolov et al., 2013c], representations with the recommended built-in k-means clustering algorithm. We evaluated the resulting clusters with respect to two criteria: a) the POS purity of each cluster, calculated automatically; and b) the semantic coherence of the lemmas in each cluster, evaluated manually by experts. The criterium applied to determine the coherence of cluster was to check if the words within the cluster held one of the following semantic relations: synonymy, hypernymy or hyponymy.

CLUTO obtained much higher results in terms of both evaluation criteria. The POS coherence of the obtained clusters was 98%, compared to 70% obtained by Word2Vec. Manual evaluation shows that 99% of the clusters obtained by CLUTO were more semantically coherent than the corresponding ones obtained by Word2Vec. These results justify the representations and parameters as adequate for the task and as comparable with the state of the art. Kovatchev et al. [2016] present a more in-depth comparison of the clustering algorithms using corpora of different sizes.

3.4.2 Pattern Evaluation

Obtaining high quality lexico-syntactic patterns is the main objective of the DISCOVER methodology. In this section, we present two different evaluations of the

obtained patterns: (1) an automatic evaluation, applying statistical association measures; and (2) a manual evaluation by expert linguists³⁵. For these evaluations, we used the sum of the patterns of both the 15,000 and 10,000 word configurations.

First, we evaluated the patterns automatically using statistical association measures and a different, much larger, corpus (Diana-Arakhnion++). In Section 3.3.1, we define two main properties of constructions: 1) Syntactic-semantic coherence and 2) Generalizability. “Syntactic-semantic coherence” entails that the words in each pattern need to be syntactically and semantically related. The “syntactic coherence” of the patterns is not evaluated explicitly, as it is considered to be a by-product of the methodology: all linked clusters from which the patterns are derived have a plausible syntactic relationship and a high connectivity score (see Section 3.3.5.1). However, we need to evaluate the semantic coherence of the patterns, that is, whether there is a semantic relation between the two lemmas. Defining and evaluating “semantic relatedness” is a non-trivial task, which often requires the use of external resources, such as WordNet and BabelNet [Navigli and Ponzetto, 2012]. However, these resources are built considering the paradigmatic relationship between words (such as synonymy, hypernymy, and hyponymy), while we are interested in evaluating syntagmatic relationships.

Evert [2008] and Pecina [2010] discuss the use of association measures for identifying collocations. They define collocations as “the empirical concept of recurrent and predictable word combinations, which are a directly observable property of natural language”. In the context of distributional semantics, this definition corresponds to “semantic coherence”.

In the DISCOVer process, we obtained two qualitatively different types of candidates-to-be-constructions: Attested-Patterns, which are observed in the corpus and Unattested-Patterns, which are obtained as a result of a generalization process that includes clustering, linking and filtering. In order to evaluate the quality of these candidates-to-be-constructions, we formulate two hypotheses and disprove their corresponding null hypotheses.

- **Hypothesis 1:** *The two lemmas in each construction are semantically related.*

Null hypothesis 1 (henceforth H_01): The degree of statistical association between the two lemmas in each of the Attested-Patterns, measured in a corpus other than the one they were extracted from, is equal to statistical chance.

³⁵An extrinsic evaluation has also been carried out in a text classification task (See Section 3.5).

- **Hypothesis 2:** *Constructions can be generalized and/or derived from other constructions through generalization.* Unattested-Patterns (derived through a generalization process) should be possible language expressions and have the property of semantic coherence.

Null hypothesis 2.1 (henceforth $H_02.1$): Unattested-Patterns are not possible language expressions. They cannot appear in a corpus.

Null hypothesis 2.2 (henceforth $H_02.2$): If Unattested-Patterns appear in a corpus, they will not have the property of semantic coherence. That is, they will have association scores equal to statistical chance.

In order to prove the two main hypotheses we needed to disprove the three null hypotheses.

For a baseline of H_01 , we extracted a list of all bigrams (BI-Patterns) from the original Diana-Arakhion corpus. Each bigram contains at least one of the 15,000 most frequent words. We removed all bigrams containing non-content words. All of the Attested-Patterns and the BI-Patterns were found and extracted from the Diana-Arakhion 100M token corpus.

For a baseline of $H_02.1$, we generated patterns by combining frequent lemmas (FL-Patterns): FL-Patterns-15 contain all combinations of the most frequent 15,000 lemmas found in the Diana-Arakhion corpus; FL-Patterns-30 contain all combinations in which one lemma is among the 15,000 most frequent lemmas and the other among the 30,000 most frequent ones; FL-Patterns-all contain all word combinations which contain at least one of the 15,000 most frequent lemmas³⁶.

We use two different statistical methods [Evert, 2008]: simple Mutual Information (MI), which is an effect size measure, and the Z-score (Z-sc), which is an evidence-based measure. Effect-size measures and evidence-based measures are qualitatively different, and for evaluation can be used complementarily. Our final experimental setup includes the following:

- Attested-Patterns, in five different test groups, based on their observed frequency in the Diana-Arakhion corpus:
 - Att-Patterns-all with an original frequency of 1 or more
 - Att-Patterns-2 with an original frequency of 2 or more
 - Att-Patterns-3 with an original frequency of 3 or more

³⁶The total number of lemmas used in the FL-Patterns (all) is 422,000.

- Att-Patterns-4 with an original frequency of 4 or more
- Att-Patterns-5 with an original frequency of 5 or more
- BI-Patterns, with an original frequency of 5 or more³⁷
- Unattested-Patterns
- FL-Patterns-15, FL-Patterns-30, FL-Patterns-all

Evaluating H_01 :

We calculated the MI and Z-sc association scores of the two words in each of the Attested-Patterns and BI-Patterns in the Diana-Arakhion++ 600M token corpus. The association score was calculated based on the sentential co-occurrence of the two words. Patterns that co-occurred less than 5 times obtained a score of 0. First, we compared the obtained association with standard thresholds, representing statistical chance: 0, 0.5, and 1 for MI; 0, 1.96, and 3.29 for Z-sc. Second, we compared the average association score of the Attested Patterns with those of the BI-Patterns.

Table 3.6 shows what percentage of the Attested-Patterns in each group obtains scores higher than statistical chance. Overall, the majority of the Attested-Patterns outperform the statistical chance baseline. The results are consistent for both the measures and their thresholds, even though they measure the association in a qualitatively different manner. It is important to note that filtering out the Attested-Patterns with a frequency of 1 significantly improves the results. We believe this factor should be taken into consideration in future experiments.

Table 3.6 Association score of Attested-Patterns compared with statistical chance

Patterns	MI			Z-sc		
	>0	>0.5	>1	>0	>1.96	>3.29
Att-Patterns-5	85%	83%	80%	85%	83%	82%
Att-Patterns-4	84%	82%	79%	84%	82%	80%
Att-Patterns-3	82%	80%	77%	82%	80%	78%
Att-Patterns-2	78%	76%	72%	78%	76%	73%
Att-Patterns-all	68%	66%	62%	68%	65%	62%

As a complementary evaluation, we directly compared the association scores of the Attested-Patterns with those of the BI-Patterns. Table 3.7 shows the average association scores for the two types of patterns³⁸. The Attested-Patterns have a much higher degree of association than the BI-Patterns. In the case of MI, the

³⁷5,285 of the BI-patterns coincide with Attested-Patterns.

³⁸The average is calculated as a simple average of all patterns of the corresponding type.

Attested-Patterns obtain scores more than two times higher than the BI-Patterns. In the case of Z-sc, the Attested-Patterns obtain scores between 30% and 100% higher than the BI-Patterns.

Table 3.7 Average association score of Attested-Patterns and BI-patterns

Patterns	Average MI	Average Z-sc
Attested-Patterns-5	3.90	52
Attested-Patterns-4	3.86	49
Attested-Patterns-3	3.80	46
Attested-Patterns-2	3.70	42
Attested-Patterns-all	3.50	35
BI-Patterns	1.72	27

The obtained results disprove H_01 and confirm Hypothesis 1. That is, we can conclude that the Attested-Patterns are semantically coherent.

Evaluating $H_02.1$:

We checked how many of the Unattested-Patterns were present in Diana-Arakhion++. As a baseline we used the FL-Patterns. Both Unattested-Patterns and FL-Patterns are not directly obtained, but are rather a result of generalization and generation using different methodologies. For each group, we calculated the percentage of the patterns that appear once and the percentage of the patterns that appear at least five times. Table 3.8 shows the results obtained.

Table 3.8 Occurrence of Unttested-Patterns and FL-Patterns

Patterns	Occurred Once	Occurred Five Times
Unattested-Patterns	54%	24%
FL-Patterns-15	24%	9%
FL-Patterns-30	11%	4%
FL-Patterns-all	4%	0.6%

Unattested-Patterns appear much more frequently than the patterns generated by simply combining frequent lemmas. 56% of the Unattested-Patterns were observed in Diana-Arakhion++. This is more than double the observance rate of the FL-Patterns-15 and five times higher than for FL-Patterns-30. 24% of the Unattested-Patterns appear in Diana-Arakhion++ with a frequency of 5 or more. This is almost three times higher than FL-Patterns-15 and six times higher than FL-Patterns-30. The results of FL-Patterns-all are much lower, showing that unfiltered pattern generation is not effective. Unattested-Patterns are linguistic patterns

given that they appear in a corpus with a much higher probability than patterns generated using a simpler frequency based methodology. These results disprove $H_02.1$.

Evaluating $H_02.2$:

We calculated the association score (MI and Z-sc) between the lemmas in each of the Unattested-Patterns that occurred at least 5 times³⁹ in Diana-Arakhnion++. We compared the scores with the same thresholds we used when evaluating H_01 . Table 3.9 shows the percentage of patterns with a score higher than the statistical chance thresholds.

Table 3.9 Association scores of Unttested-Patterns

Patterns	MI			Z-sc		
	>0	>0.5	>1	>0	>1.96	>3.29
Unattested-Patterns	93%	86%	76%	93%	80%	70%

The observed degree of association is very high. Over 90% of the observed Unattested-Patterns obtained a positive association score with respect to both measures. When comparing them with the statistical chance thresholds, the obtained results are similar to those obtained by Attested-Patterns in H_01 . The Unattested-Patterns, when observed in a different corpus, are semantically coherent. This disproves $H_02.2$.

In conclusion, the automated statistical evaluation of the patterns obtained by DISCOVer shows that: (1) Attested-Patterns are semantically coherent, as they outperform two baselines: statistical chance thresholds and BI-Patterns. These results disprove H_01 .; (2) A significant percentage (56%) of the Unattested-Patterns can be found in Diana-Arakhnion++, which is much higher than the occurrence of FL-Patterns. These results disprove $H_02.1$; (3) Whenever Unattested-Patterns occur in Diana-Arakhnion++, the statistical association between the lemmas in the patterns is much higher than the statistical chance baseline. This disproves $H_02.2$.

As we have disproved all 3 of the null hypotheses, we can conclude that the patterns obtained by the DISCOVer methodology have both properties of constructions: syntactic and semantic coherence and generalizability. Therefore they are good candidates-to-be-constructions.

We also performed a manual evaluation of the lexico-syntactic patterns. This complementary validation reinforces the results obtained in the two statistical evaluations. We prepared a dataset of 600 patterns for the manual evaluation: 300 patterns obtained by applying the DISCOVer methodology (the patterns were randomly selected from all Attested and Unattested Patterns) and 300 of the FL-

³⁹Calculating this score for patterns with lower frequency is unreliable due to the low-frequency bias in some of the measures.

Patterns-15. Three experts were asked to classify each pattern as a correct or incorrect construction. The instructions given to them were: a) evaluate whether the pattern is a possible Spanish pattern in your judgement as a native speaker; b) in case of doubt, consult the Google Search engine to check whether it is used by users. Our research questions in this evaluation were: 1) How do the experts evaluate the patterns obtained by DISCOVER?; 2) Are experts more likely to accept patterns obtained by DISCOVER than random patterns of frequent words?

The average percentage of agreement between the three annotators was 81.67% (see Table 3.10), which is considered high for a semantic evaluation task. The corresponding Fleiss Kappa score is 0.602 with expected agreement of 0.539, which is statistically significant.

Table 3.10 Interannotator agreement test

Annotators (A)	% Agreement
A1 and A2	85%
A1 and A3	80.17%
A2 and A3	79.83%
A1, A2 and A3	81.67%

The results of the evaluation are shown in Table 3.11. We use three pattern quality categories. “Strict Positive” includes patterns that were annotated as positive by all three annotators, “Positive” includes patterns that were annotated as positive by at least two annotators and “Negative” groups together patterns that were annotated as positive by one or none of the annotators. The experts accepted the majority of the DISCOVER patterns as constructions. At the same time they rejected the majority of the FL-Patterns. We also want to highlight that the percentage of “Strict Positive” patterns is very similar to the percentage of patterns that obtain a high association score. These findings confirm the results that we obtained in the automatic evaluation (See Tables 3.6 and 3.9).

Table 3.11 Expert evaluation

	DISCOVER	FL-Patterns
Strict Positive	84%	14%
Positive	93%	38%
Negative	7%	62%

3.5 Conclusions and Future Work

This article describes DISCOVer, an unsupervised methodology for automatically identifying lexico-syntactic patterns to be considered as constructions. We based this methodology on the pattern-construction hypothesis, which states that the linguistic contexts that are relevant for defining a cluster of semantically related words tend to be (part of) a lexico-syntactic construction.

Following this assumption, we developed a bottom-up language independent methodology to discover lexico-syntactic patterns in corpora. The DSM developed allows us to model the contexts of words (lemmas) taking into account their dependency directions and dependency labels. We applied a clustering process to the resulting matrix to obtain clusters of semantically related lemmas. Then we linked all the clusters that were strongly semantically related and we used them as a source of information for deriving lexico-syntactic patterns, obtaining a total number of 220,732 candidates to be constructions. We evaluated the DISCOVer methodology by applying different evaluations. First, the patterns were automatically evaluated using statistical association measures and a different, much larger, corpus. We evaluated whether the patterns we generated obtained a significantly higher association score than statistical chance. We also compared the association scores of the DISCOVer patterns with a baseline of bigrams. DISCOVer obtained better results with respect to both baselines. The patterns obtained by generalization were additionally evaluated against a baseline of randomly generated patterns. DISCOVer significantly outperforms these baselines. Second, the patterns were manually evaluated by expert linguists obtaining good results (89.33%).

This methodology only requires having at one's disposal a medium-sized corpus automatically annotated with POS tags and syntactic dependencies. Therefore, our methodology can be easily replicated with other corpora and other languages. For instance, the DISCOVer patterns were also used in a text classification task [Franco-Salvador et al., 2015]. The patterns obtained using our methodology have been compared to other representations (i.e., tf-idf, tf-idf n -grams, and enriched graph). The use of these patterns results in an accuracy of 91.69%, which outperforms the representations based on tf-idf (25.26%), tf-idf n -grams (79.26%) and an enriched graph (43.98%), proving to be the best option to represent the content of the corpus.

Furthermore, our methodology increases the descriptive power of the source corpus. First, the lexico-syntactic patterns generated constitute a structured and formalized semantic representation of the corpus. Second, the linking process enlarges the content of the initial data with new relationships not directly present in the corpus (i.e., a total of 167,443 Unattested-Patterns).

The Diana-Araknion-KB⁴⁰ can be used as a source of information to derive relevant linguistic information, such as the selection restrictions of verbs, nouns and adjectives; to disambiguate syntactic analysis in order to discard candidate parse trees; to provide a knowledge base of related words with a high degree of association measures for psycholinguistic research; and, to allow for a fine-grained corpus comparison.

The methodology presented and the results obtained, which are available in the Diana-Araknion-KB, open several lines of future research.

First, the Diana-Araknion-KB can be used as a source of information for the development of patterns at different levels of abstraction, in such a way as to obtain a hierarchy of patterns with components belonging to different levels of linguistic knowledge, that is, combining lexical, morpho-syntactic and semantic information. Second, since the same semantic category can be shared by more than one cluster, we could group them into metaclusters containing all the clusters with the same semantic category. Third, a further cluster linking process could be carried out allowing all members of a metacluster to combine with all the target clusters that are related with at least one of the members of the metacluster. Fourth, constructions could be linked in terms of transitivity to obtain larger structures. That is, if cluster A combines with cluster B, and B combines with cluster C, we have the candidate construction: A+B+C. Fifth, the methodology can be used to extract and study patterns in corpora from a specific area, such as the Biomedical domain.

To sum up, we consider that this methodology for discovering constructions outperforms the results of other proposals in the sense that it is fully automatic, language independent, and easily replicable in other corpora and languages. The quality of the results obtained and their wide range of possible applications confirm the DISCOVER methodology as a promising line of research and DSMs as a good choice for discovering linguistic knowledge.

⁴⁰Available at <http://clic.ub.edu/corpus/>

Part II

Paraphrase Typology and Paraphrase Identification

Chapter 4

WARP-Text: a Web-Based Tool for Annotating Relationships between Pairs of Texts

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó

University of Barcelona

Published at

Proceedings of the

27th International Conference on Computational Linguistics: System

Demonstrations, 2018

pp.: 132-136

Abstract We present WARP-Text, an open-source web-based tool for annotating relationships between pairs of texts. WARP-Text supports multi-layer annotation and custom definitions of inter-textual and intra-textual relationships. Annotation can be performed at different granularity levels (such as sentences, phrases, or tokens). WARP-Text has an intuitive user-friendly interface both for project managers and annotators. WARP-Text fills a gap in the currently available NLP toolbox, as open-source alternatives for annotation of pairs of text are not readily available. WARP-Text has already been used in several annotation tasks and can be of interest to the researchers working in the areas of Paraphrasing, Entailment, Simplification, and Summarization, among others.

4.1 Introduction

Multiple research fields in NLP have pairs of texts as their object of study: Paraphrasing, Textual Entailment, Text Summarization, Text Simplification, Question Answering, and Machine Translation, among others. All these fields benefit from high quality corpora, annotated at different granularity levels. However, existing annotation tools have limited capabilities to process and annotate such corpora. The most popular state-of-the-art open source tools do not natively support detailed pairwise annotation and require significant adaptations and modifications of the code for such tasks.

We present the first version of WARP-Text, an open source¹ web-based annotation tool, created and designed specifically for the annotation of relationships between pairs of texts at multiple layers and at different granularity levels. Our objective was to create a tool that is functional, flexible, intuitive, and easy to use. WARP-Text was built using PHP and MySQL standard implementation.

WARP-Text is highly configurable: the administrator interface manages the number, order, and content of the different annotation layers. The pre-built layers allow for custom definitions of labels and granularity levels. The system architecture is flexible and modular, which allows for the modification of the existing layers and the addition of new ones.

The annotator interface is intuitive and easy to use. It does not require previous knowledge or extensive annotator training. The interface has already been used in the task of annotating atomic paraphrases [Kovatchev et al., 2018a] and is currently being used on two annotation tasks in Text Summarization. The learning process of the annotators was quick and the feedback was overwhelmingly positive.

The rest of this article is organized as follows. Section 4.2 presents the Related Work. Section 4.3 describes the architecture of the interface, the annotation scheme, the usage cases, and the two interfaces: administrator and annotator. Finally, Section 4.4 presents the conclusions and the future work.

4.2 Related Work

In the last several years, the NLP community has shown growing interest in tools that are web-based, open source, and multi-purpose: WebAnno [Yimam and Gurevych, 2013], Inforex [Marcińczuk et al., 2017], and Anafora [Chen and Styler, 2013]. Other popular non web-based annotation systems include GATE [Cunningham et al., 2011] and AnCoraPipe [Bertrán et al., 2008]. These systems

¹The code is available at <https://github.com/venelink/WARP> under Creative Commons Attribution 4.0 International License.

are intended to be feature-rich and multi-purpose. However, in many tasks, it is often preferable to create a specialized annotation tool to address problems that are non-trivial to solve using the multi-purpose annotation tools. One such problem is working with multiple texts in parallel. While multi-purpose annotation tools can be adapted for such use, this often leads to a more complex annotation scheme, complicates the annotation process, requires additional annotator training and post-processing of the annotated corpora. Toledo et al. [2014] and more recently Nastase et al. [2018], Batanović et al. [2018], and Arase and Tsujii [2018] emphasize the lack of a feature-rich open-source tool for annotation of pairs of texts². Some of these authors develop simple custom-made tools with limited re-usability, designed for carrying out one specific annotation task. WARP-Text aims to address this gap in the NLP toolbox by providing a feature rich system which could be used in all these annotation scenarios.

To the best of our knowledge, the only existing multi-purpose tool that is designed to work with pairs of text and allows for detailed annotation is CoCo [España Bonet et al., 2009]. It has already been used for annotations in paraphrasing [Vila et al., 2015] and plagiarism detection [Barrón-Cedeño et al., 2013]. However, CoCo is not open source and is currently not being supported or updated.

4.3 WARP-Text

By addressing various limitations of existing tools, WARP-Text fills a gap in the state-of-the-art NLP toolbox. It offers project managers and annotators a rich set of functionalities and features: the ability to work with pairs of texts simultaneously; multi-layer annotation; annotation at different granularity levels; annotation of discontinuous scope and long-distance dependencies; and the custom definition of relationships. WARP-Text consists of two separate web interfaces: annotator and administrator. In the *administrator interface* the project manager configures the annotation scheme, defines the relationships and sets all parameters for the annotation process. The annotators work in the *annotator interface*.

WARP-Text is a tool for qualitative document annotation. It provides a wide range of configuration options and can be used for fine-grained annotation. It is best suited to medium sized corpora (containing thousands of small documents) and is not fully optimized for processing, analyzing, searching, and annotating large corpora (containing millions of documents). WARP-Text has full UTF-8 support and is language independent in the sense that it can handle documents in

²See also the discussion about looking for tools for annotating pairs of texts in the Corpora Mailing List (May 2017): <http://mailman.uib.no/public/corpora/2017-May/026526.html> - <http://mailman.uib.no/public/corpora/2017-May/026619.html>

any UTF-8 supported natural language. So far it has been used to annotate texts in English, Bulgarian (Cyrillic), and Arabic.

WARP-Text is a multi-user system and provides two different forms of interaction between the different annotators. In the *collaborative mode*, multiple annotators work on the same text and each annotator can see and modify the annotations of the others. In the *independent mode*, the annotators perform the annotation independently from one another. The different annotations can then be compared in order to calculate inter-annotator agreement.

4.3.1 Annotation Scheme

The atomic units of the annotation scheme in WARP-Text are *relationships*. The properties of the *relationships* are *label* and *scope*. The *scope* of a *relationship* is a list of continuous or discontinuous *elements* in each of the two texts. The granularity level of the scope determines the *element* type. An *element* can be the whole text, a sentence, a phrase, a token, or can be defined manually. A *layer* in WARP-Text is a set of relationships, whose scopes belong to the same granularity level³. The definition of relationships and their grouping into layers is fully configurable through the administrator interface. WARP-Text supports multi-layer annotation. That is, the same pair of texts can be annotated multiple times, at different granularity levels and using different sets of relationships.

4.3.2 Administrator Interface

The administrator interface has three main modules: a) the *dataset management module*, b) the *user management module*, and c) the *layer management module*. In the *dataset management module* the project manager can: a) import a corpus, in a delimited text format, for annotation; b) monitor the current annotation status and statistics; and c) export the annotated corpus as an SQL file or an XML file. In the *user management module* the project manager creates new users and modifies existing ones. In this module the project manager also distributes the tasks (pairs) among the annotators. In the *layer management module* the project manager configures each of the layers and determines the order of the layers in the annotation process. The project manager configures for each individual layer: 1) the granularity level; 2) the relationships that belong to the layer; 3) the sub-relationships

³There is no one-to-one correspondence between granularity level and annotation layer. Each annotation layer is a sub-task in the main annotation task. Multiple annotation layers can work at the same granularity level. For example: at layer (1) the annotator annotates the semantic relations between the tokens in the two texts; at layer (2) the annotator annotates the scope of negation and the negation cues in the two texts. Both layer (1) and layer (2) work at the token granularity level.

or properties of the relationships; 4) optional parameters such as “sentence lock” and “display previous layers”.

4.3.3 Annotator Interface

The annotator interface has three main modules: a) the *annotation statistics module*, b) the *review annotations module*, and c) the *annotation panel module*. In the *annotation statistics module* the annotator monitors the progress of the annotation and sees statistics such as the number of annotated pairs, and the remaining number of pairs. In the *review annotations module* the annotator reviews the text pairs (s)he already annotated and introduces corrections where necessary. The *annotation panel module* is the core of the annotator interface. One of our main objectives in the creation of WARP-Text was to make it easier to use for the annotators and to optimize the annotation time. For that reason we have made the *annotator panel module* as automated as possible and have limited the intervention of annotators to a minimum. The *annotation panel module* is generated dynamically, based on the user and project configuration. It loads the first text pair, assigned to the current annotator and guides the annotator through the different layers in the order specified by the project manager. Once the text pair has been annotated at all configured layers, the module updates the database, loads the next pair and repeats the process.

We illustrate the annotation process with the interface configuration that was used in the annotation of the Extended Typology Paraphrase Corpus (ETPC) [Kovatchev et al., 2018a]. The annotation scheme of ETPC consists of two layers: one layer that is configured for annotation at the text granularity level; and one layer that is configured for annotation at the token granularity level.

The screenshot shows the 'ANNOTATION' interface with a 'LOGOUT' button in the top right. It displays two text pairs for comparison. Below the texts are radio buttons for 'Semantic Relation' (Paraphrases selected, Non-Paraphrases) and 'Negation' (Yes, No selected). At the bottom, there is a 'Next' button with a checkmark icon.

Text 1:	Amrozi accused his brother , whom he called `` the witness '' , of deliberately distorting his evidence .
Text 2:	Referring to him as only `` the witness '' , Amrozi accused his brother of deliberately distorting his evidence .
Semantic Relation	<input checked="" type="radio"/> Paraphrases <input type="radio"/> Non-Paraphrases
Negation	<input type="radio"/> Yes <input checked="" type="radio"/> No

Next

Figure 4.1: Annotating relationships at textual level.

The textual layer (Figure 4.1) displays the two texts and allows the annotator to select the values for an arbitrary number of relationships between the texts. In the case of ETPC, the two textual relationships that we were interested in were:

1) “The semantic relationship between the two texts”: “Paraphrases” or “Non-paraphrases”; and 2) “The presence of negation in either of the two sentences”: “Yes” or “No”. In ETPC, both relationships had two possible options, however WARP-Text supports multiple options for each relationship. In this first layer, the scope of the relationship is the whole text.

ANNOTATION		LOGOUT	
1	Text 1:	Amrozi accused his brother , whom he called `` the witness " , of deliberately distorting his evidence .	
	Text 2:	Referring to him as only `` the witness " , Amrozi accused his brother of deliberately distorting his evidence .	
2	Semantic Relation	Paraphrases	
	Negation	No	
3	<div>✓ Previous</div>		<div>✓ Next</div>
CURRENT ANNOTATION			
	Type	Scope	Key
	L_SameP_Sub_C		
			DELETE
4	Text 1	whom	n/a
	Text 2	to him	n/a
ADD TYPE			
5	Morphology	Inflectional Changes	<div>✓ Add Type</div>

Figure 4.2: Annotating relationships at token level.

The second layer (Figure 4.2) has five functional parts, labeled in the figure with numbers from 1 to 5. The annotator can see the two texts in (1), the annotation at the previous layers in (2), and at the annotation at the current layer in (4). (3) is the navigation panel between the different layers. Finally, (5) is where the annotator can choose to add a new relationship. The list of possible relationships is defined by the project manager in the administrator interface. In the case of ETPC we organized the relationships in a two-level hierarchical system based on their linguistic meta-category. The token-layer annotation is more complex than the textual-layer annotation as it requires the annotation of scope in addition to the annotation a label⁴. When the annotator chooses a relationship, the "Add Type" button goes to the scope selection page (Figure 4.3). The scope can be discontinuous and can include elements from one of the texts only or from both. In the case of ETPC, the elements that the annotator can select are tokens. In other configurations, they can be phrases or sentences.

The flexibility of WARP-Text makes it easy to adapt for multiple tasks. The textual layer can be used in tasks such as the annotation of textual paraphrases,

⁴The token level annotation layer is an instance of the more general “atomic level annotation layer”. The organization and work flow described here are the same when the granularity level is “paragraph”, “sentence”, “phrase”, or custom defined.

ANNOTATION
LOGOUT

MARK ALL THE ELEMENTS THAT BELONG TO RELATION TYPE SAME POLARITY SUBSTITUTION (CONTEXTUAL)

Text 1:	Amrozi accused his brother , whom he called " the witness " , of deliberately distorting his evidence .	Whole text
Text 2:	Referring to him as only " the witness " , Amrozi accused his brother of deliberately distorting his evidence .	Whole text

✓
Add Type

Figure 4.3: Scope selection page.

textual entailment, or semantic similarity. The atomic level annotation layer has even more applications. As we showed in ETPC, it can be used to annotate fine-grained similarities and differences between pairs of texts. It can also be used for tasks such as manual correction of text alignment. Another possible use is, given a summary or a simplified text, to identify in the reference text the exact sentences or phrases which are summarized or simplified.

4.4 Conclusions and Future Work

In this paper we presented WARP-Text, a web-based tool for annotating relationships between pairs of texts. Our software fills an important gap as the high quality annotation of pairwise corpora at different granularity levels is needed and can benefit multiple fields in NLP. Previously available tools are not well suited for the task, require substantial modification, or are hard to configure. The main advantages of WARP-Text are that it is feature-rich, open source, highly configurable, and intuitive and easy to use.

As future work, we plan to add several functionalities to both interfaces. In the administrator interface, we plan to offer project managers tools for visualization and data analysis, and automatic calculation of inter-annotator agreement. In the annotator interface, we plan to fully explore the advantages of multi-layer architecture. By design, WARP-Text can support parent-child dependencies between layers. However, the pre-built modules available in this first release of the tool use only independent layers. That is, the annotation at one layer does not affect the configuration of the other layers. We also plan to explore the possibility of incorporating external automated pre-processing tools.

Chapter 5

ETPC - a Paraphrase Identification Corpus Annotated with Extended Paraphrase Typology and Negation

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó
University of Barcelona

Published at
Proceedings of the
Eleventh International Conference on Language Resources and Evaluation, 2018
pp.: 1384-1392

Abstract We present the Extended Paraphrase Typology (EPT) and the Extended Typology Paraphrase Corpus (ETPC). The EPT typology addresses several practical limitations of existing paraphrase typologies: it is the first typology that copes with the non-paraphrase pairs in the paraphrase identification corpora and distinguishes between contextual and habitual paraphrase types. ETPC is the largest corpus to date annotated with atomic paraphrase types. It is the first corpus with detailed annotation of both the paraphrase and the non-paraphrase pairs and the first corpus annotated with paraphrase and negation. Both new resources contribute to better understanding the paraphrase phenomenon, and allow for studying the relationship between paraphrasing and negation. To the developers of Paraphrase Identification systems ETPC corpus offers better means for evaluation and error analysis. Furthermore, the EPT typology and ETPC corpus emphasize the relationship with other areas of NLP such as Semantic Similarity, Textual Entailment, Summarization and Simplification.

Keywords Paraphrasing, Paraphrase Typology, Paraphrase Identification

5.1 Introduction

The task of Paraphrase Identification (PI) consists of comparing two texts of arbitrary size in order to determine whether they have approximately the same meaning. The most common approach to PI is as a binary classification problem, in which a system learns to make correct binary predictions (paraphrase or non-paraphrase) for a given pair of texts. The task of PI is challenging from more than one point of view. From the resource point of view, defining the task and preparing high quality corpora is a non-trivial problem due to the complex nature of “paraphrasing”. From the application point of view, for a system to perform well on PI often requires a complex ML architecture and/or a large set of manually engineered features. From the evaluation point of view, the classical task of PI does not offer many possibilities for detailed error analysis, which in turn limits the reusability and the improvement of PI systems.

In the last few years, researchers in the field of paraphrasing have adopted the approach of decomposing the meta phenomenon of “*textual paraphrasing*” into a set of “*atomic paraphrase*” phenomena, which are more strictly defined and easier to work with. “*Atomic paraphrases*” are hierarchically organized into a typology, which provides a better means to study and understand paraphrasing. While the theoretical advantages of these approaches are clear, their practical implications have not been fully explored. The existing corpora annotated with paraphrase typology are limited in size, coverage and overall quality. The only corpus of sufficient size to date annotated with paraphrase typology is the corpus by Vila et al. [2015], which contains 3900 re-annotated “*textual paraphrase*” pairs from the MRPC corpus [Dolan et al., 2004].

The use of a paraphrase typology in practical tasks has several advantages. First, “*atomic paraphrases*” are much more strict in their definition, which makes the results more useful and understandable. Second, the more detailed annotation can be useful to (re)balance binary PI corpora in terms of type distribution. Third, annotating a corpus with paraphrase types provides much better feedback to the PI systems and allows for a detailed, per-type error analysis. Fourth, enriching the corpus and improving the evaluation can provide a linguistic insight into the workings of complex machine learning systems (i.e. Deep Learning) that are traditionally hard to interpret. Fifth, corpora annotated with a paraphrase typology open the way for new research and new tasks, such as “PI by type” or “Atomic PI in context”. Finally, decomposing “*textual paraphrases*” can help relate the task of PI to tasks such as Recognizing Textual Entailment, Text Summarization, Text Simplification, and Question Answering.

In this paper, we present the Extended Typology Paraphrase Corpus (ETPC), the result of annotating the MRPC [Dolan et al., 2004] corpus with our Extended Paraphrase Typology (EPT). EPT is oriented towards practical applications and takes inspiration from several authors that work on the typology of paraphrasing and textual entailment. To the best of our knowledge, this is the first attempt to make a detailed annotation of the linguistic phenomena involved in both the positive (paraphrases) and negative (non-paraphrases) examples in the MRPC (for a total size of 5801 textual pairs). The focus on non-paraphrases and the qualitative and quantitative comparison between “*textual paraphrases*” and “*textual non-paraphrases*” provides a different perspective on the PI task and corpora.

As a separate layer of annotation, we have identified all pairs of texts that include negation and we have annotated the negation scope. This makes ETPC the first corpus that is annotated both with paraphrasing and with negation.

The rest of this article is organized as follows. Section 5.2 is devoted to the Related Work. Section 5.3 describes the proposed Extended Typology, the reasons and the practical considerations behind it. Section 5.4 explains the annotation process, the annotation scheme and instructions, the tool that we used and the corpus preprocessing. Section 5.5 presents ETPC, with its structure and type distribution. It discusses the results of the annotation and outlines some of the practical applications of the corpus. Finally, Section 5.6 concludes the article and outlines the future work.

5.2 Related Work

The task of PI is one of the classical tasks in NLP. Several corpora can be used in the task for training and/or for evaluation. Traditionally, PI is addressed using the MRPC corpus [Dolan et al., 2004]. The MRPC corpus consists of 5801 pairs, that have been manually annotated as paraphrases or non-paraphrases. More recently, Ganitkevitch et al. [2013] introduce PPDB - a very large automatic collection of paraphrases, which consists of 220 million pairs. The introduction of PPDB allowed for the training of deep learning systems, due to the significant increase of the available data. However, the quality of the PPDB pairs is much lower than those of MRPC, which makes it less reliable for evaluation. A common approach is to work on both datasets simultaneously - using the PPDB for training, and the MRPC for development and evaluation.

Closely related to the PI task is the yearly task of Recognizing Textual Entailment (RTE) [Dagan et al., 2006], which has also produced various datasets and multiple practical systems. The meta-phenomena of paraphrasing and textual entailment are very similar and are often studied together at least from a theoretical point of view. Androutsopoulos and Malakasiotis [2010] present a summary of

the tasks related to both paraphrasing and textual entailment.

The idea of decomposing paraphrasing into simpler and easier to define phenomena has been growing in popularity in the last few years. Bhagat [2009] and later Bhagat and Hovy [2013] propose a simplified framework that identifies several possible phenomena involved in the paraphrasing relation. Vila et al. [2014] propose a more complex, hierarchically structured typology that studies the different phenomena at the corresponding linguistic levels (lexical, morphological, syntactic, and discourse). More recently, Benikova and Zesch [2017] approach the problem by focusing on the paraphrasing at the level of events, understood as predicate-argument structure.

A similar decomposition tendency is noticed in the field of Textual Entailment. Garoufi [2007], Sammons et al. [2010], and Cabrio and Magnini [2014] propose different frameworks for decomposing the textual “inference” into simple, atomic phenomena. It is important to note that the similarity and the relation between paraphrasing and textual entailment is even stronger in the context of the decomposition framework and the resulting typologies. The two most exhaustive typologies: Vila et al. [2014] for paraphrasing and Cabrio and Magnini [2014] for textual entailment share the majority of their atomic phenomena as well as the overall structure and organization of the typology.

One of the advantages of the decomposition approaches is that naturally they work towards bridging the gap between the research at different granularity levels. A corpora annotated with semantic relations at both the textual and the atomic (morphological, lexical, syntactic, discourse) levels can be a valuable resource for studying the relation between them. In this same line of work, Shwartz and Dagan [2016] emphasize the importance of studying lexical entailment “in context” and the lack of resources that can enable such work. The corpora annotated with atomic paraphrase and atomic entailment phenomena can be used for that purpose without adaptation or additional annotation.

The application of paraphrase typology for the creation of resources and in practical tasks is still very limited. Most of the authors annotate a very small subsamples of around 100 text pairs to illustrate the proposed typology. The largest available corpus annotated with paraphrase types to date is the one of Vila et al. [2015]. Barrón-Cedeño et al. [2013] use this corpus to demonstrate some possible uses of the decomposition approach to paraphrasing.

5.3 Extended Paraphrase Typology

We propose the Extended Paraphrase Typology (EPT), which was created to address several of the practical limitations of the existing typologies and to provide better resources to the NLP community. EPT has better coverage than previous

typologies, including the annotation of non-paraphrases. This allows for a more in-depth understanding of the meta-phenomena and of the relation between “*textual paraphrases*” and “*atomic paraphrases*”.

5.3.1 Basic Terminology

In order to discuss the issues and limitations of existing paraphrase typologies, we first define “*paraphrasing*”, “*textual paraphrase*”, and “*atomic paraphrase*”.

We understand “*paraphrasing*” to be a specific semantic relation between two texts of arbitrary length. The two texts that are connected by a paraphrase relation have approximately the same meaning. We call them “*textual paraphrases*”. There is no limitation for “*textual paraphrases*” in terms of the nature of the linguistic phenomena involved. The concept of “*textual paraphrases*” is a practical simplification of a complex linguistic phenomenon, which is adopted in most paraphrase-related tasks, datasets, and applications. The original annotation of the MRPC and the PPDB corpora is built around the notion of textual paraphrases. Another term that we use in the article is “*textual non-paraphrases*”. With this term we refer to pairs of texts (of arbitrary length), which are not connected by a paraphrase relation.

“*Atomic paraphrases*” are paraphrases of a particular type. They must satisfy specific (linguistic) conditions, defined in the paraphrase typology. “*Atomic paraphrases*” are identified by the linguistic phenomenon which is responsible for the preservation of the meaning between the two texts. “*Atomic paraphrases*” have a (linguistically defined) scope, such as a word, a phrase, an event, or a discourse structure. The most complete typologies to date organize “*atomic paraphrases*” hierarchically, in terms of the linguistic level of the involved phenomenon. Unlike “*textual paraphrases*”, “*atomic paraphrases*” cannot be of arbitrary length. Their length is defined and restricted by their scope.

5.3.2 From Atomic to Textual Paraphrases

The relation between textual and atomic paraphrases is not easy to define and explore. It poses many challenges to the researchers, annotators, and developers of practical systems. In this section, we illustrate several issues that we want to address with the creation of the EPT and the ETPC.

The first issue to be addressed is that multiple atomic paraphrases can appear in a single textual paraphrase pair. The two texts in 1a and 1b are textual paraphrases¹. However, they include more than one atomic paraphrase²: “*magistrate*”

¹All examples in this subsection are from the MRPC corpus. When we say that the texts are textual paraphrases or textual non-paraphrases, we refer to the labels corresponding to these pairs

and “*judge*” are an instance of “*same polarity substitution*”, while “*A federal magistrate ... ordered*” and “*Zuccarini was ordered by a federal judge...*” are an instance of “*diathesis alternation*”².

1a A federal **magistrate** in Fort Lauderdale ordered him held without bail.

1b Zuccarini was ordered held without bail Wednesday by a federal **judge** in Fort Lauderdale, Fla.

Second issue is that atomic paraphrases can appear in textual pairs that are not paraphrases. The two texts in 2a and 2b as a whole are not textual paraphrases, even if they have a high degree of lexical overlap and a similar syntactic structure. However, “*Microsoft*” and “*shares of Microsoft*” are an instance of “*same polarity substitution*” - both phrases have the same role and meaning in the context of the two sentences. This demonstrates the possibility of atomic paraphrases being present in textual non-paraphrases.³

2a Microsoft fell 5 percent before the open to \$27.45 from Thursday’s close of \$28.91.

2b Shares in Microsoft slipped 4.7 percent in after-hours trade to \$27.54 from a Nasdaq close of \$28.91.

Third issue is that in certain cases, the semantic relation between the elements in an atomic paraphrase can only be interpreted within the context (as shown in the work of Shwartz and Dagan [2016]). The two texts in 3a and 3b are textual paraphrases. The out-of-context meaning of “*cargo*” and “*explosives*” differs significantly, however within the given context, they are an instance of “*same polarity substitution*”.

3a They had published an advertisement on the Internet on June 10, offering the cargo for sale, he added.

3b On June 10, the ship’s owners had published an advertisement on the Internet, offering the explosives for sale.

in MRPC.

²These types and annotation are from Vila et al. [2015].

³In fact, it is possible to find atomic paraphrases within pairs of texts connected by various relations, such as entailment, simplification, summarization, contradiction, and question-answering, among others. This is illustrated by the significant overlap of atomic types in Paraphrase Typology research and typology research in Textual Entailment.

And finally, 4a and 4b illustrate an issue that is often overlooked in theoretical paraphrase research: the linguistic phenomena behind certain atomic paraphrases do not always preserve the meaning. The meanings of “*beat*” and “*battled*” are similar, and play the same syntactic and discourse role in the structure of the texts. Therefore, the substitution of “*beat*” for “*battled*” fulfills the formal requirements of a “*same polarity substitution*”. However, after this substitution, the resulting texts are not paraphrases as they differ substantially in meaning.

4a He beat testicular cancer that had spread to his lungs and brain.

4b Armstrong, 31, battled testicular cancer that spread to his brain.

5.3.3 Objectives of EPT and Research Questions.

We argue that the objectives behind a paraphrase typology are twofold: 1) to classify and describe the linguistic phenomena involved in paraphrasing (at the atomic level); and 2) to provide the means to study the function of atomic paraphrases within pairs of texts of arbitrary size and with various semantic relations (such as, textual paraphrases, textual entailment pairs, contradictions, and unrelated texts).

Traditionally, the authors of paraphrase typologies have focused on the first objective while the latter is mentioned only briefly or ignored altogether. In our work, we want to extend the existing work on paraphrase typology in the direction of Objective 2, as we argue that it is crucial for applications. We pose four research questions, that we aim to address with the creation of EPT and ETPC:

RQ1 what is the relation between atomic and textual paraphrases considering the distribution of atomic paraphrases in textual paraphrases?

RQ2 what is the relation between atomic paraphrases and textual non-paraphrases considering the distribution of atomic paraphrases in textual non-paraphrases?

RQ3 what is the role of the context in atomic paraphrases?

RQ4 in which cases do the linguistic phenomena behind an atomic paraphrase preserve the meaning and in which they do not?

5.3.4 The Extended Paraphrase Typology

The full Extended Paraphrase Typology is shown in Table 5.1. It is organized in seven meta categories: “Morphology”, “Lexicon”, “Lexico-syntax”, “Syntax”, “Discourse”, “Other”, and “Extremes”. Sense Preserving (Sens Pres.) shows

Table 5.1 Extended Paraphrase Typology

ID	Type	Sense Pres.
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	Coordination changes	+
17	Subordination and nesting changes	+
Discourse-based changes		
18	Punctuation changes	+
19	Direct/indirect style alternations	+ / -
20	Sentence modality changes	+
21	Syntax/discourse structure changes	+
Other changes		
22	Addition/Deletion	+ / -
23	Change of order	+
24	Semantic (General Inferences)	+ / -
Extremes		
25	Identity	+
26	Non-Paraphrase	-
27	Entailment	-

whether a certain type can give rise to textual paraphrases (+), to textual non-paraphrases (-), or to both (+ / -)⁴. The typology contains 25 atomic paraphrase types (+) and 13 atomic non-paraphrase types (-). It is based on the work of Vila et al. [2014] and aims to extend it in two directions in order to address the four Research Questions.

First, we have added three new atomic paraphrase types - we split the atomic types “*same polarity substitution*” and “*opposite polarity substitution*” into two separate types based on the nature of the relation between the substituted words: “*habitual*” and “*contextual*”. We have also added the type “*same polarity substitution (named entity)*”. While the principle behind all substitutions is the same, in practice there is a significant difference whether the replaced words are connected in their habitual meaning, contextually, or refer to related named entities in the world. Instances of the new types can be seen in sentence pairs 5 (“*same polarity substitution (habitual)*”), 6 (“*same polarity substitution (contextual)*”), 7 (“*same polarity substitution (named entity)*”), 8 (“*opposite polarity substitution (habitual)*”), and 9 (“*opposite polarity substitution (contextual)*”)

5a A federal magistrate in Fort Lauderdale ordered him held without bail.

5b Zuccarini was ordered held without bail Wednesday by a federal judge in Fort Lauderdale, Fla.

6a Meanwhile, the global death toll approached 770 with more than 8,300 people sickened since the severe acute respiratory syndrome virus first appeared in southern China in November.

6b The global death toll from SARS was at least 767, with more than 8,300 people sickened since the virus first appeared in southern China in November.

7a He told The Sun newspaper that Mr. Hussein’s daughters had British schools and hospitals in mind when they decided to ask for asylum.

7b “Saddam’s daughters had British schools and hospitals in mind when they decided to ask for asylum – especially the schools,” he told The Sun.

8a Leicester failed in both enterprises.

8b He did not succeed in either case.

9a A big surge in consumer confidence has provided the only positive economic news in recent weeks.

⁴A more detailed table of EPT, with additional examples for each atomic type is available at <https://github.com/venelink/ETPC> and in Appendix A of the thesis.

- 9b Only a big surge in consumer confidence has interrupted the bleak economic news.

Second, we have introduced the “*sense preserving*” feature in 13 of the atomic types. As we have shown in the previous section (examples 4a and 4b), the same atomic linguistic transformation (such as substitution, diathesis alternation, and negation switching) can give rise to different semantic relations at textual level: paraphrasing, entailment, and contradiction, among others. This idea has already been expressed by Cabrio and Magnini [2014] in the field of Recognizing Textual Entailment. Building on this idea, we identify 13 atomic types that can, in different instances, give rise to both paraphrases and non-paraphrases. Sentence pairs 10 and 11 show an example of sense preserving and non-sense preserving “*Inflection change*” types. In 10a and 10b, both “*streets*” and “*street*” are a generalization with the meaning “*all streets*”. In a similar way, in 11b, “*boats*” has the meaning as “*all boats*”. However in 11a, “*boat*” can have the meaning “*one particular boat*”, thus the inflectional change “*boat - boats*” is not sense-preserving.

10a It was with difficulty that the course of streets could be followed.

10b You couldn’t even follow the path of the street.

11a You can’t travel from Barcelona to Mallorca with the boat.

11b Boats can’t travel from Barcelona to Mallorca.

The changes introduced in EPT allow us to work on all four Research Questions (RQs) defined in Section 5.3.3 This is a clear advantage over the existing paraphrase typologies, which are only suitable for addressing **RQ1**. For **RQ1**, we annotated all atomic types in the positive (“paraphrases”) portion of MRPC and measured their distribution. For **RQ2**, we annotated all atomic types in the negative (“non-paraphrases”) portion of MRPC and compared the distribution of the types in the positive and negative portions. For **RQ3**, the two newly added “contextual” types allow us to distinguish and compare context dependent from context independent atomic paraphrases. Finally, for **RQ4**, the addition of “sense preserving” allows us to annotate, isolate and compare the sense preserving and non-sense preserving instances of the same linguistic phenomena.

5.4 Annotation Scheme and Guidelines

We propose the Extended Paraphrase Typology (EPT) with a clear practical objective in mind: to create language resources that improve the performance, evaluation, and understanding of the systems competing on the task of PI and to open

new research directions. We used the EPT to annotate the MRPC corpus with atomic paraphrases. We annotated all 5801 text pairs in the corpus, including both the pairs annotated as paraphrases (3900 pairs) and those annotated as non-paraphrases (1901 pairs).

As a basis, we used the MRPC-A corpus by Vila et al. [2015], which already contains some annotated atomic paraphrases. Our annotation consisted of three steps, corresponding to the three different layers of annotation.

First, we annotated the non-sense preserving atomic phenomena (Section 5.4.1) in the textual non-paraphrases. Second, we annotated the sense preserving atomic paraphrase phenomena (Section 5.4.2) in both textual paraphrases and textual non-paraphrases. And third, we identified all sentences in the corpus containing negation, and explicitly annotated the negation scope (Section 5.4.4).

For the purpose of the annotation, we created a web-based annotation tool, Pair-Anno, capable of annotating aligned pairs of discontinuous scopes in two different texts⁵. As the scope of each atomic phenomena is one or more sets of tokens, prior to the annotation we automatically tokenized the corpus using NLTK [Bird et al., 2009].

5.4.1 Non-Sense Preserving Atomic Phenomena

Textual non-paraphrases in the MRPC corpus typically have a very high degree of lexical overlap and a similar syntactic and discourse structure. Normally, they differ only by a few elements (morphological, lexical, or structural), but the modification of these few elements leads to a substantial difference in the meaning of the two texts as a whole. The annotation of non-sense preserving phenomena aims to identify these key elements and study the linguistic nature of the modification.

When annotating atomic phenomena, our experts identified and annotated the type, the scope, and in some paraphrase types, the key element. Both the scope and the key are kept as a 0-indexed list of tokens. Examples 12a and 12b show a textual pair, annotated as non-paraphrase in the MRPC corpus. Table 5.2 shows the annotation of non-sense preserving atomic phenomena in 12a and 12b. The key differences are “*opposite polarity substitution (habitual)*” (type id 10) of “slip” with “rise”, and the “*same polarity substitution (named entity)*” (type id 7) of “Friday” with “Thursday”.

12a The loonie , meanwhile , continued to slip in early trading Friday .

12b The loonie , meanwhile , was on the rise again early Thursday .

⁵Screenshots of Pair-Anno can be seen at <https://github.com/venelink/ETPC>.

Table 5.2 Non-sense preserving phenomena

type	pair	s1 scope	s2 scope	s1 text	s2 text
7	146	11	11	Friday	Thursday
10	146	7	8	slip	rise

The annotation of 12a and 12b illustrates one of the issues when annotating non-sense preserving phenomena. In many textual pairs, there is more than one “key” difference. In those cases, all of the phenomena were annotated separately. Nevertheless, the annotators were instructed to be conservative and only annotate phenomena that carry substantial differences in the meaning of the two texts. Determining which differences are substantial, and which are not was the main challenge for the annotators. Due to the difficulty of the task, we selected annotators that were expert linguists with a high proficiency of English⁶.

When the two texts were substantially different and it was not possible to identify the atomic phenomena responsible for the difference, the pair was annotated with atomic type “*non-paraphrase*” (examples 13a and 13b) or “*entailment*” (examples 14a and 14b).

13a That compared with \$35.18 million, or 24 cents per share, in the year-ago period.

13b Earnings were affected by a non-recurring \$8 million tax benefit in the year-ago period.

14a The year-ago comparisons were restated to include Compaq results.

14b The year-ago numbers do not include figures from Compaq Computer.

5.4.2 Sense Preserving Atomic Phenomena

For the annotation of the sense preserving atomic phenomena, we used the same annotation scheme format as the one for the non-sense preserving phenomena. Each phenomenon is identified by a type, a scope, and, where applicable, a key. 15a and 15b show a textual pair, annotated as a paraphrase in the MRPC. An example of a single annotated atomic phenomenon can be seen in Table 5.3

15a Amrozi accused his brother , whom he called “ the witness ” , of deliberately distorting his evidence .

⁶The full annotation guidelines for both sense preserving and non-sense preserving phenomena can be found at <https://github.com/venelink/ETPC> and in Appendix A of the thesis.

- 15b Referring to him as only “ the witness ” , Amrozi accused his brother of deliberately distorting his evidence .

Table 5.3 Sense preserving phenomenon

type	pair	s1 scope	s2 scope	s1 text	s2 text
6	1	5	1, 2	whom	to him

For the 3900 text pairs already annotated by Vila et al. [2015], we worked with the existing corpus and we only re-annotated the 3 new sense preserving paraphrase types introduced in EPT. For the 1901 textual non-paraphrases, which were not annotated in MRPC-A, we performed a full annotation with all 25 sense preserving atomic types.

5.4.3 Inter-Annotator Agreement

In this section, we present the measures for calculating the inter-annotator agreement and the agreement score on the first two layers of annotation: non-sense preserving atomic phenomena and sense preserving atomic phenomena.

The measure that we use is the IAPTA TPO, introduced by Vila et al. [2015]. It is a fine-grained measure, created specifically for the task of annotating paraphrase types. It takes into account the agreement with respect to both the label and the scope of the phenomena. It is a pairwise agreement measure, obtained by calculating the Precision, Recall and F1 of one of the annotators, while using another annotator as a gold standard. There are two versions of the measure - TPO-partial, which requires that the annotators select the same label and that the scopes overlap by at least one token; and TPO-total which requires full overlap of label and scope.

The classical TPO measures are pairwise, they calculate the agreement between two annotators. When the annotation process involves more than two annotators, we first calculate the pairwise TPO measure between any two annotators and then we use one of three different techniques for calculating the overall agreement for the corpus. TPO (avg) is the most simple score, as it is the average of all pairwise TPO scores. TPO (union) is the union of all pairwise TPO agreement tables. That is, any phenomena that is annotated with the same label and the same scope by any 2 annotators is part of the TPO (union). Finally, TPO (gold) is the average F1 score of the three annotators, when treating TPO (union) as a gold standard. TPO (union) and TPO (gold) are two new measures, that we propose as part of this paper. TPO (union) represents all the “high quality” phenomena (that

is, phenomena annotated the same way by multiple annotators). TPO (gold) represents the probability that any of our annotators would annotate “high quality” phenomena.

The annotation of the sense preserving atomic paraphrases was carried out by two expert annotators, while the annotation of the non-sense preserving atomic phenomena was carried out by three expert annotators. For the purpose of calculating the inter-annotator agreement, all experts were given the same 180 text pairs (roughly 10 % of all non-paraphrase pairs in the corpus). The pairs were split in 3 equal parts and given to the annotators in three different stages of the annotation process: one at the beginning, one in the middle, and one at the end of the annotation process. Table 5.4 shows the obtained scores, where ETPC (-) stands for the non-sense preserving layer, ETPC (+) stands for the sense-preserving layer of annotation and MRPC-A is the annotation of Vila et al. [2015]. For ETPC (+) we only had two annotators, so we were not able to calculate TPO (union) and TPO (gold). Since these measures have been introduced by us in the current paper, the MRPC-A corpus by Vila et al. [2015] does not have values for them either.

Table 5.4 Inter-annotator Agreement

Measure	ETPC (-)	ETPC (+)	MRPC-A
TPO-partial (avg)	0.72	0.86	0.78
TPO-total (avg)	0.68	0.68	0.51
TPO (union)	0.77	n-a	n-a
TPO (gold)	0.86	n-a	n-a

ETPC (+) and MRPC-A are directly comparable as they measure the agreement on the same task (annotation of sense-preserving atomic phenomena). The results show much higher agreement score with respect to both TPO-partial (0.86 against 0.78) and TPO-total (0.68 against 0.51). ETPC (-) measures the agreement on a different task (annotation of non-sense preserving phenomena). The TPO-partial score of ETPC (-) is lower than both ETPC (+) and MRPC-A (0.72 against 0.86 and 0.78 respectively), however the TPO-total score is equal to that of ETPC (+) and much higher than that of MRPC-A. It is interesting to note that there is almost no difference between TPO-partial and TPO-total for ETPC (-) (0.72 against 0.68), while for ETPC (+) and MRPC-A, the difference is significant. The TPO (union) for ETPC (-) shows that 77% of all phenomena are annotated the same way by at least 2 of the annotators. The TPO (gold) indicates that the probability of any of our experts annotating a “gold” example is 86%. Considering the difficulty of the task, the obtained results indicate the high quality of the annotated corpus.

5.4.4 Annotation of Negation

During the first two steps of the annotation, we identified all sentences that contain negation. For every instance of negation we annotated the negation cues and the scope of negation. 16a and 16b illustrate an example of annotated negation.

16a (Moore had (**no** [negation marker]) immediate comment Tuesday [scope])

16b (Moore (**did not** [negation marker]) have an immediate response Tuesday [scope])

5.5 The ETPC Corpus

This section presents the results of the annotation of the ETPC corpus. Section 5.5.1 shows the results of annotating non-sense preserving phenomena. Section 5.5.2 shows the results of annotating sense preserving phenomena. Section 5.5.3 discusses the results and the Research Questions, and Section 5.5.4 lists some applications of ETPC.

5.5.1 Non-Sense Preserving Atomic Phenomena

Table 5.5 shows the distribution of the non-sense preserving phenomena. Type Relative Frequency (Type RF) shows the relative distribution of the atomic types. Occurrence Frequency (Type OF) shows the distribution of phenomena per sentence, that is in how many textual pairs each phenomenon can be found⁷. The total number of non-sense preserving phenomena is 3406 in 1901 text pairs.

Both Type Relative Frequency (RF) and Occurrence Frequency (OF) indicate that the non-paraphrase portion of the corpus is not well balanced with respect to atomic phenomena. In 260 of the text pairs (13.7%), the annotators selected “*non-paraphrase*” indicating that the two texts were substantially different. In the rest of the pairs, the most common reason for the “non-paraphrase” label at textual level was “*Addition/Deletion*” (52% RF, 65.5% OF), followed by “*Same polarity substitution (named entity)*” (27% RF, 22.5% OF), “*Same polarity substitution (contextual)*” (RF 9.3%, OF 15.5%), and “*Opposite polarity substitution (habitual)*” (RF 2.8%, OF 4.6%). These are the only types with Type Relative Frequency and Occurrence Frequency above 1%, and they constitute over 99% of all non-sense preserving atomic phenomena annotated in the corpus. Six of the atomic phenomena are represented only with a few examples, while two are not represented at all.

⁷The sum of all Occurrence Frequencies exceeds 100, as one sentence often contains more than one atomic phenomenon.

Table 5.5 Distribution of non-sense preserving phenomena

Type	Type RF	Type OF
Inflectional	0.02%	0.04%
Same Polarity (con)	9.3%	15.5%
Same Polarity (ne)	27.5%	22.5%
Opp Polarity (hab)	2.7%	4.4%
Opp Polarity (con)	0.01%	0.02%
Converse	0.01%	0.02%
Diathesis	0.01%	0.01%
Negation	0.02%	0.03%
Direct/Indirect	0%	0%
Addition/Deletion	52%	65.5%
Semantic based	0%	0%
Non-paraphrase	7.6%	13.7%
Entailment	0.02%	0.04%

5.5.2 Sense Preserving Atomic Phenomena

Table 5.6 shows the distribution of sense preserving atomic phenomena in the textual paraphrase and non-paraphrase portions of the corpus⁸. For the textual paraphrase portion, we used the numbers reported by Vila et al. [2015] with partial re-annotation to account for the new types in ETPC. For “*same polarity substitution*”, 35% of the phenomena were re-annotated as “*habitual*”, 47% as “*contextual*”, and 18% as “*named entity*”. For “*opposite polarity substitution*” 21% of the phenomena were “*contextual*” and 79% of the phenomena were “*habitual*”.

The results show that the distribution of sense-preserving phenomena is relatively consistent between the two portions of the corpus. The most notable differences between the two distributions are the frequencies of “*same polarity substitution (named entity)*”, “*synthetic/analytic*”, “*addition/deletion*”, and “*identity*”. Both distributions are not well balanced in terms of atomic types, with 8 types (“*addition/deletion*”, “*identity*”, “*same polarity substitution (contextual)*”, “*same polarity substitution (habitual)*”, “*synthetic/analytic*”, “*same polarity substitution (named entity)*”, “*change of order*”, and “*punctuation*”) responsible for over 80% of the phenomena.

⁸At the time of the submission of this paper, the annotation of the non-paraphrase portion was not finished. The reported results are for 500 annotated pairs (about 30% of the corpus). The full figures will be made available at <https://github.com/venelink/ETPC>

Table 5.6 Distribution of Sense preserving phenomena in textual paraphrases and textual non-paraphrases

Type	Non Paraphrase	Paraphrase
Inflectional	2.13%	2.78%
Modal verb	0.59%	0.83%
Derivational	0.35%	0.85%
Spelling changes	1.30%	2.91%
Same Polarity (hab)	10.55%	8.68%
Same Polarity (con)	11.15%	11.66%
Same Polarity (ne)	7.11%	5.08%
Format	1.06%	1.1%
Opp Polarity (hab)	0%	0.07%
Opp Polarity (con)	0%	0.02%
Synthetic/analytic	7.82%	3.80%
Converse	0.12%	0.20%
Diathesis	0.83%	0.73%
Negation	0%	0.09%
Ellipsis	0.47%	0.30%
Coordination	0.24%	0.22%
Subord. and nesting	1.18%	2.14%
Punctuation	2.72%	3.77%
Direct/Indirect	0.24%	0.30%
Sentence modality	0%	0%
Synt./Disc. structure	1.30%	1.39%
Addition/Deletion	20.04%	25.94%
Change of order	3.08%	3.89%
Semantic	0%	1.53%
Identity	25.02%	17.54%
Non-Paraphrase	2.49%	3.81%
Entailment	0.12%	0.37%

5.5.3 Discussion

In this section we briefly discuss the annotation results and the Research Questions that we posed in Section 5.3.3

With respect to **RQ1** and **RQ2**, we measured the raw frequency distribution of the sense preserving atomic phenomena in both the paraphrase and non-

paraphrase portions of the corpus. We make two important observations from the data. First, the corpus is not well balanced in terms of type distribution in either of the portions. It can be seen in Table 5.6 that 8 of the types are overrepresented while the rest are underrepresented. This imbalance is even more significant in terms of meta-categories. The structure meta-types “syntax” and “discourse” account for less than 10 % of all types. Second, the raw frequency distribution of atomic phenomena in textual paraphrases and textual non-paraphrases is very similar. This finding suggests that it is the non-sense preserving phenomena that are mostly responsible for the relation at textual level in this corpus. This makes the annotation of the non-sense preserving phenomena even more important for the PI task.

With respect to **RQ3**, we annotated the “*same polarity substitution (contextual)*” and “*opposite polarity substitution (contextual)*” types in all portions of the corpus. For “*same polarity substitution*”, over 40% of the sense-preserving and over 25% of the non-sense preserving instances were contextual. For “*opposite polarity substitution*”, 21% of the sense-preserving instances were annotated as contextual, while in the non-sense preserving portion we found almost no contextual instances.

With respect to **RQ4**, we measured the raw frequency distribution of the non-sense preserving phenomena. If we compare it with the distribution of sense preserving phenomena, we can see that the differences are noteworthy and we can easily differentiate between the two distributions. Non-sense preserving phenomena are even less balanced than sense preserving phenomena, with just 4 types responsible for almost all instances. The structure types “syntax” and “discourse” are not represented at all, with all frequent types being either “lexical”, “lexico-syntactic”, or “other”.

Finally, it is worth mentioning that 13% of the sentences in the textual paraphrase portion of the corpus and 12% of the sentences in the textual non-paraphrase portion contain negation. The relative distribution in the paraphrase and in the non-paraphrase portion of the corpus is consistent. The negation scope for each of these sentences has been annotated in a separate layer.

5.5.4 Applications of ETPC

The ETPC corpus has clear advantages over the currently available PI corpora, and the MRPC in particular. It is much more informative and can be used in several ways.

First, ETPC can be used as a single PI corpus. The annotation with atomic types makes it much more informative for evaluation than any other existing PI corpus. PI systems are currently evaluated in terms of binary Precision, Recall, F1 and Accuracy. ETPC provides the developers with much more detailed infor-

mation, without requiring any additional work on the developers’ side. Knowing which atomic types are involved in the correct and incorrect classification helps the error analysis and should lead to an improvement in the these systems’ performance. It also promotes reusability.

Second, ETPC can be used to provide quantitative and qualitative analysis of the MRPC corpus, as we have already shown in section 5.5.3 By having a detailed statistical analysis of the content of the corpus we can identify possible biases and promote the creation of better and more balanced corpora.

Third, ETPC can be easily split into various smaller corpora built around a certain atomic type or a class of types. Each of them can be used for a new task of Atomic Paraphrase Identification. It can be used to study the nature of the relation between atomic paraphrases and textual paraphrases.

Finally, ETPC can be used to study the role of negation in PI, a research question that, to date, has received very little attention.

5.6 Conclusions and Future Work

In this paper we presented the ETPC corpus - the largest corpus annotated with detailed paraphrase typology to date. For the annotation we used the new Extended Paraphrase Typology, a practically oriented typology of atomic paraphrases. The annotation process included three expert linguists and covered the whole 5801 text pairs from the MRPC corpus. The full corpus is publicly available in two formats: SQL and XML⁹.

ETPC is a high quality resource for paraphrase related research and the task of PI. It provides more in-depth analysis of the existing corpora and promotes better understanding of the phenomena, the data, and the task. It also identifies several problems, such as the under-representation of structure based types and the over-representation of lexical based types. ETPC sets an example for the development of new feature-rich corpora for paraphrasing research. It also promotes collaboration between similar areas, such as PI, RTE and Semantic Similarity.

Our work opens several lines of future research. First, the ETPC can be used to re-evaluate existing state-of-the-art PI systems. This detailed evaluation can lead to improvements of the existing PI systems and the creation of new ones. Second, it can be used to create new corpora for paraphrase research, which will be more balanced in terms of type distribution. Third, it can be used to study the nature of the paraphrase phenomenon and the relation between “atomic” and “tex-

⁹We have also made publicly available all complementary data, such as annotation guidelines, screenshots of the interface, detailed statistics, as well as the ETPC_Neg corpus, composed only from the paraphrase and non-paraphrase pairs containing negation (<https://github.com/venelink/ETPC>).

tual” paraphrases. Finally, the EPT and ETPC can be extended to other research areas, such as lexical and textual entailment, semantic similarity, simplification, summarization, and question answering, among others.

Chapter 6

A Qualitative Evaluation Framework for Paraphrase Identification

Venelin Kovatchev, M. Antònia Martí, Maria Salamó, and Javier Beltran
University of Barcelona

Published at
*Proceedings of the
Twelfth Recent Advances in Natural Language Processing Conference, 2019*
pp.: 569-579

Abstract In this paper, we present a new approach for the evaluation, error analysis, and interpretation of supervised and unsupervised Paraphrase Identification (PI) systems. Our evaluation framework makes use of a PI corpus annotated with linguistic phenomena to provide a better understanding and interpretation of the performance of various PI systems. Our approach allows for a qualitative evaluation and comparison of the PI models using human interpretable categories. It does not require modification of the training objective of the systems and does not place additional burden on the developers. We replicate several popular supervised and unsupervised PI systems. Using our evaluation framework we show that: 1) Each system performs differently with respect to a set of linguistic phenomena and makes qualitatively different kinds of errors; 2) Some linguistic phenomena are more challenging than others across all systems.

6.1 Introduction

In this paper we propose a new approach to evaluation, error analysis and interpretation in the task of Paraphrase Identification (PI). Typically, PI is defined as comparing two texts of arbitrary size in order to determine whether they have approximately the same meaning [Dolan et al., 2004]. The two texts in 1a and 1b are considered paraphrases, while the two texts at 2a and 2b are non-paraphrases.¹ In 1a and 1b there is a change in the wording (“*magistrate*” - “*judge*”) and the syntactic structure (“*was ordered*” - “*ordered*”) but the meaning of the sentences is unchanged. In 2a and 2b there are significant differences in the quantities (“5%” - “4.7%” and “\$27.45” - “\$27.54”).

1a A federal magistrate in Fort Lauderdale ordered him held without bail.

1b He was ordered held without bail Wednesday by a federal judge in Fort Lauderdale, Fla.

2a Microsoft fell **5 percent** before the open to **\$27.45** from Thursday’s close of \$28.91.

2b Shares in Microsoft slipped **4.7 percent** in after-hours trade to **\$27.54** from a Nasdaq close of \$28.91.

The task of PI can be framed as a binary classification problem. The performance of the different PI systems is reported using the Accuracy and F1 score measures. However this form of evaluation does not facilitate the interpretation and error analysis of the participating systems. Given the Deep Learning nature of most of the state-of-the-art systems and the complexity of the PI task, we argue that better means for evaluation, interpretation, and error analysis are needed. We propose a new evaluation methodology to address this gap in the field. We demonstrate our methodology on the ETPC corpus [Kovatchev et al., 2018a] - a recently published corpus, annotated with detailed linguistic phenomena involved in paraphrasing.

We replicate several popular state-of-the-art Supervised and Unsupervised PI Systems and demonstrate the advantages of our evaluation methodology by analyzing and comparing their performance. We show that while the systems obtain similar quantitative results (Accuracy and F1), they perform differently with respect to a set of human interpretable linguistic categories and make qualitatively different kinds of errors. We also show that some of the categories are more challenging than others across all evaluated systems.

¹Examples are from the MRPC corpus [Dolan et al., 2004]

6.2 Related Work

The systems that compete on PI range from using hand-crafted features and Machine Learning algorithms [Fernando and Stevenson, 2008, Madnani et al., 2012, Ji and Eisenstein, 2013] to end-to-end Deep Learning models [He et al., 2015, He and Lin, 2016, Wang et al., 2016, Lan and Xu, 2018b, Kiros et al., 2015, Conneau et al., 2017]. The PI systems are typically divided in two groups: Supervised PI systems and Unsupervised PI systems.

“Supervised PI systems” [He et al., 2015, He and Lin, 2016, Wang et al., 2016, Lan and Xu, 2018b] are explicitly trained for the PI task on a PI corpora. “Unsupervised PI systems” in the PI field is a term used for systems that use a general purpose sentence representations such as Mikolov et al. [2013b], Pennington et al. [2014], Kiros et al. [2015], Conneau et al. [2017]. To predict the paraphrasing relation, they can compare the sentence representations of the candidate paraphrases directly (ex.: cosine of the angle), and use a PI corpus to learn a threshold. Alternatively they can use the representations as features in a classifier.

The complexity of paraphrasing has been emphasized by many researchers [Bhagat and Hovy, 2013, Vila et al., 2014, Benikova and Zesch, 2017]. Similar observations have been made for Textual Entailment [Sammons et al., 2010, Cabrio and Magnini, 2014]. Gold et al. [2019] study the interactions between paraphrasing and entailment.

Despite the complexity of the phenomena, the popular PI corpora [Dolan et al., 2004, Ganitkevitch et al., 2013, Iyer et al., 2017, Lan et al., 2017] are annotated in a binary manner. In part it is due to lack of annotation tools capable of fine-grained annotation of relations. WARP-Text [Kovatchev et al., 2018b] fills this gap in the NLP toolbox.

The simplified corpus format poses a problem with respect to the quality of the PI task and the ways it can be evaluated. The vast majority of the state-of-the-art systems in PI provide no or very little error analysis. This makes it difficult to interpret the actual capabilities of a system and its applicability to other corpora and tasks.

Some researchers have approached the problem of non-interpretability by evaluating the same architecture on multiple datasets and multiple tasks. Lan and Xu [2018a] apply this approach to Supervised PI systems, while Aldarmaki and Diab [2018] use it for evaluating Unsupervised PI systems and general sentence representation models.

Linzen et al. [2016] demonstrate how by modifying the task definition and the evaluation the capabilities of a Deep Learning system can be determined implicitly. The main advantage of such an approach is that it only requires modification and additional annotation of the corpus. It does not place any additional burden on the developers of the systems and can be applied to multiple systems without

additional cost.

We follow a similar line of research and propose a new evaluation that uses ETPC [Kovatchev et al., 2018a]: a PI corpus with a multi-layer annotation of various linguistic phenomena. Our methodology uses the corpus annotation to provide much more feedback to the competing systems and to evaluate and compare them qualitatively.

6.3 Qualitative Evaluation Framework

6.3.1 The ETPC Corpus

ETPC [Kovatchev et al., 2018a] is a re-annotated version of the MRPC corpus. It contains 5,801 text pairs. Each text pair in ETPC has two separate layers of annotation. The first layer contains the traditional binary label (paraphrase or non-paraphrase) of every text pair. The second layer contains the annotation of 27 “*atomic*” linguistic phenomena involved in paraphrasing, according to the authors of the corpus. All phenomena are linguistically motivated and humanly interpretable.

- 3a A federal **magistrate** in Fort Lauderdale ordered him held without bail.
- 3b He was ordered held without bail Wednesday by a federal **judge** in Fort Lauderdale, Fla.

We illustrate the annotation with examples 3a and 3b. At the binary level, this pair is annotated as “paraphrases”. At the “atomic” level, ETPC contains the annotation of multiple phenomena, such as the “*same polarity substitution (habitual)*” of “magistrate” and “judge” (marked **bold**) or the “*diathesis alternation*” of “...ordered him held” and “he was ordered by...” (marked underline).

For the full set of phenomena, the linguistic reasoning behind them, their frequency in the corpus, real examples from the pairs, and the annotation guidelines, please refer to Kovatchev et al. [2018a].

6.3.2 Evaluation Methodology

We use the corpus to evaluate the capabilities of the different PI systems implicitly. That means, the training objective of the systems remains unchanged: they are required to correctly predict the value of the binary label at the first annotation layer. However, when we analyze and evaluate the performance of the systems, we make use of both the binary and the atomic annotation layers. Our evaluation framework is created to address our main research question (RQ 1):

RQ 1 Does the performance of a PI system on each candidate-paraphrase pair depend on the different phenomena involved in that pair?

We evaluate the performance of the systems in terms of their “*overall performance*” (Accuracy and F1) and “*phenomena performance*”.

“*Phenomena performance*” is a novelty of our approach and allows for qualitative analysis and comparison. To calculate “*phenomena performance*”, we create 27 subsets of the test set, one for each linguistic phenomenon. Each of the subsets consists of all text pairs that contain the corresponding phenomenon². Then, we use each of the 27 subsets as a test set and we calculate the binary classification Accuracy (paraphrase or non-paraphrase) for each subset. This score indicates how well the system performs in cases that include one specific phenomenon. We compare the performance of the different phenomena and also compare them with the “*overall performance*”.

Prior to running the experiments we verified that: 1) the relative distribution of the phenomena in paraphrases and in non-paraphrases is very similar; and 2) there is no significant correlation (Pearson $r < 0.1$) between the distributions of the individual phenomena. These findings show that the sub-tasks are non-trivial: 1) the binary labels of the pairs cannot be directly inferred by the presence or absence of phenomena; and 2) the different subsets of the test set are relatively independent and the performance on them cannot be trivially reduced to overlap and phenomena co-occurrence.

The “*overall performance*” and “*phenomena performance*” of a system compose its “*performance profile*”. With it we aim to address the rest of our research questions (RQs):

RQ 2 Which are the strong and weak sides of each individual system?

RQ 3 Are there any significant differences between the “*performance profiles*” of the systems?

RQ 4 Are there phenomena on which all systems perform well (or poorly)?

6.4 PI Systems

To demonstrate the advantages of our evaluation framework, we have replicated several popular Supervised and Unsupervised PI systems. We have selected the

²i.e. The “*diathesis alternation*” subset contains all pairs that contain the “*diathesis alternation*” phenomenon (such as the example pair 3a–3b). Some of the pairs can also contain multiple phenomena: the example pair 3a–3b contains both “*same polarity substitution (habitual)*” and “*diathesis alternation*”. Therefore pair 3a–3b will be added both to the “*same polarity substitution (habitual)*” and to the “*diathesis alternation*” phenomena subsets. Consequentially, the sum of all subsets exceeds the size of the test set.

systems based on three criteria: popularity, architecture, and performance. The systems that we chose are popular and widely used not only in PI, but also in other tasks. The systems use a wide variety of different ML architectures and/or different features. Finally, the systems obtain comparable quantitative results on the PI task. They have also been reported to obtain good results on the MRPC corpus which is the same size as ETPC. The choice of system allows us to best demonstrate the limitations of the classical quantitative evaluation and the advantages of the proposed qualitative evaluation.

To ensure comparability, all systems have been trained and evaluated on the same computer and the same corpus. We have used the configurations recommended in the original papers where available. During the replication we did not do a full grid-search as we want to replicate and thereby contribute to generalizable research and systems. As such, the quantitative results that we obtain may differ from the performance reported in the original papers, especially for the Supervised systems. However, the results are sufficient for the objective of this paper: to demonstrate the advantages of the proposed evaluation framework.

We compare the performance of five Supervised and five Unsupervised systems on the PI task, including one Supervised and one Unsupervised baseline systems. We also include Google BERT [Devlin et al., 2019] for reference.

The **Supervised PI systems** include:

- [S1] Machine translation evaluation metrics as hand-crafted features in a Random Forest classifier. Similar to Madnani et al. [2012] (*baseline*)
- [S2] A replication of the convolutional network similarity model of He et al. [2015]
- [S3] A replication of the lexical composition and decomposition system of Wang et al. [2016]
- [S4] A replication of the pairwise word interaction modeling with deep neural network system by He and Lin [2016]
- [S5] A character level neural network model by Lan and Xu [2018b]

The **Unsupervised PI systems** include:

- [S6] A binary Bag-of-Word sentence representation (*baseline*)
- [S7] Average over sentence of pre-trained Word2Vec word embeddings [Mikolov et al., 2013b]
- [S8] Average over sentence of pre-trained Glove word embeddings [Pennington et al., 2014]

[S9] InferSent sentence embeddings [Conneau et al., 2017]

[S10] Skip-Thought sentence embeddings [Kiros et al., 2015]

In the unsupervised setup we first represent each of the two sentences under the corresponding model. Then we obtain a feature vector by concatenating the absolute distance and the element-wise multiplication of the two representations. The feature vector is then fed into a logistic regression classifier to predict the textual relation. This setup has been used in multiple PI papers, more recently by Aldarmaki and Diab [2018]. While the vector representations of BERT are unsupervised, they are fine-tuned on the dataset. Therefore we put them in a separate category (System #11).

6.5 Results

6.5.1 Overall Performance

Table 6.1 Overall Performance of the Evaluated Systems

ID	System Description	Acc	F1
SUPERVISED SYSTEMS			
1	MTE features (baseline)	.74	.819
2	He et al. [2015]	.75	.826
3	Wang et al. [2016]	.76	.833
4	He and Lin [2016]	.76	.827
5	Lan and Xu [2018b]	.70	.800
UNSUPERVISED SYSTEMS			
6	Bag-of-Words (baseline)	.68	.790
7	Word2Vec (average)	.70	.805
8	GLOVE (average)	.72	.808
9	InferSent	.75	.826
10	Skip-Thought	.73	.816
11	Google BERT	.84	.889

Table 6.1 shows the “overall performance” of the systems on the 1725 text pairs in the test set. Looking at the table, we can observe several regularities. First, the deep systems outperform the baselines. Second, the baselines that we choose are competitive and obtain high results. Since both baselines make their predictions based on lexical similarity and overlap, we can conclude that the dataset is

biased towards those phenomena. Third, the supervised systems generally outperform the unsupervised ones, but without running a full grid-search the difference is relatively small. And finally, we can identify the best performing systems: **S3** [Wang et al., 2016] for the supervised and **S9** [Conneau et al., 2017] for the unsupervised. BERT largely outperforms all other systems.

The “*overall performance*” provides a good overview of the task and allows for a quantitative comparison of the different systems. However, it also has several limitations. It does not provide much insight into the workings of the systems and does not facilitate error analysis. In order to study and improve the performance of a system, a developer has to look at every correct and incorrect predictions and search for custom defined patterns. The “*overall performance*” is also not very informative for a comparison between the systems. For example **S3** [Wang et al., 2016] and **S4** [He and Lin, 2016] obtain the same Accuracy score and only differ by 0.06 F1 score. With only looking at the quantitative evaluation it is unclear which of these systems would generalize better on a new dataset.

6.5.2 Full Performance Profile

Table 6.2 shows the full “*performance profile*” of **S3** [Wang et al., 2016], the supervised system that performed best in terms of “*overall performance*”. Table 6.2 shows a large variation of the performance of **S3** on the different phenomena. The accuracy ranges from .33 to 1.0. We also report the statistical significance of the difference between the correct and incorrect predictions for each phenomena and the correct and incorrect predictions for the full test set, using the Mann–Whitney U-test³ [Mann and Whitney, 1947].

Ten of the phenomena show significant difference from the overall performance at $p < 0.1$. Note that eight of them are also significant at $p < 0.05$. The statistical significance of “*Opposite polarity substitution (habitual)*”, and “*Negation Switching*” cannot be verified due to the relatively low frequency of the phenomena in the test set.

The demonstrated variance in phenomena performance and its statistical significance address **RQ 1**: we show that the performance of a PI system on each candidate-paraphrase pair depends on the different phenomena involved in that pair or at least there is a strong observable relation between the performance and the phenomena.

The individual “*performance profile*” also addresses **RQ 2**. The profile is humanly interpretable, and we can clearly see how the system performs on various sub-tasks at different linguistic levels. The qualitative evaluation shows that **S3**

³The Mann–Whitney U-test is a non-parametric equivalence of T-test. The U-Test does not assume normal distribution of the data and is better suited for small samples.

Table 6.2 Performance profile of Wang et al. [2016]

OVERALL PERFORMANCE		
Overall Accuracy	.76	
Overall F1	.833	
PHENOMENA PERFORMANCE		
Phenomenon	Acc	p
Morphology-based changes		
Inflectional changes	.79	.21
Modal verb changes	.90	.01
Derivational changes	.72	.22
Lexicon-based changes		
Spelling changes	.88	.01
Same polarity sub. (habitual)	.78	.18
Same polarity sub. (contextual)	.75	.37
Same polarity sub. (named ent.)	.73	.14
Change of format	.75	.44
Lexico-syntactic based changes		
Opp. polarity sub. (habitual)	1.0	na
Opp. polarity sub. (context.)	.68	.14
Synthetic/analytic substitution	.77	.39
Converse substitution	.92	.07
Syntax-based changes		
Diathesis alternation	.83	.12
Negation switching	.33	na
Ellipsis	.64	.07
Coordination changes	.77	.47
Subordination and nesting	.86	.01
Discourse-based changes		
Punctuation changes	.87	.01
Direct/indirect style	.76	.5
Syntax/discourse structure	.83	.05
Other changes		
Addition/Deletion	.70	.05
Change of order	.81	.04
Contains negation	.78	.32
Semantic (General Inferences)	.80	.21
Extremes		
Identity	.77	.29
Non-Paraphrase	.81	.04
Entailment	.76	.5

performs better when it has to deal with: 1) surface phenomena such as “*spelling changes*”, “*punctuation changes*”, and “*change of order*”; 2) dictionary related phenomena such as “*opposite polarity substitution (habitual)*”, “*converse substitution*”, and “*modal verb changes*”. **S3** performs worse when facing phenomena such as “*negation switching*”, “*ellipsis*”, “*opposite polarity substitution (contextual)*”, and “*addition/deletion*”.

Table 6.3 Performance profiles of all systems

Phenomenon	Paraphrase Identification Systems										
	Supervised					Unsupervised					
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11
OVERALL	.74	.75	.76	.76	.70	.68	.70	.72	.75	.73	.84
Inflectional	.77	.76	.79	.79	.75	.79	.75	.76	.78	.80	.84
Modal verb	.84	.89	.90	.89	.91	.92	.89	.84	.81	.89	.92
Derivational	.80	.83	.72	.73	.84	.80	.88	.86	.80	.77	.87
Spelling	.85	.83	.88	.90	.89	.85	.89	.88	.85	.89	.94
Same pol. (hab.)	.74	.77	.78	.76	.76	.76	.76	.75	.76	.76	.85
Same pol. (con.)	.74	.74	.75	.74	.70	.71	.71	.71	.73	.73	.81
Same pol. (NE)	.74	.72	.73	.75	.64	.67	.65	.70	.73	.66	.80
Change Format	.80	.79	.75	.84	.85	.82	.81	.80	.80	.71	.91
Opp. pol. (hab.)	1.0	1.0	1.0	.50	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Opp. pol. (con.)	.77	.84	.68	.84	.52	.84	.61	.77	.65	.52	.71
Synth./analytic	.73	.73	.77	.77	.74	.70	.72	.71	.73	.74	.83
Converse sub.	.93	.93	.92	.86	.93	.86	.79	.79	.93	.79	.86
Diathesis altern.	.77	.85	.83	.77	.83	.89	.85	.83	.84	.81	.85
Negation switc	1.0	.67	.33	.33	.33	.67	.33	.67	.33	.67	.33
Ellipsis	.77	.71	.64	.74	.80	.65	.81	.74	.61	.71	.81
Coordination	.92	.92	.77	.92	.77	.92	.85	.85	.92	.92	.92
Subord. & Nest.	.83	.84	.86	.84	.81	.81	.85	.86	.80	.85	.93
Punctuation	.88	.90	.87	.87	.86	.87	.89	.89	.89	.88	.93
Direct/indirect	.84	.84	.76	.80	.76	.80	.80	.84	.80	.80	.92
Syntax/Disc.	.80	.83	.83	.81	.78	.81	.80	.80	.76	.78	.82
Add./Del.	.69	.68	.70	.72	.67	.64	.65	.66	.70	.67	.82
Change of order	.82	.83	.81	.81	.77	.82	.82	.82	.83	.84	.89
Contains neg.	.78	.74	.78	.79	.78	.72	.74	.78	.75	.76	.85
Semantic (Inf.)	.80	.89	.80	.81	.88	.90	.90	.92	.76	.79	.90
Identity	.74	.75	.77	.77	.73	.72	.73	.73	.76	.74	.85
Non-Paraphrase	.76	.77	.81	.75	.71	.55	.67	.68	.77	.79	.88
Entailment	.80	.80	.76	.76	.88	.80	.84	.88	.92	.88	.76

6.5.3 Comparing Performance Profiles

Table 6.3 shows the full performance profiles of all systems. The systems are identified by their IDs, as shown in Table 6.1. In addition to providing a better error analysis for every individual system, the “*performance profiles*” of the different systems can be used to compare them qualitatively. This comparison is much more informative than the “*overall performance*” comparison shown in Table 6.1. Using the “*performance profile*”, we can quickly compare the strong and weak sides of the different systems.

When looking at the “*overall performance*”, we already pointed out that **S3** [Wang et al., 2016] and **S4** [He and Lin, 2016] have almost identical quantitative results: 0.76 accuracy, 0.833 F1 for **S3** against 0.76 accuracy, 0.827 F1 for **S4**. However, when we compare their “*phenomena performance*” it is evident that, while these systems make approximately the same number of correct and incorrect predictions, the actual predictions and errors can vary.

Looking at the accuracy, we can see that **S3** performs better on phenomena such as “*Converse substitution*”, “*Diathesis alternation*”, and “*Non-Paraphrase*”, while **S4** performs better on “*Change of format*”, “*Opposite polarity substitution (contextual)*”, and “*Ellipsis*”.

We performed McNemar paired test comparing the errors of the two systems for each phenomena. Table 6.4 shows some of the more interesting results. Four of the phenomena with largest difference in accuracy show significant difference with $p < 0.1$. These differences in performance are substantial, considering that the two systems have nearly identical quantitative performance.

Table 6.4 Difference in phenomena performance between S3 [Wang et al., 2016] and S4 [He and Lin, 2016]

Phenomenon	#3	#4	p
Format	.75	.84	.09
Opp. Pol. Sub (con.)	.68	.84	.06
Ellipsis	.64	.74	.08
Non-Paraphrase	.81	.75	.07

We performed the same test on systems with a larger quantitative difference. Table 6.5 shows the comparison between **S3** and **S5** [Lan and Xu, 2018b]. Ten of the phenomena show significant difference with $p < 0.1$ and seven with $p < 0.05$. These results answer our **RQ 3**: we show that there are significant differences between the “*performance profiles*” of the different systems.

Table 6.5 Difference in phenomena performance: S3 [Wang et al., 2016] and S5 [Lan and Xu, 2018b]

Phenomenon	#3	#5	p
Derivational	.72	.84	.03
Same Pol. Sub (con.)	.75	.70	.02
Same Pol. Sub (NE)	.73	.64	.01
Format	.75	.85	.03
Opp. Pol. Sub (con.)	.68	.52	.10
Ellipsis	.64	.80	.10
Addition/Deletion	.70	.67	.02
Identity	.77	.73	.01
Non-Paraphrase	.81	.71	.01
Entailment	.76	.88	.08

6.5.4 Comparing Performance by Phenomena

The “*phenomena performance*” of the individual systems clearly differ among them, but they also show noticeable tendencies. Looking at the performance by phenomena, it is evident that certain phenomena consistently obtain lower than average accuracy across multiple systems while other phenomena consistently obtain higher than average accuracy.

In order to quantify these observations and to confirm that there is a statistical significance we performed Friedman-Nemenyi test [Demšar, 2006]. For each system, we ranked the performance by phenomena from 1 to 27, accounting for ties. We calculated the significance of the difference in ranking between the phenomena using the Friedman test [Friedman, 1940] and obtained a Chi-Square value of 198, which rejects the null hypothesis with $p < 0.01$. Once we had checked for the non-randomness of our results, we computed the Nemenyi test [Nemenyi, 1963] to find out which phenomena were significantly different. In our case, we compute the two-tailed Nemenyi test for $k = 27$ phenomena and $N = 11$ systems. The Critical Difference (CD) for these values is 12.5 at $p < 0.05$.

Figure 6.1 shows the Nemenyi test with the CD value. Each phenomenon is plotted with its average rank across the 11 evaluated systems. The horizontal lines connect phenomena which rank is within CD of each other. Phenomena which are not connected by a horizontal line have significantly different ranking. We can observe that each phenomenon is significantly different from at least half of the other phenomena.

We can observe that some phenomena, such as “*opposite polarity substitution (habitual)*”, “*punctuation changes*”, “*spelling*”, “*modal verb changes*”, and

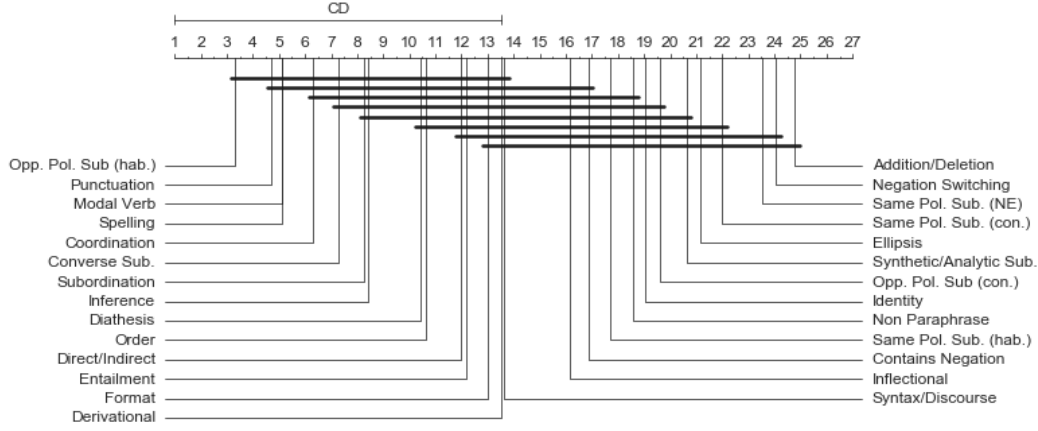


Figure 6.1: Critical Difference diagram of the average ranks by phenomena

“*coordination changes*” are statistically much easier according to our evaluation, as they are consistently among the best performing phenomena across all systems. Other phenomena, such as “*negation switching*”, “*addition/deletion*”, “*same polarity substitution (named entity)*”, “*opposite polarity substitution (contextual)*”, and “*ellipsis*” are statistically much harder, as they are consistently among the worst performing phenomena across all systems. With the exception of “*negation switching*” and “*opposite polarity substitution (habitual)*”, these phenomena occur in the corpus with sufficient frequency. These results answer our **RQ 4**: we show that there are phenomena which are easier or harder for the majority of the evaluated systems.

6.6 Discussion

In Section 6.3.2 we described our evaluation methodology and posed four research questions. The experiments that we performed and the analysis of the results answered all four of them. We briefly discuss the implications of the findings.

By addressing **RQ 1**, we showed that the performance of a system can differ significantly based on the phenomena involved in each candidate-paraphrase pair. By addressing **RQ 4**, we showed that some phenomena are consistently easier or harder across the majority of the systems. These findings empirically prove the complexity of paraphrasing and the task of PI. The results justify the distinction between the qualitatively different linguistic phenomena involved in paraphrasing and demonstrate that framing PI as a binary classification problem is an oversimplification.

By addressing **RQ 2**, we showed that each system has strong and weak sides, which can be identified and interpreted via its “*performance profile*”. This information can be very valuable when analyzing the errors made by the system or when reusing it on another task. Given the Deep architecture of the systems, such a detailed interpretation is hard to obtain via other means and metrics. By addressing **RQ 3**, we showed that two systems can differ significantly in their performance on candidate-paraphrase pairs involving particular phenomenon. These differences can be seen even in systems that have almost identical quantitative (Acc and F1) performance on the full test set. These findings justify the need for a qualitative evaluation framework for PI. The traditional binary evaluation metrics do not account for the difference in phenomena performance. They do not provide enough information for the analysis or for the comparison of different PI systems. Our proposed framework shows promising results.

Our findings demonstrate the limitations of the traditional PI task definition and datasets and the way PI systems are typically interpreted and evaluated. We show the advantages of a qualitative evaluation framework and emphasize the need to further research and improve the PI task. The “*performance profile*” also enables the direct empirical comparison of related phenomena such as “*same polarity substitution (habitual)*” and “*(contextual)*” or “*contains negation*” and “*negation switching*”. These comparisons, however, fall outside of the scope of this paper.

Our evaluation framework is not specific to the ETPC corpus or the typology behind it. The framework can be applied to other corpora and tasks, provided they have a similar format. While ETPC is the largest corpus annotated with paraphrase types to date, it has its limitations as some interesting paraphrase types (ex.: “*negation switching*”) do not appear with a sufficient frequency. We release the code for the creation and analysis of the “*performance profile*”⁴.

6.7 Conclusions and Future Work

We present a new methodology for evaluation, interpretation, and comparison of different Paraphrase Identification systems. The methodology only requires at evaluation time a corpus annotated with detailed semantic relations. The training corpus does not need any additional annotation. The evaluation also does not require any additional effort from the systems’ developers. Our methodology has clear advantages over using simple quantitative measures (Accuracy and F1 Score): 1) It allows for a better interpretation and error analysis on the individual systems; 2) It allows for a better qualitative comparison between the different

⁴https://github.com/JavierBJ/paraphrase_eval

systems; and 3) It identifies phenomena which are easy/hard to solve for multiple systems and may require further research.

We demonstrate the methodology by evaluating and comparing several of the state-of-the-art systems in PI. The results show that there is a statistically significant relationship between the phenomena involved in each candidate-paraphrase pair and the performance of the different systems. We show the strong and weak sides of each system using human-interpretable categories and we also identify phenomena which are statistically easier or harder across all systems.

As a future work, we intend to study phenomena that are hard for the majority of the systems and proposing ways to improve the performance on those phenomena. We also plan to apply the evaluation methodology to more tasks and systems that require a detailed semantic evaluation, and further test it with transfer learning experiments.

Acknowledgements

We would like to thank Darina Gold, Tobias Horsmann, Michael Wojatzki, and Torsten Zesch for their support and suggestions, and the anonymous reviewers for their feedback and comments.

This work has been funded by Spanish Ministry of Science, Innovation, and Universities Project PGC2018-096212-B-C33, by the CLiC research group (2017 SGR 341), and by the APIF grant of the first author.

Part III

Paraphrasing, Textual Entailment, and Semantic Similarity

Chapter 7

Annotating and Analyzing the Interactions between Meaning Relations

Darina Gold^{1*}, Venelin Kovatchev^{2,3*}, Torsten Zesch¹

¹Language Technology Lab, University of Duisburg-Essen, Germany

²Language and Computation Center, Universitat de Barcelona, Spain

³Institute of Complex Systems, Universitat de Barcelona, Spain

*Both authors contributed equally to this work

Published at
*Proceedings of the
Thirteenth Language Annotation Workshop, 2019*
pp.: 26-36

Abstract Pairs of sentences, phrases, or other text pieces can hold semantic relations such as paraphrasing, textual entailment, contradiction, specificity, and semantic similarity. These relations are usually studied in isolation and no dataset exists where they can be compared empirically. Here we present a corpus annotated with these relations and the analysis of these results. The corpus contains 520 sentence pairs, annotated with these relations. We measure the annotation reliability of each individual relation and we examine their interactions and correlations. Among the unexpected results revealed by our analysis is that the traditionally considered direct relationship between paraphrasing and bi-directional entailment does not hold in our data.

7.1 Introduction

Meaning relations refer to the way in which two sentences can be connected, e.g. if they express approximately the same content, they are considered paraphrases. Other meaning relations we focus on here are textual entailment and contradiction¹ [Dagan et al., 2006], and specificity.

Meaning relations have applications in many NLP tasks, e.g. recognition of textual entailment is used for summarization [Lloret et al., 2008] or machine translation evaluation [Padó et al., 2009], and paraphrase identification is used in summarization [Harabagiu and Lacatusu, 2010].

The complex nature of the meaning relations makes it difficult to come up with a precise and widely accepted definition for each of them. Also, there is a difference between theoretical definitions and definitions adopted in practical tasks. In this paper, we follow the approach taken in previous annotation tasks and we give the annotators generic and practically oriented instructions.

Paraphrases are differently worded texts with approximately the same content [Bhagat and Hovy, 2013, De Beaugrande and Dressler, 1981]. The relation is symmetric. In the following example, (a) and (b) are paraphrases.

(a) *Education is equal for all children.*

(b) *All children get the same education.*

Textual Entailment is a directional relation between pieces of text in which the information of the *Text* entails the information of the *Hypothesis* [Dagan et al., 2006]. In the following example, Text (t) entails Hypothesis (h):

(t) *All children get the same education.*

(h) *Education exists.*

Specificity is a relation between phrases in which one phrase is more precise and the other more vague. Specificity is mostly regarded between noun phrases [Cruse, 1977, Enç, 1991, Farkas, 2002]. However, there has also been work on specificity on the sentence level [Louis and Nenkova, 2012]. In the following example, (c) is more specific than (d) as it gives information on who does not get good education:

(c) *Girls do not get good education.*

(d) *Some children do not get good education.*

¹Mostly, contradiction is regarded as one of the relations within an entailment annotation.

Semantic Similarity between texts is not a meaning relation in itself, but rather a gradation of meaning similarity. It has often been used as a proxy for the other relations in applications such as summarization [Lloret et al., 2008], plagiarism detection [Alzahrani and Salim, 2010, Bär et al., 2012], machine translation [Padó et al., 2009], question answering [Harabagiu and Hickl, 2006], and natural language generation [Agirre et al., 2013]. We use it in this paper to quantify the strength of relationship on a continuous scale. Given two linguistic expressions, semantic text similarity measures the degree of semantic equivalence [Agirre et al., 2013]. For example, (a) and (b) have a semantic similarity score of 5 (on a scale from 0-5 as used in the SemEval STS task) [Agirre et al., 2013, 2014].

Interaction between Relations Despite the interactions and close connection of these meaning relations, to our knowledge, there exists neither an empirical analysis of the connection between them nor a corpus enabling it. We bridge this gap by creating and analyzing a corpus of sentence pairs annotated with all discussed meaning relations.

Our analysis finds that previously made assumptions on some relations (e.g. paraphrasing being bi-directional entailment [Madnani and Dorr, 2010, Androutsopoulos and Malakasiotis, 2010, Sukhareva et al., 2016]) are not necessarily right in a practical setting. Furthermore, we explore the interactions of the meaning relation of specificity, which has not been extensively studied from an empirical point of view. We find that it can be found in pairs on all levels of semantic relatedness and does not correlate with entailment.

7.2 Related Work

To our knowledge, there is no other work where the discussed meaning relations have been annotated separately on the same data, enabling an unbiased analysis of the interactions between them. There are corpora annotated with multiple semantic phenomena, including meaning relations.

7.2.1 Interactions between Relations

There has been some work on the interaction between some of the discussed meaning relations, especially on the relation between entailment and paraphrasing, and also on how semantic similarity is connected to the other relations.

Interaction between Entailment and Paraphrases According to Madnani and Dorr [2010], Androutsopoulos and Malakasiotis [2010], bi-directional entailment

can be seen as paraphrasing. Furthermore, according to Androutsopoulos and Malakasiotis [2010] both entailment and paraphrasing are intended to capture human intuition. Kovatchev et al. [2018a] emphasize the similarity between linguistic phenomena underlying paraphrasing and entailment. There has been practical work on using paraphrasing to solve entailment [Bosma and Callison-Burch, 2006].

Interaction between Entailment and Specificity Specificity was involved in rules for the recognition of textual entailment [Bobrow et al., 2007].

Interaction with Semantic Similarity Cer et al. [2017] argue that to find paraphrases or entailment, some level of semantic similarity must be given. Furthermore, Cer et al. [2017] state that although semantic similarity includes both entailment and paraphrasing, it is different, as it has a gradation and not a binary measure of the semantic overlap. Based on their corpus, Marelli et al. [2014] state that paraphrases, entailment, and contradiction have a high similarity score; paraphrases having the highest and contradiction the lowest of them. There also was practical work using the interaction between semantic similarity and entailment: Yokote et al. [2011] and Castillo and Cardenas [2010] used semantic similarity to solve entailment.

7.2.2 Corpora with Multiple Semantic Layers

There are several works describing the creation, annotation, and subsequent analysis of corpora with multiple parallel phenomena.

MASC The annotation of corpora with multiple phenomena in parallel has been most notably explored within the Manually Annotated Sub-Corpus (MASC) project² — It is a large-scale, multi-genre corpus manually annotated with multiple semantic layers, including WordNet senses [Miller, 1998], Penn Treebank Syntax [Marcus et al., 1993], and opinions. The multiple layers enable analyses between several phenomena.

SICK is a corpus of around 10,000 sentence pairs that were annotated with semantic similarity and entailment in parallel [Marelli et al., 2014]. As it is the corpus that is the most similar to our work, we will compare some of our annotation decisions and results with theirs.

Sukhareva et al. [2016] annotated subclasses of entailment, including *paraphrase*, *forward*, *revert*, and *null* on propositions extracted from documents on educational topics that were paired according to semantic overlap. Hence, they implicitly regarded paraphrases as a kind of entailment.

²<http://www.anc.org/MASC/About.html>

7.3 Corpus Creation

To analyze the interactions between semantic relations, a corpus annotated with all relations in parallel is needed. Hence, we develop a new corpus-creation methodology which ensures all relations of interest to be present. First, we create a pool of potentially related sentences. Second, based on the pool of sentences, we create sentence pairs that contain all relations of interest with sufficient frequency. This contrasts existing corpora on meaning relations that are tailored towards one relation only. Finally, we take a portion of the corpus and annotate all relations via crowdsourcing. This part of our methodology differs significantly from the approach taken in the SICK corpus [Marelli et al., 2014]. They don't create new corpora, but rather re-annotate pre-existing corpora, which does not allow them to control for the overall similarity between the pairs.

7.3.1 Sentence Pool

Table 7.1 List of given source sentences

Getting a high educational degree is important for finding a good job, especially in big cities.
In many countries, girls are less likely to get a good school education.
Going to school socializes kids through constant interaction with others.
One important part of modern education is technology, if not the most important.
Modern assistants such Cortana, Alexa, or Siri make our everyday life easier by giving quicker access to information.
New technologies lead to asocial behavior by e.g. depriving us from face-to-face social interaction.
Being able to use modern technologies is obligatory for finding a good job.
Self-driving cars are safer than humans as they don't drink.
Machines are good in strategic games such as chess and Go.
Machines are good in communicating with people.
Learning a second language is beneficial in life.
Speaking more than one language helps in finding a good job.
Christian clergymen learn Latin to read the bible.

In the first step, the authors create 13 sentences, henceforth *source sentences*, shown in Table 7.1. The sentences are on three topics: *education*, *technology*, and *language*. We choose sentences that can be understood by a competent speaker

without any domain-specific knowledge and which due to their complexity potentially give rise to a variety of lexically differing sentences in the next step. Then, a group of 15 people, further on called *sentence generators*, is asked to generate *true* and *false* sentences that vary lexically from the source sentence.³ Overall, 780 sentences are generated. The 13 *source sentences* are not considered in the further procedure.

For creating the *true* sentences, we ask each sentence generator to create two sentences that are true and for the *false* sentences, two sentences that are false given one source sentence. This way of generating a sentence pool is similar to that of the textual entailment SNLI corpus [Bowman et al., 2015], where the generators were asked to create true and false captions for given images. The following are exemplary true and false sentences created from one source sentence.

Source: *Getting a high educational degree is important for finding a good job, especially in big cities.*

True: *Good education helps to get a good job.*

False: *There are no good or bad jobs.*

7.3.2 Pair Generation

We combine individual sentences from the sentence pool into pairs, as meaning relations are present between pairs and not individual sentences. To obtain a corpus that contains all discussed meaning relation with sufficient frequency, we use four pair combinations:

- 1) a pair of two sentences that are true given the same source sentence (*true-true*)
- 2) a pair of two sentences that are false given the same source sentence (*false-false*)
- 3) a pair of one sentence that is true and one sentence that is false given the same source sentence (*true-false*)
- 4) a pair of randomly matched sentences from the whole sentence pool and all source sentences (*random*)

From the 780 sentences in the sentence pool, we created a corpus of 11,310 pairs, with a pair distribution as follows: 5,655 (50%) *true-true*; 2,262 (20%)

³The full instructions given to the sentence generators is included with the corpus data.

false-false, 2,262 (20%) *true-false*, and 1,131 (10%) *random*. We include all possible 5,655 *true-true* combinations of 30 true sentences for each of the 13 source sentences. For *false-false*, *true-false*, and *random* we downsample the full set of pairs to obtain the desired number, keeping an equal number of samples per source sentence. We chose this distribution because we are mainly interested in paraphrases and entailment, as well as their relation to specificity. We hypothesize that pairs of sentences that are both true have the highest potential to contain these relations.

From the 11,310 pairs, we randomly selected 520 (5%) for annotation, with the same 50-20-20-10 distribution as the full corpus. We select an equal number of pairs from each source sentence. We hypothesize that length strongly correlates with specificity, as there is potentially more information in a longer sentence than in a shorter one. Hence, for half of the pairs, we made sure that the difference in length between the two sentences is not more than 1 token.

7.3.3 Relation Annotation

We annotate all the relations in the corpus of 520 sentence pairs using Amazon Turk. We select 10 crowdworkers per task, as this gives us the possibility to measure how well the tasks has been understood overall, but especially how easy or difficult individual pairs are in the annotation of a specific relation. In the SICK corpus, the same platform and number of annotators were used.

We chose to annotate the relations separately to avoid biasing the crowdworkers who might learn heuristic shortcuts when seeing the same relations together too often. We launched the tasks consecutively to have the annotations as independent as possible. This differs from the SICK corpus annotation setting, where entailment, contradiction, and semantic similarity were annotated together.

The complex nature of the meaning relations makes it difficult to come up with a precise and widely accepted definition and annotation instructions for each of them. This problem has already been emphasized in previous annotation tasks and theoretical settings [Bhagat and Hovy, 2013]. The standard approach in most of the existing paraphrasing and entailment datasets is to use a more generic and less strict definitions. For example, pairs annotated as “paraphrases” in MRPC [Dolan et al., 2004] can have “obvious differences in information content”. This “relatively loose definition of semantic equivalence” is adopted in most empirically oriented paraphrasing corpora.

We take the same approach towards the task of annotating semantic relations: we provide the annotators with simplified guidelines, as well as with few positive and negative examples. In this way, we believe that annotation is more generic, reproducible, and applicable to any kind of data. It also relies more on the intuitions of a competent speaker than on understanding complex linguistic concepts.

Prior to the full annotation, we performed several pilot studies on a sample of the corpus in order to improve instructions and examples given to the annotators. In the following, we will shortly outline the instructions for each task.

Paraphrasing In Paraphrasing (PP), we ask the crowdworkers whether the two sentences have approximately the same meaning or not, which is similar to the definition of Bhagat and Hovy [2013] and De Beaugrande and Dressler [1981].

Textual Entailment In Textual Entailment (TE), we ask whether the first sentence makes the second sentence true. Similar to RTE Tasks [Dagan et al., 2006] - [Bentivogli et al., 2011], we only annotate for forward entailment (FTE). Hence, we use the pairs twice: in the order we ask for all other tasks and in reversed order, to get the entailment for both directions. Backward Entailment is referred to as *BTE*. If a pair contains only backward or forward entailment, it is uni-directional (UTE). If a pair contains both forward and backward entailment, it is bi-directional (BiTE). Our annotation instructions and the way we interpret directionality is similar to other crowdworking tasks for textual entailment [Marelli et al., 2014, Bowman et al., 2015].

Contradiction In Contradiction (Cont), we ask the annotators whether the sentences contradict each other. Here, our instructions are different from the typical approach in RTE [Dagan et al., 2006], where contradiction is often understood as the absence of entailment.

Specificity In Specificity (Spec), we ask whether the first sentence is more specific than the second. To annotate specificity in a comparative way is new ⁴. Like in textual entailment, we pose the task only in one direction. If the originally first sentence is more specific, it is forward specificity (FSpec), whereas if the originally second sentence is more specific than the first, it is backward specificity (BSpec).

Semantic Similarity For semantic similarity (Sim), we do not only ask whether the pair is related, but rate the similarity on a scale 0-5. Unlike previous studies [Agirre et al., 2014], we decided not to provide explicit definitions for every point on the scale.

Annotation Quality To ensure the quality of the annotations, we include 10 control pairs, which are hand-picked and slightly modified pairs from the original corpus, in each task.⁵ We discard workers who perform bad on the control pairs.⁶

⁴Louis and Nenkova [2012] labelled individual sentences as *specific*, *general*, or *cannot decide*.

⁵The control pairs are also available online at https://github.com/MeDarina/meaning_relations_interaction

⁶Only 2 annotators were discarded across all tasks. To have an equal number of annotations for each task, we re-annotated these cases with other crowdworkers.

7.3.4 Final Corpus

For each sentence pair, we get 10 annotations for each relation, namely paraphrasing, entailment, contradiction, specificity, and semantic similarity. Each sentence pair is assigned a binary label for each relation, except for similarity. We decide that if the majority (at least 60% of the annotators) voted for a relation, it gets the label for this relation.

Table 7.8 shows exemplary annotation outputs of sentence pairs taken from our corpus. For instance, sentence pair #4 contains two relations: forward entailment and forward specificity. This means that it has uni-directional entailment and the first sentence is more specific than the second. The semantic similarity of this pair is 2.7.

Inter-Annotator Agreement We evaluate the agreement on each task separately. For semantic similarity, we determine the average similarity score and the standard deviation for each pair. We also calculate the Pearson correlation between each annotator and the average score for their pairs. We report the average correlation, as suggested by SemEval [Agirre et al., 2014] and SICK.

For all nominal classification tasks we determine the majority vote and calculate the % of agreement between the annotators. This is the same measure used in the SICK corpus. Following the approach used with semantic similarity, we also calculated Cohen’s *kappa* between each annotator and the majority vote for their pairs. We report the average *kappa* for each task.⁷

Table 7.2 Inter-annotator agreement for binary relations

✓denotes a relation being there

✗denotes a relation not being there

	%	κ	%✓	%✗	control
PP	.87	.67	.83	.90	.98
TE	.83	.61	.75	.89	.89
Cont	.94	.71	.84	.95	.95
Spec	.80	.56	.81	.82	.89

Table 7.2 shows the overall inter-annotator agreement for the binary tasks. We report: 1) the average %-agreement for the whole corpus; 2) the average κ score; 3) the average %-agreement for the pairs where the majority label is “yes”; 4) the average %-agreement for the pairs where the majority label is “no”; 5) the average

⁷We are aware that κ does not fit the restrictions of our task very well and also that it is usually not averaged. However, we wanted to report a chance corrected measure, which is non-trivial in a crowd-sourcing setting, where each pair is annotated by a different set of annotators.

% agreement between the annotators and the expert-provided “control labels” on the control questions.

The overall agreement for all tasks is between .80 - .94, which is quite good given the difficulty of the tasks. Contradiction has the highest agreement with .94. It is followed by the paraphrase relation, which has an agreement of .87. The agreements of the entailment and specificity relations are slightly lower, which reflects that the tasks are more complex. SICK report agreement of .84 on entailment, which is consistent with our result.

The agreement is higher on the control questions than on the rest of the corpus. We consider it the upper boundary of agreement. The agreement on the individual binary classes shows that, except for the specificity relation, annotators have a higher agreement on the absence of relation.

Table 7.3 Distribution of Inter-annotator agreement

	50%	60%	70%	80%	90%	100%
PP	.11	.12	.13	.20	.24	.20
TE	.17	.19	.17	.16	.19	.10
Cont	.04	.07	.18	.23	.23	.25
Spec	.22	.18	.21	.13	.13	.12

Table 7.3 shows the distribution of agreement for the different relations. We take all pairs for which at least 50% of the annotators found the relation and shows what percentage of these pairs have inter-annotator agreement of 50%, 60%, 70%, 80%, 90%, and 100%. We can observe that, with the exception of contradiction, the distribution of agreement is relatively equal. For our initial corpus analysis, we discarded the pairs with 50% agreement and we only considered pairs where the majority (60% or more) of the annotators voted for the relation. However, the choice of agreement threshold an empirical question and the threshold can be adjusted based on particular objectives and research needs.

The average standard deviation for semantic similarity is 1.05. SICK report average deviation of .76, which is comparable to our result, considering that they use a 5 point scale (1-5), and we use a 6 point one (0-5). Pearson’s r between annotators and the average similarity score is 0.69 which is statistically significant at $\alpha = 0.05$.

Distribution of Meaning Relations Table 7.4 shows that all meaning relations are represented in our dataset. We have 160 paraphrase pairs, 195 textual entailment pairs, 68 contradiction pairs, and 381 specificity pairs. There is only a small number of contradictions, but this was already anticipated by the different pairings. The distribution is similar to Marelli et al. [2014] in that the set is

Table 7.4 Distribution of meaning relations within different pair generation patterns

	all	T/T	F/F	T/F	rand.
PP	31%	49%	27%	2%	6%
TE	38%	60%	36%	2%	2%
Cont.	13%	0%	10 %	56%	0%
Spec	73%	79%	72%	66%	63%
ØSim	2.27	2.90	2.39	1.32	0.77

slightly leaning towards entailment⁸. Furthermore, the distribution of uni- and bi-directional entailment with our and the SICK corpus are similar: they are nearly equally represented.⁹

Distribution of Meaning Relations with Different Generation Pairings Table 7.4 shows the distribution of meaning relations and the average similarity score in the differently generated sentence pairings. In the true/true pairs, we have the highest percentage of paraphrase (49%), entailment (60%), and specificity (79%). In the false/false pairs, all relations of interest are present: paraphrases (27%), entailment (36%), and specificity (72%). Unlike in true/true pairs, false/false ones include contradictions (10%). True/false pairs contain the highest percentage of contradiction (85%). There were also few entailment and paraphrase relations in true/false pairs. In the random pairs, there were only few relations of any kind. The proportion of specificity is high in all pairs.

This different distribution of phenomena based on the source sentences can be used in further corpus creation when determining the best way to combine sentences in pairs. In our corpus, the balanced distribution of phenomena we obtain justifies our pairing choice of 50-20-20-10.

Lexical Overlap within Sentence Pairs As discussed by Joao et al. [2007], a potential flaw of most existing relation corpora is the high lexical overlap between the pairs. They show that simple lexical overlap metrics pose a competitive baseline for paraphrase identification. Due to our creation procedure, we reduce this problem. In Table 7.5, we quantified it by calculating unigram and bigram BLEU score between the two texts in each pair for our corpus, MRPC and SNLI, which

⁸As opposed to contradiction. However, as contradiction and entailment were annotated exclusively, it is not directly comparable.

⁹In SICK 53% of the entailment is uni-directional and 46% are bi-directional, whereas we have 44% uni-directional and 55% bi-directional.

are the two most used corpora for paraphrasing and textual entailment. The BLEU score is much lower for our corpus than for MRPC and SNLI.

Table 7.5 Comparison of BLEU scores between the sentence pairs in different corpora

	MRPC	SNLI	Our corpus
unigram	61	24	18
bigram	50	12	6

Relations and Negation Our corpus also contains multiple instances of relations that involve negations and also double negations. Those examples could pose difficulties to automatic systems and could be of interest to researchers that study the interaction between inference and negation. Pairs #1, #2, and #9 in Table 7.8 are examples for pairs containing negation in our corpus.

7.4 Interactions between Relations

We analyze the interactions between the relations in our corpus in two ways. First, we calculate the correlation between the binary relations and the interaction between them and similarity. Second, we analyze the overlap between the different binary relations and discuss interesting examples.

7.4.1 Correlations between Relations

We calculate correlations between the binary relations using the Pearson correlation. For the correlations of the binary relations with semantic similarity, we discuss the average similarity and the similarity score scales of each binary relation.

7.4.1.1 Correlation of Binary Meaning Relations

In Table 7.6, we show the Pearson correlation between the meaning relations. For entailment, we show the correlation for uni-directional (UTE), bi-directional (BTE), and any-directional (TE).

Paraphrases and any-directional entailment are highly similar with a correlation of .75. Paraphrases have a much higher correlation with bi-directional entailment (.70) than with uni-directional entailment (.20). Prototypical examples of

pairs that are both paraphrases and textual entailment are pairs #1 and #2 in Table 7.8. Furthermore, both paraphrases and entailment have a negative correlation with contradiction, which is expected and confirms the quality of our data.

Specificity does not have any strong correlation with any of the other relations, showing that it is independent of those in our corpus.

Table 7.6 Correlation between all relations

	TE	UTE	BiTE	Cont	Spec	∅ Sim
PP	.75	.20	.70	-.25	-.01	3.77
TE		.57	.66	-.30	-.01	3.59
UTE			-.23	-.17	-.04	3.21
BiTE				-.20	-.01	3.89
Cont					-.09	1.45
Spec						2.27

7.4.1.2 Binary Relations and Semantic Similarity

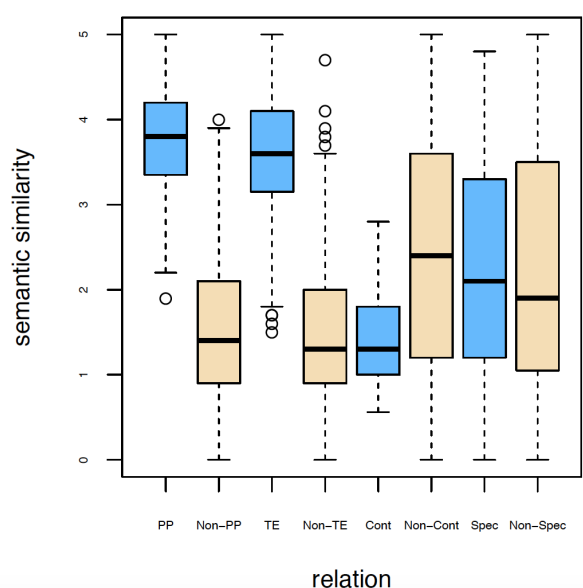


Figure 7.1: Similarity scores of sentences annotated with different relations

We look at the average similarity for each relation (see Table 7.6) and show boxplots between relation labels and similarity ratings (see Figure 7.1). Table 7.6

shows that bi-directional entailment has the highest average similarity, followed by paraphrasing, while contradiction has the lowest.

Figure 7.1 shows plots of the semantic similarity for all pairs where each relation is present and all pairs where it is absent. The paraphrase pairs have much higher similarity scores than the non-paraphrase pairs. The same observation can be made for entailment. The contradiction pairs have a low similarity score, whereas the non-contradiction pairs do not have a clear tendency with respect to similarity score. In contrast to the other relations, pairs with and without specificity do not have any consistent similarity score.

7.4.2 Overlap of Relation Labels

Table 7.7 shows the overlap between the different binary labels. Unlike Pearson correlation, the overlap is asymmetric - the % of paraphrases that are also entailment (UTE in PP) is different from the % of entailment pairs that are also paraphrases (PP in UTE). Using the overlap measure, we can identify interesting interactions between phenomena and take a closer look at some examples.

Table 7.7 Distribution of overlap within relations

	PP	UTE	BiTE	Contra	Spec
In PP		28 %	64 %	0	73 %
In UTE	52 %		-	0	73 %
In BiTE	94 %	-		0	72 %
In Contra	0	0	0		63 %
In Spec	30 %	17 %	21 %	11 %	

7.4.2.1 Entailment and Paraphrasing Overlap

In a more theoretical setting, bi-directional entailment is often defined as being paraphrases [Madnani and Dorr, 2010, Androutsopoulos and Malakasiotis, 2010, Sukhareva et al., 2016]. This implies that paraphrases equal bi-directional entailment. In our corpus, we can see that only 64% of the paraphrases are also annotated as bi-directional entailment. An example of a pair that is annotated both as paraphrase and as bi-directional entailment is pair #10 in Table 7.8. However, in the corpus we also found that 28 % of the paraphrases are only uni-directional entailment, while in 8% annotators did not find any entailment. An example of a pair where our annotators found paraphrasing, but not entailment is sentence pair #5 in Table 7.8. The agreement on the paraphrasing for this pair was 80%, the agreement on (lack of) forward and backward entailment was 80% and 70% respectively. Although the information in both sentences is nearly identical, there is

no entailment, as “having a higher chance of getting smth” does not entail “getting smth” and vice versa.

Table 7.8 Annotations of sentence pairs on all meaning relations taken from our corpus

#	Sentence 1	Sentence 2	PP	FTE	BTE	Cont	FSpec	BSpec	Sim
1	The importance of technology in modern education is overrated.	Technology is not mandatory to improve education	✓	✓	✓				2.8
2	Machines cannot interact with humans.	No machine can communicate with a person.	✓	✓	✓				4.9
3	The modern assistants make finding data slower.	Today’s information flow is greatly facilitated by digital assistants.				✓		✓	1.9
4	The bible is in Hebrew.	Bible is not in Latin.		✓			✓		2.7
5	All around the world, girls have higher chance of getting a good school education.	Girls get a good school education everywhere.	✓					✓	4.7
6	Reading the Bible requires studying Latin.	The Bible is written in Latin.		✓	✓			✓	3.6
7	Speaking more than one language can be useful.	Languages are beneficial in life.	✓	✓	✓			✓	4.4
8	You can find a good job if you only speak one language.	People who speak more than one language could only land pretty bad jobs.			✓				2.3
9	All Christian priests need to study Persian, as the Bible is written in Ancient Greek.	Christian clergymen don’t read the bible.						✓	0.9
10	School makes students antisocial.	School usually prevents children from socializing properly.	✓	✓	✓			✓	3.9

If we look at the opposite direction of the overlap, we can see that 52% of the uni-directional and 94% of the bi-directional entailment pairs are also paraphrases. This finding confirms the statement that bi-directional entailment is paraphrasing (but not vice versa).

There is also a small portion (6%) of bi-directional entailments that were not annotated as paraphrases. An example of this is pair #6 in Table 7.8. Although both sentences make each other true, they do not have the same content.

Neither paraphrasing nor entailment had any overlap with contradiction, which further verifies our annotation scheme and quality.

These findings are partly due to the more “relaxed” definition of paraphrasing adopted here. Our definition is consistent with other authors that work on paraphrasing and the task of paraphrase identification, so we argue that our findings are valid with respect to the practical applications of paraphrasing and entailment and their interactions.

7.4.2.2 Overlap with Specificity

Specificity has a nearly equal overlap within all the other relations. In the pairs annotated with paraphrase or entailment, 73% are also annotated with specificity. The high number of pairs that are in a paraphrase relation, but also have a difference in specificity is interesting, as it seems more natural for paraphrases to be on the same specificity level. One example of this is pair #7 in Table 7.8. Although they are paraphrases (with 100% agreement), the first one is more specific, as it 1) specifies the ability of speaking a language and 2) says “more than one language”.

There are also 27% of uni-directional entailment relation pairs that are not in any specificity relation. One example of this is pair #8 in Table 7.8. Although the pair contains uni-directional entailment (backward entailment), none of the sentences is more specific than the other.

If we look at the other direction of the overlap, we can observe that in 62% of the cases involving difference in specificity, there is no uni-directional nor bi-directional entailment. An example of such a relation pair is pair #9 in Table 7.8. The two sentences are on the same topic and thus can be compared on their specificity. The first sentence is clearly more specific, as it gives information on what needs to be learned and where the Bible was written, whereas the second one just gives an information on what Christian clergymen do. These findings indicate that entailment is not specificity.

7.4.3 Discussion

Our methodology for generating text pairs has proven successful in creating a corpus that contains all relations of interest. By selecting different sentence pairings, we have obtained a balance between the relations that best suit our needs.

The inter-annotator agreement was good for all relations. The resulting corpus can be used to study individual relations and their interactions. It should be emphasized that our findings strongly depend on our decisions concerning the annotations setup, the guidelines in particular. When examining the interactions between the different relations, we found several interesting tendencies.

Findings on the Interaction between Entailment and Paraphrases We showed that paraphrases and any-directional entailment had a high correlation, high overlap, and a similarly high semantic similarity. Almost all bi-directional entailment pairs are paraphrases. However, only 64% of the paraphrases are bi-directional entailment, indicating that paraphrasing is the more general phenomena, at least in practical tasks.

Findings on Specificity With respect to specificity, we found that it does not correlate with other relations, showing that it is independent of those in our corpus. It also shows no clear trend on the similarity scale and no correlation with the difference in word length between the sentences. This indicates that specificity cannot be automatically predicted using the other meaning relations and requires further study.

In the examples that we discuss, we focus on interesting cases, which are complicated and unexpected (ex.: paraphrases that are not entailment or entailment pairs that do not differ in specificity). However, the full corpus also contains many conventional and non-controversial examples.

7.5 Conclusion and Further Work

In this paper, we made an empirical, corpus-based study on interactions between various semantic relations. We provided empirical evidence that supports or rejects previously hypothesized connections in practical settings. We release a new corpus that contains all relations of interest and the corpus creation methodology to the community. The corpus can be used to further study relation interactions or as a more challenging dataset for detecting the different relations automatically¹⁰.

Some of our most important findings are:

- 1) there is a strong correlation between paraphrasing and entailment and most paraphrases include at least uni-directional entailment;
- 2) paraphrases and bi-directional entailment are not equivalent in practical settings;
- 3) specificity relation does not correlate strongly with the other relations and requires further study;
- 4) contradictions (in our dataset) are perceived as dis-similar.

As a future work, we plan to: 1) study the specificity relation in a different setting; 2) use a linguistic annotation to determine more fine-grained distinctions between the relations; 3) and annotate the rest of the 11,000 sentences in a semi-automated way.

¹⁰The full corpus, the annotation guidelines, and the control examples can be found at https://github.com/MeDarina/meaning_relations_interaction. The annotation guidelines are also available in Appendix B of the thesis.

Acknowledgements

We would like to thank Tobias Horsmann and Michael Wojatzki and the anonymous reviewers for their suggestions and comments. Furthermore, we would like to thank the sentence generators for their time and creativity. This work has been partially funded by Deutsche Forschungsgemeinschaft within the project ASSURE. This work has been partially funded by Spanish Ministry of Economy Project TIN2015-71147-C2-2, by the CLiC research group (2017 SGR 341), and by the APIF grant of the second author.

Chapter 8

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

Venelin Kovatchev¹², Darina Gold³,

M. Antònia Martí¹², Maria Salamó¹², Torsten Zesch³

¹Language and Computation Center, Universitat de Barcelona, Spain

²Institute of Complex Systems, Universitat de Barcelona, Spain

³Language Technology Lab, University of Duisburg-Essen, Germany

Accepted for publication at

Proceedings of the

Twelfth International Conference on Language Resources and Evaluation, 2020

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.

8.1 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].

¹In 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.

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, 2014, Agirre 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 children.*
 b) *Education is equal for all kids.*
- 6 a) *All children receive the same education.*
 b) *The same education is provided to all children.*

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. [2019b] empirically demonstrate that in the case of Paraphrase Identification (PI), the different “paraphrase types” are processed in a different way by automated PI systems.

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 **Single Human-Interpretable Typology for Annotating Meaning Relations** (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 8.2 lists the Related Work. Section 8.3 presents the typology, the objectives behind it and the process of selection of the types. Section 8.4 describes the annotation process - the corpus, the annotation guidelines, and the annotation interface. Section 8.5 shows the results of the annotation. Section 8.6 discusses the implications of the findings and the way our results relate to our objectives and research questions. Finally, Section 8.7 concludes the paper and addresses the future work.

8.2 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, 2014], and textual similarity [Agirre 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. [2019b] 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 compositional 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 [2018a] 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.

8.3 Shared Typology for Meaning Relations

This section is organized as follows. Section 8.3.1 presents the problem of decomposing meaning relations. Section 8.3.2 describes our proposed typology and the rationale behind it. Section 8.3.3 formulates our research questions.

8.3.1 Decomposing Meaning Relations

The goal behind the Single **H**uman-Interpretable Typology for **A**nnotating **M**eaning **R**elations (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.

8.3.2 The SHAREl Typology

Table 8.1 shows the SHAREl Typology and its 34 different types, organized in 8 categories. The first 6 categories (morphology, lexicon, lexico-syntactic, syntax,

Table 8.1 The SHARel Typology

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)

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.

The distinction between linguistic and reason-based types is introduced by Sammons et al. [2010] and Cabrio and Magnini [2014] 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 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 [2014]. 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 [2014]. 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]

²The annotation guidelines and examples for all types can be seen at <https://github.com/venelink/sharel> and in Appendix C of the thesis.

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 8.2 Comparing typologies of textual meaning relations

Typology	Relation	All	Ling.	Reason.	Hierarchy
Sammons et al. [2010]	TE, CNT	22	13	9	No
LoBue and Yates [2011]	TE, CNT	20	0	20	No
Cabrio and Magnini [2014]	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
<i>SHARel</i>	TE, CNT	34	24	8	Yes
	PP, SP, TS				

Table 8.2 lists some properties of the existing typologies of 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.

8.3.3 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.

8.4 Corpus Annotation

This section is organized as follows: Section 8.4.1 describes the corpus that we chose to use in the annotation. Section 8.4.2 presents the annotation setup. Finally, in Section 8.4.3 we report the annotation agreement.

8.4.1 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]. 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.

8.4.2 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 [2014], 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.

8.4.3 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 8.3. 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).

³The annotation guidelines and the annotated corpus are available at <https://github.com/venelink/sharel>

Table 8.3 Inter-annotator Agreement

	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

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. 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**).

8.5 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 8.5.1 analyzes and compares the frequency distribution of the different types in pairs with the following textual relations: Paraphrasing, Textual Entailment, and Contradiction. Section 8.5.2 discusses the Specificity relation and the types involved in it.

8.5.1 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 8.4 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.

Table 8.4 Type Frequencies

ID	Type	Paraph.	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 pol. sub. (habitual)	25.2 %	17 %	26 %	8.68 %
6	Same pol. sub. (contextual)	9.7 %	6.3 %	5.5 %	11.66 %
7	Same pol. sub. (named ent.)	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 pol. sub. (habitual)	2.7 %	3.5 %	7.5 %	0.07 %
10	Opposite pol. sub. (context.)	0.5 %	0.9 %	1.2 %	0.02 %
11	Synthetic/analytic sub.	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	0.25 %	0	0	2.14 %
Discourse-based changes					
18	Punctuation changes	0	0	0	3.77 %
20	Direct/indirect style altern.	0	0	0	0.30 %
21	Sentence modality changes	0	0	0	0
22	Syntax/discourse structure	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	Funct. and Mutual Exclus.	0	0.4 %	0	n/a
30	Named Ent. Reasoning	0	0	0	n/a
31	Numerical Reasoning	0	0	0	n/a
32	Temp. 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 %

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 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 [2014], 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.

8.5.2 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.

Table 8.5 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.

Table 8.5 Decomposition 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 %

8.6 Discussion

In Section 8.3, we posed two Research Questions that we wanted to address within this paper. We answered both of them in sections 8.4 and 8.5. 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. [2019b] 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.

8.7 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.

Chapter 9

Conclusions

In this final chapter, I look back at what has been done in this thesis. In Section 9.1, I summarize the main contributions of my research. I discuss the importance of my findings in the context of the research on textual meaning relations. In Section 9.2, I present a description of the different resources, created as part of this thesis and released to the community. In Section 9.3, I outline some open issues for future research on textual meaning relations and the way forward to addressing them.

9.1 Contributions and Discussion of the Results

The contributions of this thesis can be grouped into four thematic categories, corresponding to the four thesis objectives formulated in Section 1.2.

Empirical applications of paraphrase typology

The first **objective** of this thesis is *To use linguistic knowledge and paraphrase typology in order to improve the evaluation and interpretation of the automated paraphrase identification systems*. This objective has been addressed in the articles Kovatchev et al. [2018a], Kovatchev et al. [2018b], and Kovatchev et al. [2019b], presented in Chapters 4, 5, and 6.

Traditionally, the task of Paraphrase Identification is framed as a binary classification problem. It requires manually or semi-automatically annotated data for training and testing. The performance of the automated systems is evaluated using Accuracy and F1 measures. The state-of-the-art systems that work in Paraphrase Identification are mostly based on complex deep learning architectures and trained on large amounts of data. The linguistic intuitions and resources are relatively less important for these systems.

In this thesis I found evidence that the linguistic intuitions from the theoretical research on Paraphrase Typology can successfully be incorporated in the empirical task of Paraphrase Identification. The **contributions** of this part of the thesis are two empirical applications of paraphrase typology:

- A statistical corpus-based analysis presented in Chapter 5. I measured and compared the distribution of the different paraphrase types in the MRPC corpus. The analysis of the results shows that the type distribution is severely imbalanced: some paraphrase types appear in the majority of the examples, while other types are underrepresented. I hypothesize that this imbalance introduces a potential bias in the datasets.
- A “qualitative evaluation framework” presented in Chapter 6. I evaluated and compared the performance of 11 different automated paraphrase identification systems on each of the paraphrase types. The results depicted that Accuracy and F1 measures fail to capture important aspects of the performance of the automated systems. I showed that the performance of the evaluated systems varies significantly based on the types involved in each candidate-paraphrase pair. I also showed that systems with quantitatively similar performance can make qualitatively different predictions and errors.

This work gave a new perspective on the task of Paraphrase Identification. I argued that the “binary” definition of the task is oversimplified as it does not account for the different linguistic and reason-based phenomena involved in paraphrasing. The experiments showed that linguistic knowledge, in particular Paraphrase Typology, can be beneficial when analyzing the quality of the corpora and the performance of the automated systems.

In-depth knowledge about paraphrasing

The second **objective** of this thesis is *To empirically validate and quantify the difference between the various linguistic and reason-based phenomena involved in paraphrasing*. This objective has been addressed in the articles Kovatchev et al. [2018a] and Kovatchev et al. [2019b], presented in Chapters 5 and 6.

At the beginning of this dissertation, the research on Paraphrase Typology was predominantly theoretical. The different authors proposed lists of phenomena involved in the textual meaning relations and provided examples for each different type. However, there were no empirical experiments that could measure the practical implications of the theoretical difference between the paraphrase types.

In this thesis I prepared and carried out the first empirical experiment aimed at validating the theoretical concepts of the research on Paraphrase Typology. The

following **contributions** advance the research on Paraphrase Typology and provide novel, more in-depth knowledge about paraphrasing:

- A new paraphrase typology, presented in Chapter 5. I extended the existing typologies in such a way so that they could be applied both to paraphrase and non-paraphrase pairs. All paraphrase typologies prior to this thesis were only focused on texts that hold a paraphrasing relation. Extending the coverage of the typology to non-paraphrasing pairs was crucial for the empirical evaluation.
- A statistical corpus-based analysis presented in Chapter 5. I measured and compared the frequency distribution of paraphrase types in the paraphrase and non-paraphrase pairs in the ETPC corpus.
- An analysis of machine learning experiments presented in Chapter 6. I analyzed the performance of 11 different automated paraphrase identification systems. The data showed that the performance of the automated systems varies significantly based on the paraphrase types involved in each candidate paraphrase pair. These results suggest that paraphrase types are processed differently by automated paraphrase identification systems.

This thesis explored novel directions within the research on Paraphrase Typology. I presented the first empirical experiment that quantifies the difference in processing paraphrase types. The data allows to identify types that are easier or harder for the automated paraphrase identification systems. The proposed methodology is not limited to the paraphrasing textual meaning relation. It can easily be extended to other relations such as textual entailment or semantic textual similarity.

An empirical study on multiple textual meaning relations

The third **objective** of this thesis is *To empirically determine the interactions between Paraphrasing, Textual Entailment, Contradiction, and Semantic Similarity in a corpus of multiple textual meaning relations..* This objective has been addressed in the article Gold et al. [2019], presented in Chapter 7.

Textual meaning relations, such as Paraphrasing, Textual Entailment, and Semantic Textual Similarity are a popular topic within Natural Language Processing and Computational Linguistics. Traditionally, these meaning relations are studied in isolation. The research on them and the analysis of the interactions between them was very limited at the start of this dissertation.

In this thesis, I took a new look of the problem and broadened the area of study. I carried out a joint study on multiple textual meaning relations. The results

showed that it is possible to address the analysis of several meaning relations at the same time. The findings of the thesis emphasize that such analysis can benefit each individual relation. The following **contributions** enable the research in a novel direction, focused on multiple textual meaning relations:

- A new corpus creation methodology presented in Chapter 7. I proposed a novel methodology for creating a corpus that contains multiple textual meaning relations: paraphrasing, textual entailment, contradiction, textual similarity, and textual specificity.
- A new corpus presented in Chapter 7. To the best of my knowledge this is the first corpus containing pairs independently annotated for paraphrasing, textual entailment, contradiction, textual similarity, and textual specificity. Each meaning relation was annotated independently by 10 different annotators, to ensure the quality of the corpus.
- A statistical corpus analysis presented in Chapter 7. I measured and compared the frequency of each meaning relation in the corpus. I also analyzed the interactions, correlations, and overlap between the different textual meaning relations in the corpus. To the best of my knowledge, this is the first empirical comparison between paraphrasing, textual entailment, contradiction, textual similarity, and textual specificity.

The findings of this thesis have improved the understanding on important issues associated with each individual textual meaning relation and the way they interact with each other. Thanks to this study, some theoretical hypotheses and assumptions that exist in the literature have been empirically confirmed. I also reported some unexpected results:

- There is a negative statistical correlation between contradiction and textual entailment; and between contradiction and paraphrasing.
- There is a strong positive statistical correlation between uni-directional textual entailment and paraphrasing.
- In the corpus of study, paraphrasing is not equal to bi-directional textual entailment. This finding contradicts pre-existing theoretical hypotheses and assumptions.
- The data indicates that there is no statistical correlation between textual specificity and the other textual meaning relations. This also contradicts pre-existing hypothesis claiming that specificity should be strongly correlated with textual entailment.

- The analysis showed that paraphrasing, textual entailment, and contradiction have a strong statistical correlation with the degree of textual semantic similarity. Contrary to some previous studies, in my experiments pairs that contradict each other are perceived as similar.

This thesis emphasized the importance of a joint study on multiple textual meaning relations. The proposed methodology for corpus creation and analysis and the new corpus open new directions for future research.

A shared typology of textual meaning relations

The fourth **objective** of this thesis is *To propose and evaluate a novel shared typology of meaning relations. The shared typology would then be used as a conceptual framework for a joint research on meaning relations..* This objective has been addressed in the article Kovatchev et al. [2020], presented in Chapter 8.

In the recent years, several of the researchers working on paraphrasing, textual entailment, and semantic textual similarity independently argued that a single label is not sufficient to express a complex textual meaning relation. To address this problem they proposed various typologies, that is, lists of linguistic and reasoning phenomena involved in each textual meaning relation. At the beginning of this dissertation, each typology was focused on a single textual meaning relation and was not applicable to other relations.

This thesis showed that it is possible to have a single typology for multiple meaning relations. It also emphasized the advantages of a shared typology. The following **contributions** facilitate the further the research on a shared typology of textual meaning relations:

- The SHARel typology presented in Chapter 8. I propose a new typology, that is applicable to multiple textual meaning relations: paraphrasing, textual entailment, contradiction, textual specificity, and semantic similarity.
- A corpus based study presented in Chapter 8. I empirically validated the applicability of SHARel in a corpus annotation. The different meaning relations were compared in terms of the phenomena involved in each one of them. This comparison is more informative than measuring binary correlation or overlap.

This thesis has expanded the research on textual meaning relations. The SHARel typology is a step forward from the existing typologies - it is linguistically motivated and hierarchically organized. It contains both linguistic and reason-based types and has a wider coverage than any other typology. A shared typology of textual meaning relations provides some valuable insight into the workings of each

individual relation. Furthermore the shared typology can be used as a conceptual framework for in-depth comparison between the different meaning relations. It also greatly facilitates the transfer of knowledge and resources between paraphrasing, textual entailment, and semantic similarity research.

9.2 Resources

Throughout my research, I have created several language resources that I have made available to the Natural Language Processing and Computational Linguistics community. These resources are also part of the contributions of this thesis:

- The Extended Paraphrase Typology (EPT) and annotation guidelines with examples for creating a corpus annotated with EPT.
- The Single Human-Interpretable Typology for Annotating Meaning Relations (SHARel) and annotation guidelines with examples for creating a corpus annotated with SHARel.
- The WARP-Text web-based annotation interface for a fine-grained annotation of pairs of text.
- The Extended Typology Paraphrase Corpus (ETPC) - the first Paraphrase Identification corpus annotated with paraphrase types.
- The first corpus explicitly annotated with Paraphrasing, Textual Entailment, Contradiction, Semantic Textual Similarity, and Textual Specificity.
- The first corpus of multiple meaning relations, annotated with SHARel.

During my research and in collaboration with the Language Technology Lab at the University of Duisburg-Essen, I co-organized the first RELATIONS workshop¹ [Kovatchev et al., 2019a], bringing together researchers on textual meaning relations. The workshop was collocated with the 13th International Conference on Computational Semantics (IWCS) in Gothenburg, Sweden, May 23 2019.

9.3 Future Research Directions

This thesis gives two new perspectives on the research of textual meaning relations within Natural Language Processing and Computational Linguistics.

¹<https://sites.google.com/view/relations-2019>

- The importance of linguistic knowledge in the automatic processing of meaning relations.
- The presentation of the first empirical research on multiple textual meaning relations.

My work opens several new directions for future research.

In Part II of this thesis I applied paraphrase typology to the task of Paraphrase Identification. One of my findings was that the existing corpora for Paraphrase Identification is not well balanced in terms of paraphrase types. I argued that this imbalance can introduce a bias in the task and decrease the generalizability of automated PI systems. A promising research direction in this area is to work towards creating better corpora for the recognition tasks on textual meaning relations. Following what I did in my dissertation, I am currently working on a new corpus for Recognizing Textual Entailment for Spanish. The objective behind the corpus creation is to obtain a corpus more balanced in terms of certain under-represented phenomena, such as negation and named entity reasoning. My work would also help expand the research on textual meaning relations to languages other than English.

In Chapter 6 I presented the advantages of a qualitative evaluation framework over traditional measures such as Accuracy and F1. An important future research in this area would be to extend the qualitative evaluation framework to other empirical tasks on textual meaning relations, such as recognizing textual entailment or semantic textual similarity. My work in Part III of this thesis and in particular the SHARel typology is a first step towards extending the coverage of the qualitative evaluation. I showed that the typology can be applied to multiple textual meaning relations. The next step in this research direction would be the creation of a larger corpus and the corresponding software package for a general qualitative evaluation framework for textual meaning relations.

The qualitative evaluation of Paraphrase Identification systems showed that some phenomena, such as negation, ellipsis, and named entity reasoning are challenging across all evaluated systems. As a future research, I believe that each of these phenomena has to be analyzed in more detail in the context of the importance it has for the automatic processing of textual meaning relations. In a continuation of my thesis, I am currently investigating the role of negation for paraphrase identification, recognizing textual entailment, semantic textual similarity, and question answering. The preliminary results indicate that negation is extremely challenging across multiple automated systems.

Finally, my work in Part III of this thesis opens several new directions on the joint research of multiple textual meaning relations. One potential area is the creation of datasets and automated systems for the simultaneous processing of

multiple textual meaning relations. Some preliminary experiments that I carried out on the corpus presented in Chapter 7 indicate it is possible to use one meaning relation to predict the others. The next step in this research direction would be the creation of larger datasets with multiple textual meaning relations and the creation of more sophisticated automated systems.

Bibliography

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 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, 2012. Association for Computational Linguistics. URL <http://dl.acm.org/citation.cfm?id=2387636.2387697>.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. * SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 32–43, 2013.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91, 2014.

Eneko Agirre, Aitor Gonzalez-Agirre, Iñigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uria. SemEval-2016 task 2: Interpretable semantic textual similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 512–524, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1082. URL <https://www.aclweb.org/anthology/S16-1082>.

Hanan Aldarmaki and Mona Diab. Evaluation of unsupervised compositional representations. In *Proceedings of COLING 2018*, 2018.

Salha Alzahrani and Naomie Salim. Fuzzy semantic-based string similarity for extrinsic plagiarism detection. *Braschler and Harman*, 1176:1–8, 2010.

- Ion Androutsopoulos and Prodromos Malakasiotis. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38: 135–187, 2010.
- Yuki Arase and Jun’ichi Tsujii. Spade: Evaluation dataset for monolingual phrase alignment. In *Proceedings of LREC-2018*, 2018.
- Antti Arppe, Gaetanelle Gilquin, Dylan Glynn, Martin Hilpert, and Arne Zeschel. Cognitive corpus linguistics: five points of debate on current theory and methodology. *Corpora*, 5(1):1–27, 2010.
- Sören Auer, Chris Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2008. doi: doi: 10.1007/978-3-540-76298-0_52.
- Timothy Baldwin and Su Nam Kim. Multiword Expressions. *Handbook of natural language processing*, 2:267–92, 2010.
- Daniel Bär, Torsten Zesch, and Iryna Gurevych. Text reuse detection using a composition of text similarity measures. *Proceedings of COLING 2012*, pages 167–184, 2012.
- Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, and Danilo Giampiccolo. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 01 2006.
- M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.
- Marco Baroni. Composition in distributional semantics. *Language and Linguistics Compass*, 7(10):511–22, 2013.
- Marco Baroni and Alessandro Lenci. Distributional memory: A general framework for corpus-based semantics. *Comput. Linguist.*, 36(4):673–721, December 2010.
- Marco Baroni, Brian Murphy, Eduard Barbu, and Massimo Poesio. Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 34(2):222–54, 2010.

- Alberto Barrón-Cedeño, Marta Vila, M. Antònia Martí, and Paolo Rosso. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947, 2013.
- S. Bartsch. *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag, 2004.
- Vuk Batanović, Miloš Cvetanović, and Boško Nikolić. Fine-grained semantic textual similarity for serbian. In *Proceedings of LREC-2018*, 2018.
- Darina Benikova and Torsten Zesch. Same same, but different: Compositionality of paraphrase granularity levels. In *Proceedings of RANLP 2017*, 2017.
- Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of the Second Text Analysis Conference, TAC 2009, Gaithersburg, Maryland, USA, November 16-17, 2009*, 2009.
- Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The sixth PASCAL recognizing textual entailment challenge. In *Proceedings of the Third Text Analysis Conference, TAC 2010, Gaithersburg, Maryland, USA, November 15-16, 2010*, 2010.
- Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The seventh PASCAL recognizing textual entailment challenge. In *Proceedings of the Fourth Text Analysis Conference, TAC 2011, Gaithersburg, Maryland, USA, November 14-15, 2011*, 2011.
- Manuel Bertrán, Oriol Borrega, Marta Recasens, and Bàrbara Soriano. Anco-rape: A tool for multilevel annotation. *Procesamiento del Lenguaje Natural*, 41, 2008. URL <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2577/1116>.
- Rahul Bhagat. *Learning Paraphrases from Text*. PhD thesis, Los Angeles, CA, USA, 2009. AAI3368694.
- Rahul Bhagat and Eduard H. Hovy. What is a paraphrase? *Computational Linguistics*, 39(3):463–472, 2013.
- Chris Biemann and Eugenie Giesbrecht. Distributional semantics and compositionality 2011: Shared task description and results. In *Proceedings of the workshop on distributional semantics and compositionality*, pages 21–8. Association for Computational Linguistics, 2011.

- Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009. ISBN 0596516495, 9780596516499.
- Daniel Bobrow, Dick Crouch, Tracy Halloway King, Cleo Condoravdi, Lauri Karttunen, Rowan Nairn, Valeria de Paiva, and Annie Zaenen. Precision-focused textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 16–21, 2007.
- Gemma Boleda and Katrin Erk. Distributional semantic features as semantic primitives – or not, 2015. URL <https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10240/10025>.
- Wauter Bosma and Chris Callison-Burch. Paraphrase substitution for recognizing textual entailment. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 502–509. Springer, 2006.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 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, 2015.
- Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.*, 49:1–47, 2014. doi: 10.1613/jair.4135. URL <https://doi.org/10.1613/jair.4135>.
- Elena Cabrio and Bernardo Magnini. Decomposing semantic inferences, 2014.
- T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation*, 3(1):1–27, 1974.
- Julio J. Castillo and Marina E. Cardenas. Using sentence semantic similarity based on WordNet in recognizing textual entailment. In *Ibero-American Conference on Artificial Intelligence*, pages 366–375. Springer, 2010.
- Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, 2017.
- Wei-Te Chen and Will Styler. Anafora: A web-based general purpose annotation tool. In *HLT-NAACL, 2013*. URL <http://dblp.uni-trier.de/db/conf/naacl/naacl2013.html#ChenS13>.

- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *CoRR*, abs/1705.02364, 2017. URL <http://arxiv.org/abs/1705.02364>.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.
- Mathias Creutz. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL <https://www.aclweb.org/anthology/L18-1218>.
- W. Croft and D.A. Cruse. *Cognitive Linguistics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 2004. ISBN 9780521667708.
- D. Alan Cruse. The pragmatics of lexical specificity. *Journal of linguistics*, 13(2): 153–164, 1977.
- Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damjanovic, Thomas Heitz, Mark A. Greenwood, Horacio Saggion, Johann Petrak, Yaoyong Li, and Wim Peters. *Text Processing with GATE (Version 6)*. 2011. ISBN 978-0956599315. URL <http://tinyurl.com/gatebook>.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 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, 2006. Springer-Verlag. ISBN 3-540-33427-0, 978-3-540-33427-9. doi: 10.1007/11736790_9. URL http://dx.doi.org/10.1007/11736790_9.
- Robert De Beaugrande and Wolfgang U Dressler. *Introduction to text linguistics*. Routledge, 1981.
- Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=1248547.1248548>.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, June 2019. doi: 10.18653/v1/N19-1423. URL <https://www.aclweb.org/anthology/N19-1423>.
- Bill Dolan, Chris Quirk, and Chris Brockett. 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 2004. COLING.
- William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL <https://www.aclweb.org/anthology/I05-5002>.
- Marie Dubremetz and Joakim Nivre. Extraction of Nominal Multiword Expressions in French. *EACL 2014*, page 72, 2014.
- Cecily Jill Duffield, Jena D Hwang, and Laura A Michaelis. Identifying assertions in text and discourse: the presentational relative clause construction. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 17–24. Association for Computational Linguistics, 2010.
- Mürvet Enç. The semantics of specificity. *Linguistic inquiry*, pages 1–25, 1991.
- Katrin Erk. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653, 2012.
- Cristina España Bonet, Marta Vila Rigat, Horacio Rodríguez, and Antonia Martí. Coco, a web interface for corpora compilation. 2009.
- Stefan Evert. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2:223–33, 2008.
- Meghdad Farahmand and Ronaldo Martins. A Supervised Model for Extraction of Multiword Expressions Based on Statistical Context Features. *EACL 2014*, page 10, 2014.
- Donka F. Farkas. Specificity distinctions. *Journal of semantics*, 19(3):213–243, 2002.
- Samuel Fernando and Mark Stevenson. A semantic similarity approach to paraphrase detection. *Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloquium*, 2008.

- Charles J Fillmore, Russell Lee-Goldman, and Russell Rhodes. The Framenet constructicon. *Sign-based Construction Grammar. CSLI, Stanford, CA*, 2012.
- Andrew Finch, Young-Sook Hwang, and Eiichiro Sumita. Using machine translation evaluation techniques to determine sentence-level semantic equivalence. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL <https://www.aclweb.org/anthology/I05-5003>.
- J. R. Firth. A synopsis of linguistic theory 1930-55. 1952-59:1-32, 1957.
- Markus Forsberg, Richard Johansson, Linnéa Bäckström, Lars Borin, Benjamin Lyngfelt, Joel Olofsson, and Julia Prentice. From construction candidates to constructicon entries. an experiment using semi-automatic methods for identifying constructions in corpora. *Constructions and Frames*, 6(1):114-35, 2014. ISSN 1876-1933.
- Marc Franco-Salvador, Rangel Francisco, Rosso Paolo, Taulé Mariona, and Martí M. Antónia. Language variety identification using distributed representations of words and documents. In *Proceedings of the 6th International Conference of CLEF on Experimental IR meets Multilinguality, Multimodality and Interaction*, Lectures Notes in Computer Science. Springer Verlag, 2015.
- M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86-92, March 1940.
- Pablo Gamallo, Alexandre Agustini, and Gabriel P Lopes. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107-146, 2005.
- Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *HLT-NAACL*, pages 758-764. The Association for Computational Linguistics, 2013.
- Konstantina Garoufi. Towards a better understanding of applied textual entailment: Annotation and evaluation of the rte-2 dataset. Master’s thesis, Saarland University, September 2007.
- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL@ACL 2007 Workshop on Textual Entailment and Paraphrasing, Prague, Czech Republic, June 28-29, 2007*, pages 1-9, 2007.

Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magnini, Ido Dagan, Elena Cabrio, and Bill Dolan. The fourth PASCAL recognizing textual entailment challenge. In *Proceedings of the First Text Analysis Conference, TAC 2008, Gaithersburg, Maryland, USA, November 17-19, 2008*, 2008.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 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 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2103. URL <https://www.aclweb.org/anthology/P18-2103>.

Darina Gold, Venelin Kovatchev, and Torsten Zesch. Annotating and analyzing the interactions between meaning relations. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 26–36, Florence, Italy, August 2019. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W19-4004>.

A. E. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. Cognitive Theory of Language and Culture. University of Chicago Press, 1995. ISBN 9780226300863.

A. E. Goldberg. *Constructions at work*. Oxford University Press, 2006.

Adele E Goldberg. Argument structure constructions versus lexical rules or derivational verb templates. *Mind & Language*, 28(4):435–65, 2013.

Stefan Th. Gries and Nich C. Ellis. Statistical measures for usage-based linguistics. *Language Learning*, (65):1–28, 2015.

Stefan Th. Gries, Beate Hampe, and Doris Schönefeld. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, (16):635–76, 2005.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL <https://www.aclweb.org/anthology/N18-2017>.

- Sanda Harabagiu and Andrew Hickl. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 905–912. Association for Computational Linguistics, 2006.
- Sanda Harabagiu and Finley Lacatusu. Using topic themes for multi-document summarization. *ACM Transactions on Information Systems (TOIS)*, 28(3):13, 2010.
- Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.
- Hua He and Jimmy Lin. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2016.
- Hua He, Kevin Gimpel, and Jimmy Lin. Multi-perspective sentence similarity modeling with convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1576–1586. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1181. URL <http://aclweb.org/anthology/D15-1181>.
- Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval*, page 33, USA, 2010. Association for Computational Linguistics.
- Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 2015.
- Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL, page 873, USA, 2012. Association for Computational Linguistics.
- Jena D Hwang, Rodney D Nielsen, and Martha Palmer. Towards a domain independent semantics: Enhancing semantic representation with construction grammar. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 2010.

Shankar Iyer, Nikhil Dandekar, and Kornl Csernai. First quora dataset release: Question pairs, 2017.

Yangfeng Ji and Jacob Eisenstein. Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 891–896, 2013. URL <http://aclweb.org/anthology/D/D13/D13-1090.pdf>.

Cordeiro Joao, Dias Gaël, and Brazdil Pavel. New functions for unsupervised asymmetrical paraphrase detection. *Journal of Software*, 2(4):12–23, 2007.

George Karypis. CLUTO a clustering toolkit. Technical Report 02-017, Dept. of Computer Science, University of Minnesota, 2002.

K. Kesselmeier, T. Kiss, A. Müller, C. Roch, T. Stadfeld, and J. Strunk. Mining for preposition-noun constructions in german. In *Workshop on Extracting and Using Constructions in Natural Language Processing*, NODALIDA 2009, 2009.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc., 2015. URL <http://papers.nips.cc/paper/5950-skip-thought-vectors.pdf>.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. Domain agnostic real-valued specificity prediction. In *AAAI*, 2019.

Venelin Kovatchev, Maria Salamó, and M. Antònia Martí. Comparing distributional semantics models for identifying groups of semantically related words. *Procesamiento del Lenguaje Natural*, 57:109–116, 2016.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. Etpc - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of LREC-2018*, 2018a.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 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 2018b. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/C18-2029>.
- Venelin Kovatchev, Darina Gold, and Torsten Zesch, editors. *RELATIONS - Workshop on meaning relations between phrases and sentences*, Gothenburg, Sweden, May 2019a. Association for Computational Linguistics. URL <https://aclanthology.org/W19-0800>.
- Venelin Kovatchev, M. Antonia Marti, Maria Salamo, and Javier Beltran. A qualitative evaluation framework for paraphrase identification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 568–577, Varna, Bulgaria, September 2019b. INCOMA Ltd. doi: 10.26615/978-954-452-056-4_067. URL <https://www.aclweb.org/anthology/R19-1067>.
- Venelin Kovatchev, Darina Gold, M. Ant3nia Marti, Maria Salamo, and Torsten Zesch. Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA), 2020. ISBN 979-10-95546-00-9.
- Zornitsa Kozareva and Andr3s Montoyo. Paraphrase identification on the basis of supervised machine learning techniques. In *Proceedings of the 5th International Conference on Advances in Natural Language Processing, FinTAL, 2006*, page 524, 533, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540373349. doi: 10.1007/11816508_52. URL https://doi.org/10.1007/11816508_52.
- Gerhard Kremer, Katrin Erk, Sebastian Pad3, and Stefan Thater. What substitutes tell us - analysis of an “all-words” lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1057. URL <https://www.aclweb.org/anthology/E14-1057>.
- Wuwei Lan and Wei Xu. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of COLING 2018*, 2018a.
- Wuwei Lan and Wei Xu. Character-based neural networks for sentence pair modeling. In *Proceedings of the 2018 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2018b.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 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, 2017.

T.K. Landauer, D.S. McNamara, S. Dennis, and W. Kintsch. *Handbook of Latent Semantic Analysis*. University of Colorado Institute of Cognitive Science Series. Lawrence Erlbaum Associates, 2007. ISBN 9780805854183.

Gabriella Lapesa and Stefan Evert. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2(0):531–545, 2014. ISSN 2307-387X. URL <https://transacl.org/ojs/index.php/tacl/article/view/457>.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *Rivista di Linguistica*, 20(1):1–31, 2008.

Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

Ran Levy, Liat Ein-Dor, Shay Hummel, Ruty Rinott, and Noam Slonim. TR9856: A multi-word term relatedness benchmark. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 419–424, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2069. URL <https://www.aclweb.org/anthology/P15-2069>.

Dekang Lin and Patrick Pantel. Dirt@ sbt@ discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–8. ACM, 2001.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. URL <http://aclweb.org/anthology/Q16-1037>.

Elena Lloret, Oscar Ferrández, Rafael Munoz, and Manuel Palomar. A Text Summarization Approach under the Influence of Textual Entailment. In *NLPCS*, pages 22–31, 2008.

- Peter LoBue and Alexander Yates. 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, 2011. Association for Computational Linguistics. ISBN 978-1-932432-88-6. URL <http://dl.acm.org/citation.cfm?id=2002736.2002805>.
- Annie Louis and Ani Nenkova. A corpus of general and specific sentences from news. In *LREC*, pages 1818–1821, 2012.
- Nitin Madnani and Bonnie J Dorr. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387, 2010.
- Nitin Madnani, Joel Tetreault, and Martin Chodorow. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 182–190, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. ISBN 978-1-937284-20-6. URL <http://dl.acm.org/citation.cfm?id=2382029.2382055>.
- H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60, 03 1947. doi: 10.1214/aoms/1177730491. URL <https://doi.org/10.1214/aoms/1177730491>.
- Michał Marcińczuk, Marcin Oleksy, and Jan Kočoń. Inforex - a collaborative system for text corpora annotation and analysis. In *Proceedings of RANLP-2017*, September 2017. URL https://doi.org/10.26615/978-954-452-049-6_063.
- Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–30, 1993.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, Roberto Zamparelli, et al. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223, 2014.
- Maria Antònia Martí, Mariona Taulé, Venelin Kovatchev, and Maria Salamó. Discover: Distributional approach based on syntactic dependencies for discovering constructions. *Corpus Linguistics and Linguistic Theory*, 2019.

- Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI, 2006, page 775–780. AAAI Press, 2006. ISBN 9781577352815.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’13, pages 3111–3119, USA, 2013b. Curran Associates Inc. URL <http://dl.acm.org/citation.cfm?id=2999792.2999959>.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751, 2013c.
- George Miller. *WordNet: An electronic lexical database*. MIT press, 1998.
- George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995. ISSN 0001-0782.
- Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1439, 2010.
- Hermann Moisl. *Cluster Analysis for Corpus Linguistics*. De Gruyter Mouton, 2015.
- K. Muischnek and H. Sajkan. Using collocation-finding methods to extract constructions and estimate their productivity. In *Workshop on Extracting and Using Constructions in Natural Language Processing*, NODALIDA 2009, 2009.
- Brian Murphy, Partha Pratim Talukdar, and Tom M Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. In *COLING*, pages 1933–50, 2012.
- Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 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 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/C18-1198>.

- Vivi Nastase, Devon Fritz, and Anette Frank. Demodify: A dataset for analyzing contextual constraints on modifier deletion. In *Proceedings of LREC-2018*, 2018.
- Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December 2012. ISSN 0004-3702.
- P.B. Nemenyi. *Distribution-free Multiple Comparisons*. PhD thesis, Princeton University, 1963.
- Yoshiki Niwa and Yoshihiko Nitta. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th Conference on Computational Linguistics*, volume 1 of *COLING '94*, pages 304–309, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.
- Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. Idioms. *Language*, pages 491–538, 1994.
- Matthew Brook O'Donnell and Nick Ellis. Towards an inventory of english verb argument constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, EUCCCL '10, pages 9–16, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D Manning. Textual entailment features for machine translation evaluation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 37–41. Association for Computational Linguistics, 2009.
- Lluís Padró and Evgeny Stanilovsky. Freeling 3.0: Towards wider multilinguality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *LREC*, pages 2473–9. European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7.
- Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2070. URL <https://www.aclweb.org/anthology/P15-2070>.

- Pavel Pecina. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44:137–58, 2010. ISSN 1574-020X.
- Anselmo Peñas, Álvaro Rodrigo, and Felisa Verdejo. Sparte, a test suite for recognising textual entailment in spanish. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 275–286, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32206-1.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018. URL <http://arxiv.org/abs/1802.05365>. cite arxiv:1802.05365Comment: NAACL 2018. Originally posted to openreview 27 Oct 2017. v2 updated for NAACL camera ready.
- Carlos Ramisch, Aline Villavicencio, and Christian Boitet. Multiword expressions in the wild?: the mwetoolkit comes in handy. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 57–60. Association for Computational Linguistics, 2010.
- Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer Berlin Heidelberg, 2002.
- Mark Sammons, V. G. Vinod Vydiswaran, and Dan Roth. "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, 2010.
- Federico Sangati and Andreas van Cranenburgh. Multiword expression identification with recurring tree fragments and association measures. In *Proceedings of NAACL-HLT*, pages 10–18, 2015.
- Ekaterina Shutova, Lin Sun, and Anna Korhonen. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics, 2010.
- Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Sridhar Narayanan. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123, 2017.

- Vered Shwartz and Ido Dagan. Adding context to semantic data-driven paraphrasing. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 108–113, Berlin, Germany, August 2016. Association for Computational Linguistics.
- Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, and Christopher D. Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proceedings of the 24th International Conference on Neural Information Processing Systems, NIPS*, page 801, Red Hook, NY, USA, 2011. Curran Associates Inc. ISBN 9781618395993.
- Robyn Speer and Catherine Havasi. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/1072_Paper.pdf.
- Anatol Stefanowitsch and Stefan Th. Gries. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2):209 – 43, 2003.
- Anatol Stefanowitsch and Stefan Th. Gries. Corpora and grammar. *Corpus Linguistics*, 2008.
- Maria Sukhareva, Judith Eckle-Kohler, Ivan Habernal, and Iryna Gurevych. Crowdsourcing a Large Dataset of Domain-Specific Context-Sensitive Semantic Verb Relations. In *LREC*, 2016.
- Assaf Toledo, Stavroula Alexandropoupou, Sophie Chesney, Sophia Katrenko, Heidi Klockmann, Pepijn Kokke, Benno Kruit, and Yoad Winter. Towards a semantic model for textual entailment. In Cleo Condoravdi, Valeria de Paiva, and Annie Zaenen, editors, *Linguistic Issues in Language Technology vol. 9*. 2014.
- Michael Tomasello. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1-2):61–82, 2000.
- Peter D Turney. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research (JAIR)*, 33:615–55, 2008.
- Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January 2010.

Elena Tutubalina. Clustering-based approach to multiword expression extraction and ranking. In *Proceedings of NAACL-HLT*, pages 39–43, 2015.

M. Vila, M. A. Martí, and H. Rodríguez. "is this a paraphrase? what kind? paraphrase boundaries and typology. ". pages 205–218, 2014.

Marta Vila, Manuel Bertran, M. Antònia Martí, and Horacio Rodríguez. Corpus annotation with paraphrase types: new annotation scheme and inter-annotator agreement measures. *Language Resources and Evaluation*, 49(1):77–105, 2015. ISSN 1574-0218.

Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st International Conference on World Wide Web, WWW '12 Companion*, page 1063, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312301. doi: 10.1145/2187980.2188242. URL <https://doi.org/10.1145/2187980.2188242>.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 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 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL <https://www.aclweb.org/anthology/D19-1221>.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop Black-boxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL <https://www.aclweb.org/anthology/W18-5446>.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32*, pages 3261–3275. Curran Associates, Inc., 2019. URL <http://papers.nips.cc/paper/8589-superglue-a-stickier-benchmark-for-general-purpose-language.pdf>.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. Sentence similarity learning by lexical decomposition and composition. *CoRR*, abs/1602.07019, 2016. URL <http://arxiv.org/abs/1602.07019>.

David Wible and Nai-Lung Tsao. StringNet As a Computational Resource for Discovering and Investigating Linguistic Constructions. In *Proceedings of the NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, EUCCCL '10, pages 25–31, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL <https://www.aclweb.org/anthology/N18-1101>.

Ludwig Wittgenstein. *Philosophical Investigations*. (Translated by Anscombe, G.E.M.). Basil Blackwell, 1953.

Alison Wray and Mick Perkins. The functions of formulaic language: an integrated model. *Language and Communication*, 20(1):1–28, 2000.

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

Seid Muhie Yimam and Chris Biemann. Par4Sim – adaptive paraphrasing for text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 331–342, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/C18-1028>.

Seid Muhie Yimam and Iryna Gurevych. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *In Proceedings of ACL-2013 System Demonstrations*, pages 1–6, 2013.

Ken-ichi Yokote, Shohei Tanaka, and Mitsuru Ishizuka. Effects of Using Simple Semantic Similarity on Textual Entailment Recognition. In *TAC*, 2011.

Willem Zuidema. What are the productive units of natural language grammar?: a dop approach to the automatic identification of constructions. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 29–36. Association for Computational Linguistics, 2006.

Appendix A

Annotation Guidelines for ETPC

A.1 Presentation

This document sets out the guidelines for the paraphrase typology annotation task, using the Extended Paraphrase Typology. The task consists of annotating candidate paraphrase pairs (including positive and negative examples of paraphrasing) with a textual paraphrase label, the paraphrase types they contain, and negation. These guidelines have been used to annotate the Microsoft Research Paraphrase Corpus (MRPC), thus giving rise to the Extended Typology Paraphrase Corpus (ETPC). For the purpose of the annotation, we have developed a web based annotation tool, the WARP-Text interface.

This document is divided in five blocks: general considerations about the task and theoretical definitions (Section A.2); tagset definition (section A.3); guidelines for annotating non-paraphrases (section A.4); annotating negation (section A.5).

Marks and symbols used in this document:

- Fragments in the examples that should be annotated are underlined.
When no fragment is underlined, it means that it is the whole example that should be tagged.
- The so-called “key elements” are in **bold**.

A.1.1 Credits

This document has been adapted and extended from the paraphrase typology annotation guidelines of Vila and Marti (2012).

A.2 The task

Paraphrasing stands for sameness of meaning between different wordings. For example, the pair of sentences in (a) are different in form but have the same meaning. Our **paraphrase typology (ETPC)** classifies paraphrases according to the linguistic nature of this difference in wording.

- a) John said “I like candies”/John said that he liked sweets.
- b) John said “I like candies”/John said that he liked onion.

The task described in these guidelines consists of annotating a Paraphrase Identification corpus (MRPC) with the Extended Paraphrase Typology (EPT). A Paraphrase Identification corpus contains textual paraphrase pairs (ex.: (a)), as well as textual non-paraphrase pairs (ex.: (b)). Our annotation task consists of two sub-tasks:

Annotating atomic paraphrases within textual paraphrase pairs (a) and textual non-paraphrase pairs (b). The textual pairs are generally complex in the sense that they contain multiple atomic paraphrases. We call these atomic paraphrases paraphrase phenomena and they are what should be annotated with the typology. The paraphrase pair in (a) contains two paraphrase phenomena: the direct/indirect style alternation and a synonymy substitution.

Annotating atomic non-paraphrases within textual non-paraphrase pairs (b). The non-paraphrase pair in (b) contains one atomic non-paraphrase: the substitution of “candies” with “onion”.

In the annotation process, three main decisions should be made:

- 1) determine whether a candidate pair is a textual paraphrase (Section A.2.1)
- 2) If **non-paraphrase**, determine the key differences between the two texts:
 - choose the tag that best describes the phenomenon behind each difference (Section A.2.2)
 - determine the scope of every atomic non-paraphrase (Section A.2.3)
- 3) Determine the similarities between the two texts:
 - choose the tag that best describes the phenomenon behind each similarity (Section A.2.2)
 - determine the scope of every atomic paraphrase (Section A.2.3)

A.2.1 Is This a Paraphrase Pair

The first step in the annotation process is determining whether a candidate paraphrase pair is actually a paraphrase. We consider paraphrases those pairs having

the same or an equivalent propositional content. We consider non-paraphrases those pairs that have substantial difference in the propositional content. For example, a) will be annotated as “paraphrases”, while b) will be annotated as “non-paraphrases”.

- a) Amrozi accused his brother, whom he called "the witness", of deliberately distorting his evidence.
Referring to him as only "the witness", Amrozi accused his brother of deliberately distorting his evidence.
- b) Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for \$2.5 billion.
Yucaipa bought Dominick's in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998.

Since the Extended Paraphrase Typology (ETP) can annotate atomic paraphrases (similarities) as well as atomic non-paraphrases (dissimilarities), both textual paraphrases and textual non-paraphrases will be subsequently annotated with the paraphrase typology. The subsequent annotation with paraphrase types will allow for distinguishing between paraphrase and non-paraphrase fragments within these sentences.

A.2.2 The Tagset

Our tagset is based on the Extended Paraphrase Typology shown in Table A.1. It is organized in seven meta categories: “Morphology”, “Lexicon”, “Lexico-syntax”, “Syntax”, “Discourse”, “Other”, and “Extremes”. Sense Preserving (Sens Pres.) shows whether a certain type can give rise to textual paraphrases (+), to textual non-paraphrases (-), or to both (+ / -). The typology contains 25 atomic paraphrase types (+) and 13 atomic non-paraphrase types (-).

The subclasses (morphology, lexicon, syntax and discourse based changes) follow the classical organisation in formal linguistic levels from morphology to discourse. Our paraphrase types are grouped in classes according to the nature of the underlying linguistic mechanism: (i) those types where the paraphrase arises at the morpho-lexicon level, (ii) those that are the result of a different structural organization and (iii) those types arising at the semantics level. Although the class stands for the trigger change, paraphrase phenomena in each class can entail changes in other parts of the sentence. For instance, a morpho-lexicon based change (derivational) like the one in (a), where the verb *failed* is exchanged for its nominal form *failure*, has obvious syntactic implications; however, the paraphrase is triggered by the morphological change. A structure based change (diathesis)

like the one in (b) entails an inflectional change in *hear/was heard* among others. Finally, paraphrases in semantics are based on a different distribution of semantic content across the lexical units with, on many occasions, a complete change in the form (c).

Table A.1 Extended Paraphrase Typology

ID	Type	Sense Pres.
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	Coordination changes	+
17	Subordination and nesting changes	+
Discourse-based changes		
18	Punctuation changes	+
19	Direct/indirect style alternations	+ / -
20	Sentence modality changes	+
21	Syntax/discourse structure changes	+
Other changes		
22	Addition/Deletion	+ / -
23	Change of order	+
24	Semantic (General Inferences)	+ / -
Extremes		
25	Identity	+
26	Non-Paraphrase	-
27	Entailment	-

- a) how the headmaster failed / the failure of the headmaster
- b) We were able to hear the report of a gun on shore intermittently / the report of a gun on shore was still heard at intervals
- c) I'm guessing we won't be done for some time / I've got a hunch that we 're not through with that game yet

Miscellaneous changes comprise types not directly related to one single class. Finally, in paraphrase extremes, two special cases of paraphrase phenomena should be considered: they consist of the extremes of the paraphrase continuum, which goes from the highest level of paraphrasability (identity) to the lowest limits of the paraphrase phenomenon (entailment). Non-paraphrase fragments within paraphrase pairs are also part of the class paraphrase extremes.

As some of the names of our types explicitly reflect (e.g. ADDITION / DELETION), they are **bidirectional**: in a paraphrase pair, they can be applied from the first member of the pair to the second and vice versa.

ETP contains both "sense preserving" atomic phenomena (atomic paraphrases) and "non sense preserving" atomic phenomena (atomic non-paraphrases). While some phenomena are considered to (almost) always preserve the meaning (ex.: abbreviation, habitual same polarity substitution), other phenomena are not innately preserving the meaning and can lead both to paraphrasing and to non-paraphrasing at the textual level (ex.: In (d) and (e) the involved phenomena is the same - "inflectional change", however in (d) the two texts are paraphrases, while in (e) they are not). The "sense preserving" feature is required for the annotation of the "non-paraphrases".

- d It was with difficulty that the course of streets could be followed.
You couldn't even follow the path of the street.
- e You can't travel from Barcelona to Mallorca with the boat.
Boats can't travel from Barcelona to Mallorca.

A.2.3 The Scope

The scope refers to the selection of the tokens to be annotated within each tag. In what follows, we first define the type of units we are willing to annotate (Section A.2.3.1), the criteria followed in the scope selection (Section A.2.3.2) and when the punctuation marks should be included (Section A.2.3.3).

A.2.3.1 Kind of Units to Be Annotated

We annotate **linguistic units**, not strings that do not correspond to a full linguistic unit. These linguistic units can go from the word to the (multiple-)sentence level.

In the paraphrase pair in (a), although a change takes place between the snippets *here by* and *it is there in*, two paraphrase mappings have to be established between *here* and *there* (1), and *by virtue of* and *in virtue of* (2), two different pairs of linguistic units.

- a) Here₁ by virtue of₂ humanity's vestures.
It is there₁ in virtue of₂ the vesture of humanity in which it is clothed.

However, selecting full linguistic units is not always possible or adequate from the paraphrase annotation point of view. In the following, we set out some exceptions to the above rule:

1. Cases in which only one member of the paraphrase pair corresponds to a linguistic unit. In (b), a SEMANTICS BASED CHANGE occurs between the underlined fragments. In the first sentence, it consists in a full linguistic unit, namely a causal clause; in the second sentence, the semantic content in the first appears divided into a nominal phrase and part of a verbal phrase, i.e., the verb *has impressed*. This nominal phrase plus the verb, although they do not constitute a full linguistic unit, are the scope of the phenomenon in the second sentence

- b) There is a pattern of regularity and order in the entire cosmos, due to some hints that science provides us.
A presiding mind has impressed the stamp of order and regularity upon the whole cosmos.

2. Cases in which none of the members of the paraphrase pair correspond to a linguistic unit. The prototypical example of this situation are contractions, within the SPELLING tag. In (c), *I* constitutes a nominal phrase and *will* is part of a verbal phrase. As the contraction is produced between these two pieces, they and only they constitute the scope of the phenomenon.

- c) I will go to the cinema.
I'll go to the cinema.

3. Cases of identical (see Section A.2.3.2)

A.2.3.2 Scope Annotation Criteria

The way the scope should be annotated depends on the class of the tag. Three criteria should be followed:

1. Morpho-lexicon based changes, semantics based changes and miscellaneous changes: only the linguistic units affected by the trigger change are tagged.

- a) I dislike rash motorists .
I dislike rash drivers .
- b) He rarely makes us smile .
He has little to do with making us smile .

2. Structure based changes: the whole linguistic unit suffering the syntactic or discourse reorganization is tagged (light green rectangle in Figure 2). If the reorganization takes place within a phrase, the phrase is tagged. If the reorganization takes place within a clause, the clause is tagged. If the reorganization takes place within a sentence, the sentence is tagged. If the reorganization takes place between different phrases/clauses/sentences (mainly coordination and subordination phenomena), all and only the phrases/clauses/sentences affected are tagged. In the case of clause changes, if the reorganizations takes place within the subordinate clause, only this one is annotated (not the main clause) and vice versa.

Moreover, all structure based changes (except from diathesis alternations) have a **key element** that gives rise to the change and/or distinguishes it from others. This key element is also annotated. First, the whole linguistic unit (including the key element) is tagged, and then the key element is annotated independently.

In (d), an active/passive alternation takes place (DIATHESIS tag). As the change takes place within the subordinate clause, only this clause is tagged. In (e), a change in the subordination form takes place (SUBORDINATION & NESTING tag). As the change affects the way the two clauses (the main and the subordinate) are connected, the whole sentence is tagged. The connective mechanisms (the conjunction and the gerund clause) are annotated as key elements.

- d) When she sings that song, everything seems possible.
When that song is sang, everything seems possible.
- e) **When** we hear that song, everything seems possible.
Hearing that song, everything seems possible.

3. Entailment and non-paraphrase tags: the affected linguistic unit is tagged. The example in (f) is a case of ENTAILMENT; the example in (g) is a NON-PARAPHRASE.

- f) Google was in talks to buy YouTube.
Google bought YouTube.

- g) Mary and Wendy went to the cinema .
Mary and Wendy like each other .

4. Identical tag: Once all other phenomena are annotated, snippets which are identical in both sentences may remain. We should annotate as IDENTICAL these **snippet (not linguistic unit)** residues (h). In this case, we do not follow the linguistic unit criteria (Section A.2.3.1).

Only one (discontinuous) identical tag will be used in each pair of sentences.

Punctuation marks will also be annotated as IDENTICAL if they effectively are.

- h) The two argued that only a new board would have had the credibility to restore El Paso to health.
The two believed that only a new board would have had the credibility to restore El Paso to health.

Finally, it should be noted that tags overlap on many occasions. In (i), a SAMEPOLARITY tag overlaps an ORDER one.

- i) shaking his head wisely .
sagely shaking his head.

A.2.3.3 Should Punctuation Marks Be Included?

When a whole phrase/cause/sentence is annotated, **the closing (and opening) punctuation mark (if any) is(are) included**. Some examples are (a) and (b), which are cases of DIATHESIS and ADDITION/ELETION, respectively. In contrast, in (c) and (d), the commas are not included as they are not the opening and closing punctuation marks of the paraphrase phenomenon tagged (SAMEPOLARITY), but of a bigger unit.

- a) This song (John sang it last year in the festival) will be a great success.
This song (it was sung by John last year in the festival) will be a great success.
- b) His judgment have kept equal pace in that conclusion.
His judgment and interest may , however , have kept equal pace in that conclusion.

- c) Before leaving and before saying goodbye , I looked around.
Before leaving and before the bye bye moment , I looked around.
- d) My sisters, lovely girls, live in Melbourne.
My sisters, nice girls, live in Melbourne.

A.3 Tagset Definition

In the following, the annotation specifics are presented. For each tag, we provide (1) the definition and (2) examples both for “positive sense preserving” and “negative sense preserving” instances, where applicable.

A.3.1 Morphology based changes

Morphology based changes stand for those paraphrases that take place at the morphology level of language. Some changes in this class arise at the morphology level, but entail significant structural implications in the sentence. Only the linguistic unit affected by the trigger morphology change is annotated.

A.3.1.1 Inflectional changes

Definition: Inflectional changes consist in changing inflectional affixes of words. In the case of verbs, this type includes all changes within the verbal paradigm.

Negative sense preserving inflectional changes lead to significant changes in the meaning of the whole text, thus giving raise to non-paraphrases.

- Positive sense preserving:
It was with difficulty that the course of **streets** could be followed.
You couldn’t even follow the path of the **street**.
- Negative sense preserving:
You can’t travel from Barcelona to Mallorca with the **boat**.
Boats can’t travel from Barcelona to Mallorca.

A.3.1.2 Modal verb changes

Definition: The MODAL VERB tag stands for changes of modality using modal verbs.

- Positive sense preserving:
I was still lost in conjectures who they **might** be.
I was pondering who they **could** be.

A.3.1.3 Derivational changes

Definition: The DERIVATIONAL tag stands for changes of category by adding derivational affixes to words. These changes comprise a syntactic reorganization in the sentence where they occur.

- Positive sense preserving:
I have heard many accounts of him all **differing** from each other.
I have heard many **different** things about him.

Although drivers and driving are linked by a derivational process, in the following example this type is classified as SAME-POLARITY, and not as a DERIVATIONAL, because there is not an actual change of category, both are acting as nouns.

- I dislike rash drivers.
I dislike rash driving.

A.3.2 Lexicon based changes

Lexicon based change tags stand for those paraphrases that arise at the lexical level.

Always the **smallest** possible lexical unit has to be annotated. In (a), we should not consider one single paraphrase phenomenon because it can be divided into two lexical units pairs: often-debated/much-disputed (1) and issue/question (2). These SAME-POLARITY substitutions are independent paraphrase phenomena, as we could substitute often-debated by much-disputed, leaving issue unchanged (much-disputed issue). Thus, two different SAME-POLARITY tags should be used. In contrast, in (b), lies and is revealed should not be tagged on their own as SAME-POLARITY substitutions, as they are semantically embedded in the wider lexical units lies its appeal and its appeal is revealed, respectively. The tag used in this case is, again, SAME-POLARITY.

- a) often-debated₁ issue₂
much-disputed₁ question₂
- b) Here by virtue of humanity's vestures, lies its appeal .
Here by virtue of humanity's vestures, its appeal is revealed .

Auxiliaries and infinitive marks are not tagged within the lexical unit in question. Only the verb to be, when it is part of a passive voice, should be included in the scope (c).

- c) The viewpoint of these lands had been altered .
The whole aspect of the land had changed.

A.3.2.1 Spelling changes

Definition: This type comprises spelling changes and changes in the lexical form in general. Spelling is always sense preserving. Some examples:

1. Spelling

- a) color / colour

2. Acronyms

- b) North Atlantic Treaty Organization / NATO

3. Abbreviations

- c) Mister / Mr.

4. Contractions

- d) you have / you've

5. Hyphenation

- e) flow-accretive / flow accretive

A.3.2.2 Same Polarity Substitution

Definition: The SAME-POLARITY tag is used when a lexical unit is changed for another one with approximately the same meaning. Both lexical (a) and functional (b) units are considered within this type. Sameness of category is not a requisite to belong to this type (c).

- a) The pilot took off despite the stormy weather .
The plane took off despite the stormy weather .
- b) Despite the stormy weather
In spite of the stormy weather
- c) He rarely makes us smile .
He has little to do with making us smile.

When prepositions are part of a larger lexical unit, changes or deletions of these prepositions are tagged as SAME-POLARITY and annotated together with the lexical unit where they are embedded (d).

- d) do away / do away with

SAME-POLARITY may be used to tag several linguistic mechanisms, the following among them:

1. Synonymy

- e) I like your house .
I like your place .

2. General/specific

- f) I dislike rash motorists .
I dislike rash drivers .

3. Exact/approximate

- g) They were 9 .
They were around 10 .

4. Metaphor

- h) I was staring at her shinning teeth .
I was staring at her shinning pearls .

5. Metonymy

- i) I read a book written by Shakespeare .
I read a Shakespeare

6. Expansion/compression: expressing the same content with multiple pieces and/or in a more detailed way.

- j) Ended up causing a calm aura
Caused a rather sober and subdued air

7. Word/definition

- k) Heart attacks have experienced an increase in the last decades.
Sudden coronary thromboses have experiences an increase in the last decades.

8. Translation

- l) Jean-Francois Revel, in History of the Western Philosophy
Jean-Francois Revel, in Histoire de la philosophie occidentale

9. Idiomatic expressions

- m) It is raining cats and dogs .
It is raining a lot .

10. Part/whole

- n) Yesterday I cut my finger .
Yesterday I cut my hand

In the EPT, we distinguish between **three different kinds same-polarity substitution**: habitual, contextual, and named entity. The kind of same-polarity substitution depends on the nature of the relation between the substituted text.

Same Polarity Substitution (habitual)

The SAME-POLARITY (HABITUAL) tag is used when a lexical unit is changed for another one with approximately the same **dictionary** meaning. The substituted units have a similar meaning outside of the particular context as well as within the context. Same-polarity (habitual) is always **sense preserving**:

- Positive sense preserving:
A federal magistrate in Fort Lauderdale ordered him held without bail.
Zuccarini was ordered held without bail Wednesday by a federal judge in Fort Lauderdale, Fla.

Same Polarity Substitution (contextual)

The SAME-POLARITY (CONTEXT) tag is used when a lexical unit is changed for another one with approximately the same meaning **within the given context**. The substituted units have different out-of-context meaning. The negative sense preserving SAME-POLARITY is always contextual (unless it requires named entity reasoning). In the case of **negative sense preserving** same polarity substitution, the meaning of the units is similar, but not the same - it includes key differences and/or incompatibilities that give rise to non-paraphrasing at the level of the two texts.

- Positive sense preserving:
Meanwhile, the global death toll approached 770 with more than 8,300 people sickened since the severe acute respiratory syndrome virus first appeared in southern China in November.
The global death toll from SARS was at least 767, with more than 8,300 people sickened since the virus first appeared in southern China in November.

- Negative sense preserving:
The loonie, meanwhile, continued to slip in early trading Friday.
The loonie, meanwhile, was on the rise again early Thursday.

Same Polarity Substitution (Named Entity)

The SAME-POLARITY (NE) tag is used when a lexical unit is changed for another one with approximately the same meaning. Both replaced units are **named entities or properties of named entities**. Some degree of world knowledge and named entity reasoning is required to correctly determine whether the substitution is sense preserving or not. In the case of **negative sense preserving** same polarity substitution, the meaning of the units is similar, but not the same - it includes key differenced and/or incompatibilities that give raise to non-paraphrasing at the level of the two texts.

- Positive sense preserving:
He told The Sun newspaper that Mr. Hussein's daughters had British schools and hospitals in mind when they decided to ask for asylum.
Saddam 's daughters had British schools and hospitals in mind when they decided to ask for asylum – especially the schools, he told The Sun.
- Negative sense preserving:
Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for \$2.5 billion .
Yucaipa bought Dominick's in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998.

A.3.2.3 Change of Format

Definition: This tag stands for changes in the format. Format is always sense preserving. Some examples:

1. Digits/in letters

- a) 12 / twelve

2. Case changes

- b) Chapter 3 / CHAPTER 3

3. Format changes

- c) 03/08/1984 / Aug 3 1984

A.3.3 Lexico-syntactic based changes

Lexico-syntactic based change tags stand for those paraphrases that arise at the lexical level but are also entailing significant structural implications in the sentence. Similar to lexicon changes always the **smallest** possible lexical unit has to be annotated.

A.3.3.1 Opposite polarity substitution

Definition: OPPOSITE-POLARITY stands for changes of one lexical unit for another one with opposite polarity. In order to maintain the same meaning, other changes have to occur. Two phenomena are considered within this type:

1. Double change of polarity A lexical unit is changed for its antonym or complementary. In order to maintain the same meaning, a double change of polarity has to occur within the same sentence: another antonym (a) or complementary substitution (b), or a negation (c). In the case of double change of polarity, the two changes of polarity have to be tagged as a single (and possibly discontinuous, like in b) phenomenon and using a single tag.

a) John lost interest in the endeavor .
John developed disinterest in the endeavor .

b) Only 20% of the students were late .
Most of the students were on time .

c) He did not succeed in either case .
He failed in both enterprises .

2. Change of polarity and argument inversion An adjective is changed for its antonym in comparative structures. In order to maintain the same meaning, an argument inversion has to occur (d). In the case of change of polarity and argument inversion, only the antonym adjectives are tagged.

d) The neighboring town is poorer in forest resources than our town.
Our town is richer in forest resources than the neighboring town.

In the EPT, we distinguish between **two different kinds opposite-polarity substitution**: habitual and contextual. The kind of opposite-polarity substitution depends on the nature of the relation between the substituted text.

Opposite polarity substitution (habitual)

The OPP-POLARITY (HABITUAL) tag is used when a lexical unit is changed for another one with approximately the opposite **dictionary** meaning. The substituted units have an opposite meaning outside of the particular context as well as within the context. The **negative sense preserving** Opposite Polarity Substitution appears in two different situations. First, the case where the meaning of the two units is not completely opposite - it includes key differences and/or incompatibilities that give raise to non-paraphrasing at the level of the two texts. Second, the case where the meaning of the two units are the same, but the other changes (double change of polarity or argument inversion) are not found.

- Positive sense preserving:
Leicester failed in both enterprises.
He did not succeed in either case.
- Negative sense preserving:
John loved his new car.
He hated that car.

Opposite polarity substitution (contextual)

The OPP-POLARITY (CONTEXT) tag is used when a lexical unit is changed for another one with approximately the opposite meaning **within the given context**. The substituted units have different out-of-context meaning. The **negative sense preserving** Opposite Polarity Substitution appears in two different situations. First, the case where the meaning of the two units is not completely opposite - it includes key differences and/or incompatibilities that give raise to non-paraphrasing at the level of the two texts. Second, the case where the meaning of the two units are the same, but the other changes (double change of polarity or argument inversion) are not found.

- Positive sense preserving:
A big surge in consumer confidence has provided the only positive economic news in recent weeks.
Only a big surge in consumer confidence has interrupted the bleak economic news.
- Negative sense preserving:
Johnson welcomed the new proposal.
Johnson did not approve of the new proposal.

A.3.3.2 Synthetic/Analytic substitution

Definition:SYNTHETIC/ANALYTIC stands for those changes of synthetic structures to analytic structures and vice versa. It should be noted, however, that sometimes “syntheticity” or “analyticity” is a matter of degree. Consider examples (a) and (b). In (a), we would probably consider as analytic the genitive structure. In (b), in contrast, the genitive structure would probably be the synthetic one. Genitive structures are not synthetic or analytic by definition, but more or less synthetic/analytic compared to other structures. Thus, we could redefine this group as a change in the degree of syntheticity/analyticity.

a) the Met show / the Met’s show

b) Tina’s birthday / The birthday of Tina

SYNTHETIC/ANALYTIC is always **sense preserving** and comprises phenomena such as:

1. Compounding/decomposition A compound is decomposed through the use of a prepositional phrase (a). The alternation adjectival/prepositional phrase (b) and single word/adjective+noun alternations (c) are also considered here.

a) The gamekeeper preferred to make wildlife television documentaries .
The gamekeeper preferred to make television documentaries about wildlife .

b) Chemical life-cycles of the sexes
Life-cycles for chemistry for genders

c) One of his works holding the title "Liber Cosmographicus De Natura Locorum" belongs to a category of physiography .
One of his works bearing the title of "Liber Cosmographicus De Natura Locorum" is a species of physical geography .

2. Alternations affecting genitives and possessives Alternations between genitive/prepositional phrases (d), possessive/prepositional phrases (e), genitive/nominal phrases (f), genitive/adjectival phrases (g), etc.

d) Tina’s birthday / The birthday of Tina

e) His reflection / The reflection of his own features

f) the Met show / the Met’s show

g) Russia’s Foreign Ministry / the Russian Foreign Ministry

N.B.: A distinction has to be established between this type and DERIVATIONAL. Some DERIVATIONAL cases also contain genitive alternations (h), but these alternations are part of a wider derivational change. In the cases of genitive alternations classified as SYNTHETIC/ANALYTIC, the alternation is an isolated and independent phenomenon.

- h) Mary teaches John .
Mary is John's teacher .

N.B.: Cases of 1 (compounding/decomposition) and 2 (alternation involving genitives and possessives) in which the alternation takes place with a clause (with a verb) are not considered here but in SUBORDINATION & NESTING (i)

- i) Volcanoes **which** are now extinct / extinct volcanoes

3. Synthetic/analytic superlative

- j) He's smarter than everybody else .
He's the smartest .

4. Light/generic element addition: Changing a synthetic form A for an analytic form BA by adding a more generic element (B is more generic than A). A has to have the same lemma/stem in both member of the pair as in (k). Moreover, although the category of the phrase A and the phrase BA may differ, the change does not have structural consequences outside A or BA. In (l), although the adverbial phrase *cheerfully* is changed to the prepositional phrase *in a cheerful way*, the rest of the sentence remains unchanged. Finally, the order of the A and B units can be BA (k) or AB (l).

- k) John boasted about his work.
John spoke boastfully about his work.
- l) Marilyn carried on with her life cheerfully .
Marilyn carried on with her life in a cheerful way .

N.B.: When B is the verb to be and there is a change of category of A through a derivational process, the phenomenon is tagged as DERIVATIONAL (m)

- m) Sister Mary was helpful to Darrell .
Sister Mary helped Darrell .

5. Specifier addition: This type is parallel to the previous one, but the added element B is not more generic, but focuses on one of the components or characteristics of A (n), emphasises A (o) or determines A (p).

- n) I had to drive through fog to get there .
I had to drive through a wall of fog to get there .
- o) We are meeting at 5 . We are meeting at 5 o'clock .
- p) Translation is what they need .
The translation is what they need .

N.B.: Contrary to SAME-POLARITY or SEMANTICS BASED CHANGES, where words vary from one member of the paraphrase pair to the other, in synthetic/analytic substitutions

- although a change of category may take place, lexical word stems are the same (1 and 2) or
- a support element is added, but other lexical word stems are the same(4 and 5).

A.3.3.3 Converse substitution

Definition: A lexical unit is changed for its converse. In order to maintain the same meaning, an argument inversion has to occur. The **negative sense preserving** converse substitution occurs when the arguments are not inverted.

- Positive sense preserving:
The Geological society of London in 1855 awarded to him the Wollaston medal.
Resulted in him receiving the Wollaston medal from the Geological society in London in 1855.
- Negative sense preserving:
Last Monday, John bought the new black car from his friend Sam.
Last week, John sold his black car to Sam, his friend from high school.

A.3.4 Syntax based changes

Syntax based change tags stand for those changes that involve a syntactic reorganization in the sentence. This type basically comprises changes within a single sentence; and changes in the way sentences, clauses or phrases are connected. The phrase/clause/sentence(s) suffering the modification is(are) tagged. All syntax tags but DIATHESIS have key elements that should be annotated as well.

A.3.4.1 Diathesis alternation

Definition: DIATHESIS gathers the diathesis alternations in which verbs can participate. The whole linguistic unit suffering the syntactic reorganization is tagged. The **negative** sense preserving diathesis alternation occurs when the arguments are not properly changed or inverted.

- Positive sense preserving:
The guide drew our attention to a gloomy little dungeon.
Our attention was drawn by our guide to a little dungeon.
- Negative sense preserving:
The president gave a speech about his plan to change the Constitution.
The president was given a speech about his plan to change the Constitution.

A.3.4.2 Negation switching

Definition: Changing the position of the negation within a sentence. The whole linguistic unit suffering the modification is tagged (not only the negation scope). Negation marks are tagged as key elements. The **negative sense preserving** negation switching occurs when the scope of negation in the two texts is significantly different and that changes the overall meaning. A special case of negative sense preserving negation switching is when one of the texts (sentences) is affirmative, and the other is negative.

- Positive sense preserving:
In order to move us, it needs **no** reference to any recognized original.
One does **not** need to recognize a tangible object to be moved by its artistic representation.
- Negative sense preserving:
Frege did **not** say that Hesperus is Phosphorus.
Frege said that Hesperus is **not** Phosphorus.

A.3.4.3 Ellipsis

Definition: This tag includes linguistic ellipsis, i.e., those cases in which the elided snippets can be recovered through linguistic mechanisms. In (a), in the first member of the pair the idea of “being able to change to” is expressed twice; in the second member of the pair it is only expressed once due to elision. The whole linguistic unit suffering the modification is tagged (not only the elided snippets). All appearances of the elided snippet in both sentences are tagged as key elements: the idea of “being able to change to” in (a). Ellipsis is always **sense preserving**.

- a) - Thus, chemical force **can become** electrical current and that current **can change back** into chemical being.
- So we **can change** chemical force into the electric current, or the current into chemical force.

N.B.: When the elided snippets cannot be recovered solely through linguistic mechanisms, they must be considered DELETIONS.

A.3.4.4 Coordination changes

Definition: Changes in which one of the members of the pair contains coordinated snippets. This coordination is not present (in (a) it changes to a juxtaposition) or changes its position and/or form (b) in the other member of the pair. Only the coordinated or juxtaposed linguistic units are tagged. Only the coordination (not juxtaposition) marks are tagged as key elements. Coordination changes are always **sense preserving**.

- a) I like pears **and** apples.
I like pears. I like apples
- b) Older plans **and** contemporary ones
Old **and** contemporary plans

N.B.: When the alternation takes place between, on the one hand, coordinated or juxtaposed units and, on the other hand, subordinated or nested units, the phenomenon is tagged as SUBORDINATION & NESTING.

A.3.4.5 Subordination and Nesting changes

Definition: Changes in which one of the members of the pair contains a subordination (a) or a nesting (b). This subordination or nesting is not present (in (a) and (b) it changes to a juxtaposition) or changes the position and/or form (c) in the other member of the pair. Nesting is understood as a general term meaning that something is embedded in a bigger unit. Only the linguistic units involved in the subordination or nesting, as well as the coordinated and juxtaposed units, are tagged. In case a conjunction, a relative pronoun or a preposition are present, they are tagged as the key elements (a and c). In case they are not present, the whole subordinated or nested snippet is tagged (b). Juxtaposition or coordination elements are not tagged as key elements. Subordination and Nesting changes are always **sense preserving**.

- a) A building, **which** was devastated by the bomb, was completely destroyed.
A building was devastated by the bomb. It was completely destroyed.

- b) Patrick Ewing scored **a personal season high** of 41 points.
Patrick Ewing scored 41 points. It was a personal season high
- c) The conference venue is in the building **whose** roof is red .
The conference venue is in the building with red roof .

A.3.5 Discourse based changes

These tags stand for those changes that take place at the discourse level of language. This type gathers phenomena that are very different in nature, though all having in common that consist in structural changes not affecting the argumental elements in the sentence. The phrase/clause/sentence(s) suffering the modification is(are) tagged. Moreover, a key element should be tagged in all discourse based tags.

A.3.5.1 Punctuation changes

Definition: Changes in the punctuation (a). Cases consisting of linguistic mechanisms parallel to punctuation like (b) are also considered here. The changing punctuation signs are tagged as key elements. The whole linguistic unit(s) suffering the modification is(are) tagged. Punctuation is always **sense preserving**.

- a) This, as I see it, is wrong .
This – as I see it – is wrong.
- b) - You will purchase a return ticket to Streatham Common and a platform ticket at Victoria station .
- At Victoria Station you will purchase **(1)** a return ticket to Streatham Common and **(2)** a platform ticket

Sometimes occurs that several changes in the punctuation take place at the same time. These multiple changes are considered as a single phenomenon if they take place at the same level (between phrase, between clause or between sentence), like in (c). If they belong to different levels, they are tagged as separate phenomena: two changes in the punctuation take place in (d), repeated in (e), but they are annotated as independent paraphrase phenomena: one of them is tagged in (d) and the other in (e).

- c) I know she is coming. She will be fine; I know it .
I know she is coming; she will be fine. I know it .
- d) I need to buy a couple of things. Then, I will come .
I need to buy a couple of things; then I will come .

- e) I need to buy a couple of things. Then , I will come .
 I need to buy a couple of things. then I will come .

A.3.5.2 Direct/Indirect style alternations

Definition: Changing direct style for indirect style, and vice versa. The whole linguistic unit suffering the modification is tagged. The conjunction in the indirect style is tagged as key element. If no conjunction is present, the whole subordinate clause is tagged. The **negative sense preserving** Direct/Indirect Style alternations do not trigger the appropriate changes for pronoun resolution.

- Positive sense preserving:
 She is mine, said the Great Spirit.
 The Great Spirit said **that** she is hers.
- Negative sense preserving:
I'm on my way!, said Peter and hung up his phone .
 Peter called Ana to tell her **that** she is on her way .

A.3.5.3 Sentence modality changes

Definition: Cases in which there is a change of modality (a). We are referring strictly to changes between affirmative, interrogative, exclamatory and imperative sentences. The whole unit suffering the modification is tagged. The elements that change are tagged as key elements. Modality change is always **sense preserving**.

- a) **Can** I make a reservation?
 I'd **like to** make a reservation.

N.B.: In MODAL VERB tags, in contrast, only modal verb alternations are involved.

A.3.5.4 Syntax/Discourse Structure

Definition: This tag is used to annotate other changes in the structure of the sentences not considered in the syntax and discourse based tags above: (a), (b) and (c). The linguistic unit(s) suffering the modification is(are) tagged. The elements that change are tagged as key elements.

- a) John wore his best suit to the dance last night .
It was John **who** wore his best suit to the dance last night .

- b) He wanted to eat **nothing but** apples .
All he wanted to eat **were** apples.
- c) **You are very** courageous .
You have shown how courageous **you are** .

A.3.6 Other changes

This class gathers those changes that are related to more than one of the classes and subclasses in our typology, as they can take place in any of them.

A.3.6.1 Addition/Deletion

Definition: Deletion of lexical and functional units. In the **negative sense preserving** case, the deletion leads to a significant modification of the meaning. Only the linguistic unit deleted is tagged. When a functional unit is deleted together with a lexical unit, this functional unit is included in the scope. Normally, the scope of Addition/Deletion is only in one of the two texts, as opposed to the other types, which are pairwise.

- Positive sense preserving:
One day, she took a hot flat-iron, removed my clothes, and held it on my naked back until I howled with pain.
 As a proof of bad treatment, she took a hot flat-iron and put it on my back after removing my clothes.
- Negative sense preserving:
 Legislation making it harder for consumers to erase their debts in bankruptcy court won overwhelming House approval in March.
 Legislation making it harder for consumers to erase their debts in bankruptcy court won speedy, House approval in March and was endorsed by the White House.

A.3.6.2 Change of order

Definition: This tag includes any type of change of order from the word level to the sentence level: (a), (b) and (c). Change of order is always **sense preserving**.

- a) She used to only eat hot dishes.
 She used to eat only hot dishes.
- b) “I want a beer”, he said.
 “I want a beer”, said he.

- c) They said : “We believe that the time has come for legislation to make public places smoke-free” .
“The time has come to make public places smoke-free,” they wrote in a letter to the Times newspaper.

A.3.6.3 Semantic (General Inferences)

Definition: SEMANTICS BASED CHANGES tag stands for changes that imply a different lexicalisation pattern of the same content units. Typically the semantic relation between the two can only be determined through (common sense) reasoning. In the **negative sense preserving** case, the reasoning identifies contradiction and/or incompatibility.

- Positive sense preserving:
 Uncle Tarek was born Aribert Ferdinand Heim.
 The real name of Tarek Hussein Farid was Aribert Ferdinand Heim.
- Negative sense preserving:
 Children were among the victims of a plane crash that killed as many as 17 people Sunday in Butte, Montana.
 17 adults died in a plane crash in Montana.

A.3.7 Extremes

The following types stand for the extremes of the paraphrase continuum: identity on the one hand, and entailment and non-paraphrase on the other.

A.3.7.1 Identity

Definition: We annotate as IDENTICAL those linguistic units that are exactly the same in wording. Identical is always **sense preserving**.

- The two argued that only a new board would have had the credibility to restore El Paso to health.
The two believed that only a new board would have had the credibility to restore El Paso to health.

A.3.7.2 Non-paraphrase

Definition: Non-paraphrase includes fragments which do not have the same meaning (a), as well as cases in which we need extralinguistic information in order to establish a link between the members of the paraphrase pair: cases of

same illocutive value but different meaning (b), cases of subjectivity (c), cases of potential coreference (d), (e) and (f), etc.

- a) The two had argued that you shouldn't go there .
He and Zilkha believed that this is unfair .
- b) I want some fresh air.
Could you open the window?
- c) The U.S.-led invasion of Iraq .
The U.S.-led liberation of Iraq.
- d) They got married last year .
They got married in 2004 .
- e) I live here . I live in Barcelona .
- f) They will come later .
They will come this afternoon

N.B.: Paraphrase and coreference overlap considerably. Those cases that may corefer, but at the same time are paraphrases, should be annotated as paraphrases.

In cases (d), (e) and (f), the linguistic information is not enough to link the two members of the pair, we need to know which point in the time or in the space are we taking as reference. Thus, they are annotated as nonparaphrases. Cases in (g), (h) and (i) can be linked only through linguistic information (a year in the past, a 'city' type of entity, a masculine singular entity, respectively). Thus, they are annotated as paraphrases.

- g) They got married last year .
They got married a year ago .
- h) I live in Barcelona .
I live in a city .
- i) I love John .
I love him .

N.B.: Although sometimes a non-paraphrase fragment may actually affect the meaning of the full sentence, only the fragment in question will be tagged as NON-PARAPHRASE (j) and the rest of the sentence will be annotated independently of this fact.

- j) Mike and Lucy decided to leave .
Mark decided to leave .

N.B.: When two linguistic units having a different meaning are not aligned formally nor informatively, they should be tagged as two different ADDITION/DELETION cases (1 and 2 in k), not as NON-PARAPHRASES.

- k) Yesterday,₁ Google failed .
 Google failed because of the crisis₂.

A.3.7.3 Entailment

Definition: Fragments having an entailment relation. **N.B.:** It should be noted that entailment relations are present in many paraphrase types (e.g. general/specific in SAME-POLARITY or ADDITION/DELETION). We will only use the ENTAILMENT tag when there is a substantial difference in the information content. Entailment is always **negative sense preserving**.

- Google was in talks to buy Youtube .
 Google bought Youtube

A.4 Annotating non-paraphrases

Annotating non-paraphrases (negative examples of paraphrasing in the MRPC corpus) is a non-trivial task that has not been carried out for other paraphrase typology corpora. The non-paraphrases in the MRPC corpus have many of the properties of paraphrases, they have a very high degree of lexical and syntactic similarity. In a) we can see an example of a non-paraphrase pair. The two sentences talk about the same NEs (Yucaipa and Dominick) in the same syntactic-semantic roles of the same actions (buying, selling, owning). At the same time, there are key differences between the two sentences – the price of the sale in the first sentence is \$2.5 billion, while in the second it is \$1.8 billion.

- a) Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for \$2.5 billion.
 Yucaipa bought Dominick's in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998.

Due to the complex nature of the non-paraphrasing, the annotation of these pairs goes in three steps

- 1) (Re)evaluation of the paraphrasing or non-paraphrasing relation between the two sentences as a whole (this is the first step for both paraphrases and non-paraphrases).

- 2) (After the pair has been annotated as non-paraphrases) Annotation of the non-sense-preserving phenomena, responsible for the non-paraphrasing label of the pair.
- 3) Annotation of the sense-preserving phenomena, responsible for the high degree of similarity between the two sentences.

An example annotation of the pair in a) follows:

- 1) The relation between the two sentences is non-paraphrases
- 2) The non-sense-preserving phenomena responsible for the “non-paraphrase” label of the pair is “Lexical Substitution (Named Entities)”:
Yucaipa owned Dominick’s before selling the chain to Safeway in 1998 for \$2.5 billion .
Yucaipa bought Dominick’s in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998.
- 3) The sense-preserving phenomena, responsible for the high degree of similarity are:
 - a. Same polarity substitution (contextual)
Yucaipa owned Dominick’s before selling the chain to Safeway in 1998 for \$2.5 billion.
Yucaipa bought Dominick’s in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998.
 - b. Entailment
Yucaipa owned Dominick’s before selling the chain to Safeway in 1998 for \$2.5 billion.
Yucaipa bought Dominick’s in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998.
 - c. Inflectional changes
Yucaipa owned Dominick’s before selling the chain to Safeway in 1998 for \$2.5 billion.
Yucaipa bought Dominick’s in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998.
 - d. Order
Yucaipa owned Dominick’s before selling the chain to Safeway in 1998 for \$2.5 billion.
Yucaipa bought Dominick’s in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998.

e. Addition/Deletion

Yucaipa owned Dominick's before₁ selling the chain to Safeway in 1998 for \$2.5 billion.

Yucaipa bought Dominick's in 1995 for \$693 million and₂ sold it to Safeway for \$1.8 billion in 1998.

f. Identity

Yucaipa owned Dominick's before selling the chain to Safeway in 1998 for \$2.5 billion . Yucaipa bought Dominick's in 1995 for \$693 million and sold it to Safeway for \$1.8 billion in 1998 .

A.5 Annotating negation

Annotating negation within paraphrases is a novel approach. For the pilot annotation we will mark the scope as negation and the negation cue as a “key”.

- We did **not** drive up to the door but got down near the gate of the avenue .

Appendix B

Annotation Guidelines for Gold et al. [2019]

In this task each text pair is annotated independently for Paraphrasing, Textual Entailment, Contradiction, Textual Specificity, and Textual Similarity. At each annotation step, annotators are asked to determine the presense or absense of a single textual meaning relation. For Textual Entailment and Textual Specificity, each pair is shown twice, with the order of the texts changed to address the directionality of the relations. The instructions provided to the annotators are the following.

B.1 Paraphrasing

Background: We want to study the meaning relation between two texts. Thus you are asked to determine whether the two sentences mean (approximately) the same or not.

Task: In this task you are presented with **two sentences**. You are required to decide whether the two sentences **have approximately the same meaning or not**.

In the case of pronouns (he, she, it, mine, his, our, ...) being used, you can assume they reference proper names, if your common sense does not suggest otherwise (e.g. “Linda” is a female name and can be referenced by “she, her, ...”, but not “he, his, ...”).

Examples of the choce “approximately the same meaning”:

- John goes to work every day with the metro.
- He takes the metro to work every day.

In the content of the task, we assume that “He” and “John” are the same person.

- Mary sold her Toyota to Jeanne.
- Jeanne bought her Toyota from Mary Smith.

In the content of the task, we assume that “Mary Smith” and “Mary” are the same person.

Examples of the choice of “not the same meaning”:

- Mary sold her Toyota to Jeanne.
- Mary had a blue Toyota.

The two texts are related, but are not the same.

- John Smith takes the metro to work every day.
- John works from home every Tuesday.

The two texts are not closely related except for the person (John).

B.2 Textual Entailment

Background: We want to research causal relationships between sentences, which will help in information retrieval or summarization tasks. Thus, you are asked to determine whether given that the first sentence is true, the second sentence is also true.

Task: In this task, you are presented with **two sentences**. You are required to decide whether **if Sentence 1 is true, this also makes Sentence 2 true**.

In the case of pronouns (he, she, it, mine, his, our, ...) being used, you can assume they reference proper names, if your common sense does not suggest otherwise (e.g. “Linda” is a female name and can be referenced by “she, her, ...”, but not “he, his, ...”).

Examples for the option “Sentence 1 causes Sentence 2 to be true”:

In that case, the first sentence causes the second sentence to be true, as assuming that John bought a car, it means that he has a car now.

- John bought a car from Mike.
- John has a car.

In that case, the first sentence causes the second sentence to be true, as the first sentence says that both boys and girls play games, it also contains the information that boys play games.

- Boys and girls play games.
- Boys play games.

Examples for the option “Sentence 1 does not cause Sentence 2 to be true”:

*If the second sentence makes the first sentence true (but the first doesn't make the second true), choose the option “Sentence 1 **does not** cause Sentence 2 to be true”:*

- John has a car.
- John bought a car from Mike.

*If you cannot tell if the first sentence causes the second sentence to be true, choose the option “Sentence 1 **does not** cause Sentence 2 to be true”:*

- He works as a teacher in Peru.
- He is an English teacher.

B.3 Contradiction

Background: We want to study the meaning relation between two texts. Thus you are asked to determine whether the two sentences contradict each other.

Task: In this task you are presented with **two sentences**. You are required to decide whether the **two sentences contradict each other**. Two contradicting sentences can't be true at the same time.

In the case of pronouns (he, she, it, mine, his, our, ...) being used, you can assume they reference proper names, if your common sense does not suggest otherwise (e.g. “Linda” is a female name and can be referenced by “she, her, ...”, but not “he, his, ...”).

Examples for the option “the sentences contradict each other”:

- John bought a new house near the beach.
- John didn't buy the house near the beach.

The second sentence directly contradicts the first one they can't both be true.

- Mary is on a vacation in Florida.
- Mary is at the office, working.

The two sentences can't be true at the same time Mary is either on vacation in Florida, or at the office. She can't be in two places.

Examples for the option “the sentences do not contradict each other”:

- Mary is on vacation in Florida.
- John is at the office.

John and Mary are two different persons. There is no contradiction. Both statements can be true.

B.4 Similarity

Background: We want to study the meaning relation between two texts. Thus you are asked to determine how similar two texts are.

Task: In this task you are presented with **two sentences**. You are required to decide **how similar the two sentences are on a scale from 0 (completely dissimilar) to 5 (identical)**.

In the case of pronouns (he, she, it, mine, his, our, ...) being used, you can assume they reference proper names, if your common sense does not suggest otherwise (e.g. “Linda” is a female name and can be referenced by “she, her, ...”, but not “he, his, ...”).

Example for Similarity 0:

- John goes to work every day with the metro.
- The kids are playing baseball on the field.

The two texts are completely dissimilar.

Example for Similarity 1-2:

- John goes to work every day with the metro.
- John sold his Toyota to Sam.

The two texts have some common elements, but are overall not very similar.

Example for Similarity 3-4:

- Mary is writing the report on her Lenovo laptop.
- Mary has a Lenovo laptop.

The two texts have a lot in common, but also have differences.

Example for Similarity 5:

- Mary was feeling blue.
- Mary was sad.

The two texts are (almost) identical.

B.5 Specificity

Background: We want to research whether displaying more specific sentences is helpful in information retrieval or summarization tasks. Thus, you are asked to determine whether the 1st sentence is more specific than the 2nd. The specificity of sentence is defined as a measure of how broad or specific its information level is.

Task: In this task, you are presented with **two sentences**. You are required to decide whether **the 1st sentence IS more specific than the 2nd**. If this is not the case, choose the option **the 1st sentence IS NOT more specific than the 2nd**.

Examples for the option “Sentence 1 IS more specific”

- I like cats.
- I like animals.

As the 1st sentence gives the more specific information on which animal is liked, it is more specific. Hence, you have to choose the option that the 1st sentence is more specific.

- The cute cafe has great coffee.
- The cafe sells coffee.

As the 1st sentence gives the more specific information on both the cafe and the coffee, it is more specific. Hence, you have to choose the option that the 1st sentence is more specific.

Examples for the option “Sentence 1 IS NOT more specific”

- I like animals.
- I like cats.

As the 2nd sentence gives the more specific information on which animal is liked, it is more specific. Hence, you have to choose the option that the 1st sentence is not more specific.

- I like dogs.
- I like cats.

Now, as in both cases the liked animal is mentioned, they have the same level of specificity. Hence, you have to choose the option that the 1st sentence is not more specific.

- I like black dogs.
- He saw a blind cat.

Now, as the information is very diverse, it is impossible to say which sentence is more specific. Hence, you have to choose the option that the 1st sentence is not more specific.

Appendix C

Annotation Guidelines for Kovatchev et al. [2020]

C.1 Presentation

This document sets out the guidelines for the annotation of atomic types using the Extended Typology for Relations. The task consists of annotating pairs of text that hold a textual semantic relation (paraphrasing, entailment, contradiction, similarity) with a textual label, and the atomic phenomena they contain. These guidelines have been used to annotate the ETRC corpus. For the purpose of the annotation, the WARP-Text annotation tool has been used.

N.B.: The task definition, tagset definition and annotation of linguistic phenomena in these Guidelines overlap with those for the ETPC corpus. The reader is encouraged to consult the ETPC guidelines presented in Appendix A or the full SHARel guidelines available online. Here I only provide the guidelines for the reason-based types.

C.2 Annotating reason-based Phenomena

Reason-based phenomena account for relations that cannot be expressed and processed using only linguistic knowledge. Like the linguistic phenomena, the reason-based phenomena can be sense-preserving or non-sense preserving. Our goals with the annotation of reason-based phenomena are twofold:

- 1) we want to make a precise and explicit annotation of the units involved in the inference
- 2) we want to determine the kind of reason-based and background knowledge required.

1a and 1b show an example of an “existential” reason-based – “speaking X” entails “X exists”. 2a and 2b show an example of “causal” reason-based – “X is written in Y (language)” entails “reading X requires Y (language)” .

1a Speaking more than one language is imperative today.

1b There is more than one language.

2a Reading the Bible requires studying Latin.

2b The Bible is written in Latin.

When annotating reason-based phenomena, there are several important things:

- Annotate all possible phenomena separately. The aim is to annotate every token that is not already annotated as linguistic or addition-deletion.
- The scope of some phenomena can overlap. That means some tokens may be part of multiple scopes.
- When choosing the scope, we choose the smallest scope possible. Unlike the sense-preserving, in this part of the annotation, the goal is to choose the most specific scope possible. For example, in 1a and 1b we could annotate “Speaking more than one language” and “There is more than one language”, but in order to be as specific as possible, we choose to only annotate “Speaking” and “There is”.
- When choosing the scope, if possible, try to annotate whole linguistic units without breaking them. For example in 2a and 2b, we could only annotate “Reading requires” and “is written in”
- Like in the linguistic phenomena – the sense preserving reason-based phenomena need not relate units that have similar syntactic or semantic role; however, the non-sense preserving reason-based phenomena must relate units that have similar syntactic or semantic role.

C.3 List of reason-based phenomena

1. Cause and Effect: T causes H to be true [neg. sense preserving: T causes H to be FALSE]

- a. “Once a person is welcomed into an organization, they belong to that organization”

2. Conditions and Properties: A very general type where H containing facts (and properties) implied by T [neg. sense preserving: H contains facts and properties that contradict the implied from T (ex.: “There is only one language”)]

- a. Existential – T entails H exists (pre-requirement)
- b. “To become a naturalized citizen, one must not have been born there” (pre-requirement)
- c. “The type of thing that adopts children is person” (argument type)
- d. “When a person is an employee, that organization pays his salary” (simultaneous conditions)

3. Functionality: Relationships which are functional [neg. sense preserving: mutual exclusivity – types of things that do not participate in the same relationship]

- a. “A person can only have one father (or two arms)” (+)
- b. “Government and media sectors usually do not employ the same person” (-)

4. Transitivity: If R is transitive and R(a,b) and R(b,c) are true, then R(a,c)

- a. “The “support” is transitive. If Putin supports United Russia party, and United Russia party supports Medvedev, then Putin supports Medvedev”

5. Numerical Reasoning

6. Named Entity Reasoning: reasoning that goes beyond substitution; relations between multiple entities (i.e. not just Trump – president, but rather Trump – Clinton)

7. Temporal and Spatial Reasoning

8. Other (World Knowledge)