
SemRelData

Multilingual Contextual Annotation and Analysis of Semantic Relations between Nominals

Masterthesis von Darina Benikova

22/06/2015



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich Gesellschafts- und Geisteswissenschaften
Institut für Sprach- und Literaturwissenschaften

SemRelData:

Multilingual Contextual Annotation and Analysis of Semantic Relations between Nominals

Vorgelegte Masterthesis von Darina Benikova

1. Gutachten: Dr. Sabine Bartsch

2. Gutachten: Prof. Dr. Chris Biemann

Tag der Einreichung:

Erklärung zur Abschlussarbeit gemäß § 22 Abs. 7 APB der TU Darmstadt

Hiermit versichere ich, Darina Benikova, die vorliegende Masterthesis ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

In der abgegebenen Thesis stimmen die schriftliche und elektronische Fassung überein.

Datum:

Unterschrift:

Thesis Statement pursuant to § 22 paragraph 7 of APB TU Darmstadt

I herewith formally declare that I have written the submitted thesis independently. I did not use any outside support except for the quoted literature and other sources mentioned in the paper. I clearly marked and separately listed all of the literature and all of the other sources which I employed when producing this academic work, either literally or in content. This thesis has not been handed in or published before in the same or similar form.

In the submitted thesis the written copies and the electronic version are identical in content.

Date:

Signature:

Acknowledgement

I would hereby like to express my gratitude towards my supervisors, Dr. Sabine Bartsch, for maintaining my love of linguistics since my first semester, and Prof. Dr. Chris Biemann, for his continuous moral, scientific and financial support.

My thanks go also to the annotators of this project, Sarah Holschneider, Ilya Klyashtorny, Angelina Romashova and Xaver Ott, without whom the creation of the dataset and the subsequent analysis would not be possible.

I would also like to thank the whole LT group for their advice and technological support.

Furthermore, I would like to thank my parents, who have supported me in many ways throughout my whole life and my boyfriend, Christian Gold, who tolerated my behaviour in the stressful times of my studies and prepared my dinner every day during my thesis.

Zusammenfassung

Diese Thesis konzentriert sich auf kontextuelle Annotation und Analyse klassischer semantischer Relationen zwischen Nominalen in diversen Genres, wie enzyklopädischen, literarischen und nachrichtenbasierten Texten, und Sprachen, wie Englisch, Deutsch und Russisch.

Es wird angenommen, dass klassische semantische Relationen eine Rolle in linguistischer Wissensrepräsentation spielen. Der Hauptfokus dieser Thesis liegt auf der Analyse dieser Rolle im Kontext verschiedener Sprachen und Genres.

Im ersten Teil des Projektes wurden Synonyme, Hypernyme, Hyponyme, Co-hyponyme, Holonyme, und Meronyme in einer zweifachen Annotation nach Richtlinien, die in einem iterativen Prozess als Nebenprodukt erzeugt wurden, hinzugefügt. Die Annotation wurde mit Cohen's κ ausgewertet. Das κ wurde nicht durch Faktoren wie Sprache, Genre oder Textgröße beeinflusst. Allerdings konnte eine zeitabhängige Verbesserung der Inter-Annotator Übereinstimmung festgestellt werden, die auf die iterative Richtlinienverbesserung zurückzuführen ist. Es konnte auch gezeigt werden, dass Annotatoren zwar in der Klassifizierung einig sind, sich allerdings nicht auf die Erkennung von Relationen einigen können. Dies zeigt, dass die Konzepte klar sind, die Erkennung dieser sich aber schwierig gestaltet.

Im zweiten Teil wurde der annotierte Datensatz analysiert. Um Ähnlichkeiten und Unterschiede in der Verteilung von klassischen semantischen Relationen zwischen Nomen zu finden, wurden χ^2 -Test zwischen dem Sprachen-, Genre- und Enzyklopädie-subset durchgeführt. Es konnte gezeigt werden, dass es signifikante Unterschiede in der Verteilung semantischer Relationen und deren Typen in den meisten dieser Faktoren gibt.

Die Evaluation des Datensatzes wurde mithilfe von WordNet und seiner Pendanten, GermaNet und RuTes, vorgenommen. Etwa 50% der in dem Datensatz annotierten Relationen wurden in einer der anderen Datenbanken gefunden.

Die Ergebnisse dieser Thesis sind zweierlei: Annotation und Analyse. Einerseits wurde gezeigt, dass semantische Relationen annotierbare, sprachunabhängige Konzepte sind, auf deren Basis eine Einigung in einer zweifachen Annotation gefunden werden kann. Andererseits konnte gezeigt werden, dass semantische Relationen und die von ihnen verbundenen Entitäten eine wichtige semantische Rolle in dem zugehörigen Kontext von sowohl enzyklopädischen als auch literarischen Texten spielen.

Abstract

This thesis is focused on contextual annotation and analysis of classical semantic relations between nominals in various genres, such as encyclopaedic, literary and news texts, and languages, such as English, German and Russian.

It is assumed that classical semantic relations play a role in linguistic knowledge representation. The main purpose of this thesis is to analyse this role in the context of different languages and genres.

In the first part of this project, synonyms, hypernyms, hyponyms, co-hyponyms, holonyms, and meronyms were subject to double annotation according to guidelines, which were iteratively improved and are a side-product of this thesis. The annotation was evaluated using Cohen's κ . The κ did not vary according to factors such as language, genre, or text size. However, as a result of the iterative guideline improvement, a time-dependent inter-annotator agreement could be demonstrated. It was also shown that annotators mostly agree on the classification, but not on the detection of such relations, which shows that the concepts are clear, but the detection is difficult.

In the second part, the annotated dataset was analysed. In order to find similarities and differences in the distribution of classical semantic relations between nominals, χ^2 -tests were performed between the language, genre and categories in the encyclopaedic subset. It could be shown that there are significant differences in the distribution of semantic relations and their types between most of all these factors.

The evaluation of the dataset was performed using WordNet and its counterparts, GermaNet and RuTes. About 50% of the relations in the dataset created in this thesis were also present in one of the other databases.

The results of this thesis are twofold: annotation and analysis. On the one hand, it was shown that semantic relations are annotatable – language independent concepts that can be agreed upon in double annotation. On the other hand, it could be shown that semantic relations and the entities associated with them play an important semantic role in the corresponding contexts in both literary and encyclopaedic texts.

Table of Content

List of Tables	10
List of Figures	13
Glossary	14
1..... Introduction	17
1.1. Thesis Structure	18
2..... State of the Art	20
2.1. Semantic Relations	20
2.2. Classical Semantic Relations	22
2.2.1. Synonymy	24
2.2.2. Hyperonymy and Hyponymy	25
2.2.1. GermaNet	27
2.2.2. Holonymy and Meronymy	28
2.3. Implementations of Semantic Relation Classification	30
2.4. Hearst Patterns	30
2.5. Knowledge Bases containing Semantic Relations	31
2.5.1. WordNet	32
2.5.2. RuTes	33
2.5.3. BabelNet	34
2.5.4. NELL 35	
2.6. Semantic Web Ontologies	36
2.6.1. DBpedia	37
2.6.2. Freebase	37
2.6.3. YAGO38	
2.7. Concluding Remarks on Existing Resources	38
3..... Methods and Approaches	39
3.1. Cohen's κ	40
3.2. χ^2 -Test	41
4..... Collection of Dataset	43
4.1. Representativeness	43
4.2. Quality	45
4.3. Comparability	45
4.4. Copyright	46
5..... Preprocessing	48
6..... Annotation	49
6.1. Introduction	49
6.2. WebAnno	49
6.3. Annotation Process	50
6.3.1. Project Upset Description	50
6.3.2. Annotation of Documents	50
6.3.3. Curation and Export	51
6.4. Creation of Guidelines	52
6.4.1. Noun Compound Definition	52

6.4.2.	SemRelData's Iterative Relation Definition	54
6.5.	Inter-annotator Agreement	56
6.5.1.	Annotator as a Factor Influencing Annotator Agreement	57
6.5.2.	Paragraph Size as a Factor Influencing Annotator Agreement	58
6.5.3.	Time as a Factor Influencing Annotator Agreement	58
6.5.4.	Genre as a Factor Influencing Annotator Agreement	59
6.5.5.	Conclusions of Influential Factors on Annotator Agreement	59
6.5.6.	Conclusions on the Difference between Annotator and Curator Agreement	59
7.....	Postprocessing and Statistics	60
7.1.	Postprocessing of SemRelData	60
8.....	SemRelData Statistics and Characteristics	62
9.....	Results of the Analysis	63
9.1.	Comparison with Knowledge Bases	63
9.1.1.	Comparison with WordNet	65
9.1.2.	Comparison with GermaNet	65
9.1.3.	Comparison with RuTes	66
9.1.4.	Conclusion of the Comparison with other Knowledge Bases	66
9.2.	Comparison to Pattern-created Taxonomy	67
9.3.	Comparisons between Languages	68
9.3.1.	Comparison of Semantic Relation Density	68
9.3.2.	Comparison of Semantic Relation Types	69
9.3.3.	Conclusion of Comparison between Languages	70
9.4.	Comparisons between Genres	71
9.4.1.	Comparison of Semantic Relation Density	71
9.4.2.	Comparison of Semantic Relation Types	71
9.4.3.	Conclusion of Comparisons between Genres	72
9.5.	Comparison between Categories in Wikipedia-subset	72
9.5.1.	Comparison of Semantic Relation Density	72
9.5.2.	Comparison of Semantic Relation Types	72
9.5.3.	Conclusion of Comparison between Categories	73
9.6.	Comparison of Entities with the Highest Number of Relations	73
9.6.1.	Entities with the Highest Number of Relations within the Encyclopaedic Subset	74
9.6.2.	Entities with the Highest Number of Relations within the Literary Subset	74
9.6.3.	Conclusion of Entities with the Highest Number of Relations	75
10...	Conclusion	77
10.1.	Conclusion of the Annotation Task	77
10.2.	Conclusion of Relation Statistics	77
10.3.	Conclusion of Universality of Semantic Relations	78
10.4.	Conclusion of the Contextual Approach	79
10.5.	Conclusion of the Function of Semantic Relations	79
10.6.	Final Conclusion	80
11...	Further Work	81
	Reference List	84

A. ... Appendix	89
A.1. Annotator tests	115
A.1.1. English version	115
A.1.2. German version	116
A.1.3. Russian version	117
A.2. Guidelines	118
Introduction	119
Noun Compounds	119
1..... Bidirectional relations	120
1.1. Synonyms	120
1.2. Co-Hyponyms	122
2..... Uni-directional relations	124
2.1. Hypernyms	124
2.2. Holonyms	126
3..... General Rules	129
References	131

List of Tables

Table 2.1 Instantiations of relation by contrast (Murphy, 2003, p. 45).....	23
Table 2.2 Dimensions of synonymy (Murphy, 2003, p. 146).....	25
Table 2.3 Subclasses of holonymy expressed through part-of (Winston et al., 1987, p. 421)	29
Table 2.4 Size comparison between different databases	31
Table 2.5 Exemplary extract of the relations of <i>trousers</i> in WordNet, with hyperonymic relations of all terms.....	33
Table 2.6 Exemplary extract of the relations of <i>trousers</i> in BabelNet.....	34
Table 2.7 Relations of <i>trousers</i> in category <i>clothing</i> in NELL	36
Table 2.8 Exemplary semantic relations of <i>Paul McCartney</i> in DBpedia.....	37
Table 2.9 Exemplary semantic relations of <i>Paul McCartney</i> in Freebase	38
Table 2.10 Exemplary semantic relations of <i>Paul McCartney</i> in YAGO	38
Table 3.1 Exemplary contingency table	40
Table 3.2 Landis and Koch's scale of κ agreement	41
Table 3.3 Significance level and p-value correlation as presented by Bortz and Weber (2005)	42
Table 4.1 Number of tokens and noun compound in the individual genres	43
Table 4.2 Number of tokens and noun compound in the individual languages	43
Table 4.3 Table of all news article titles that were used for this dataset.....	44
Table 4.4 Table of all encyclopaedic articles that were used for this dataset.....	44
Table 4.5 Aggregated table of all literary works that were used for this dataset in the English translation	45
Table 6.1 Exemplary snippet of a curated .tsv file	51
Table 6.2 κ agreement of all annotators and the curator	57
Table 6.3 Annotator agreement sorted by time spans	58
Table 6.4 Annotator agreement sorted by genre	59
Table 8.1 Statistics of SemRelData. The 1 st column presents the number of noun compounds, the 2 nd column presents the number of tokens, the 3 rd column presents the number of annotated relations and the 4 th column presents the number of transitive relations.....	62
Table 9.1 Relations of <i>hose_en.tsv</i> in SemRelData.....	63
Table 9.2 Disagreement analysis of WordNet and SemRelData in 50 random relations	65
Table 9.3 Disagreement analysis of GermaNet and SemRelData in 50 random relations.....	66
Table 9.4 Disagreement analysis of RuTes and SemRelData in 50 random relations.....	66
Table 9.5 Disagreement analysis of relations that were only in the set extracted with the enhanced Hearst Patterns.....	68

Table 9.6 Contingency table denoting the number of noun compounds and transitive relations in the language subsets	69
Table 9.7 Distribution of Semantic Relation Types in different languages	69
Table 9.8 Proportional distribution of semantic relation types in different languages	69
Table 9.9 Contingency table denoting the number of noun compounds and transitive relations in the genre subsets	71
Table 9.10 Distribution of semantic relation types in different genres	71
Table 9.11 Proportional distribution of semantic relation types in different genres	71
Table 9.12 Contingency table denoting the number of noun compounds and transitive relations in the category subsets	72
Table 9.13 Distribution of semantic relation types in different subcategories of the encyclopaedic subset	72
Table 9.14 Proportional distribution of semantic relation types in different categories	72
Table 9.15 Frequency distribution of nominals within the relations of SemRelData of the word describing the subject of the article	74
Table 9.16 Distribution of nominals of SemRelData of the word describing the subject of the article ..	74
Table 9.17 Distribution of frequency placement of most frequent entities among classes	75
Table 9.18 Distribution of nominals in SemRelData of the word classified according to their function	75
Table A.1 Sources of literary subset	89
Table A.2 Sources of encyclopaedic subset	90
Table A.3 Sources of news texts	92
Table A.4 Contingency table of Annotator 1 and Annotator 2	92
Table A.5 Contingency table of Annotator 1 and Annotator 4	92
Table A.6 Contingency table of Annotator 2 and Annotator 3	92
Table A.7 Contingency table of Annotator 2 and Annotator 4	92
Table A.8 Contingency table of Annotator 1 and Curator	93
Table A.9 Contingency table of Annotator 2 and Curator	93
Table A.10 Contingency table of Annotator 3 and Curator	93
Table A.11 Contingency table of Annotator 4 and Curator	93
Table A.12 Paragraph size/k correlation	96
Table A.13 Detailed disagreement analysis of WordNet and SemRelData in 50 random relations	98
Table A.14 Detailed disagreement analysis of GermaNet and SemRelData in 50 random relations...	99
Table A.15 Detailed disagreement analysis of RuTes and SemRelData in 50 random relations	100
Table A.16 Result of Hearst-Pattern application on the raw data of SemRelData	101

Table A.17 Detailed disagreement analysis of pattern-based approach and SemRelData in 50 random relations	102
Table A.18 Detailed analysis of entities with the highest number of relations in the encycloaedic subset	105
Table A.19 Detailed analysis of frequent nouns in the encyclopaedic subset	107
Table A.20 Detailed analysis of entities with the highest number of relations in the literary subset (1) is person/character; 2) is description of person/character; 3) is description of location; 4) is feeling/condition; 5) is other)	110
Table A.21 Detailed analysis of frequent nouns in the literary subset (1) is named entity 2) is person/character; 3) is description of person/character; 4) is description of location; 5) is feeling/condition; 6) is other)	114

List of Figures

Figure 1 Example of default output of <i>trousers</i> in WordNet	32
Figure 2 Image of BabelNet output to the search term <i>trousers</i>	34
Figure 3 Ontology of <i>trouser</i> up to the second level of semantic relatedness on BabelNet (only classical semantic relations considered).....	35
Figure 4 Example of a pre-annotated sentence in the first part of the annotation	48
Figure 5 Example of pre-annotated sentence in the second part of the annotation	48
Figure 6 Prototypical workflow as implemented in WebAnno (Yimam et al., 2014)	50
Figure 7 Annotation of <i>hose_en.tsv</i> for SemRelData, showing all four possible semantic relation tags	50
Figure 8 Example 1.1.1 of synonymy in guidelines.....	55
Figure 9 Example 1.2.1 of co-hyponymy with the in common hypernym family member in guidelines	55
Figure 10 Example 2.1.1 of hyperonymy in guidelines	55
Figure 11 Example 2.3.4 of holonymy in guidelines.....	56
Figure 12 Paragraph size/k correlation	58
Figure 13 Postprocessing relation extraction rules	60
Figure 14 Graphical visualization of relations of <i>handbag</i> . Annotated relations are marked in black; annotations added in the post-processing are marked in green. Reverse annotations are not displayed.	61
Figure 15 Graphical visualization of relations of <i>hose_en.tsv</i> in SemRelData. Synonyms are marked in green, hyponyms are marked in orange, and meronyms are marked in black.....	64
Figure 16 All relations in <i>hose_en</i>	97

Glossary

Term	Definition
annotation	<p>A linguistic annotation is a notion that adds analytic or descriptive information on raw language data (Bird & Liberman, 2000).</p> <p>An exemplary annotation that is performed in this thesis is the annotation of noun compounds.</p> <p>For example, if the raw language data is <i>orange tree</i>, the annotation would mark it as a noun compound.</p> <p>Both the notion and the process of marking the notions are referred to as annotation.</p>
annotator	<p>An annotator is a person who performs an annotation.</p>
curator	<p>A curator is a person who performs the final annotation, by checking the results of several annotators against each other and also by adding new annotations.</p>
contingency table	<p>A contingency table, also known as cross tabulation, cross tab or confusion matrix, is a table, in which the entries in the rows tabulate the data to one variable, whereas the entries in the columns tabulate another ("Contingency Table", 2015).</p> <p>Here, such tables are used for the calculation of two metrics:</p> <p>In the calculation of κ this table is used for the calculation of agreement between two annotators, the row entries tabulating the annotations of one annotator, the columns the annotations of the other.</p> <p>In the calculation of χ^2 this table is used for the study of correlation and distribution of the semantic relations and their types.</p>
entity	<p>An entity is a particular and separate unit. Here, an entity is marked with the help of an annotation ("Entity", 2015).</p> <p>For example, <i>orange tree</i> is annotated as a noun compound and constitutes an entity.</p>
inter-annotator agreement	<p>This measure reflects the consensus of the annotation of the same text by two annotators.</p>
iterative	<p>An iterative process brings a result successively closer to a desired result through repetition ("Iteration", 2015).</p>
label	<p>Here, a label is the class that the raw language data has been annotated with.</p> <p>For example, in the sentence <i>An orange is a tree</i>, the label of the relation annotation of <i>orange</i> and <i>tree</i> would be <i>hypernym</i>.</p>
lemma	<p>In morphology, lemma is the word form which is not inflected. It is a dictionary form of a set of words, forming the head word of this set in a dictionary.</p> <p>For example, <i>find</i> is the lemma for <i>found</i>, <i>finds</i> and <i>finding</i>.</p> <p>The automatic process of finding lemmas is called lemmatization.</p>

lexicalized	A free, grammatically irregular composition of words that has been transformed in a formal or semantically idiomatic expression is called lexicalized ("Lexicalization", 2015).
macro-averaging	<p>Macro-averaging is the process of averaging of already calculated values.</p> <p>In the case of kappa calculation, the macro averaged value is calculated in the following way:</p> <ol style="list-style-type: none"> 1. The κ of all files of two annotators are calculated separately. 2. These κs are added up and divided by the number of files. The result of the division is the macro-averaged κ.
micro-averaging	<p>Micro-averaging is the process of averaging of raw values.</p> <p>In the case of κ calculation, the micro averaged value is calculated in the following way:</p> <ol style="list-style-type: none"> 1. All files annotated by two annotators are merged in one file. 2. The κ of this file is the micro-average κ.
morphology	Morphology is the linguistic field which concerns itself with word structure.
named entity	Named entity is a term for proper noun. It denotes names of persons, places, organizations and others.
natural language	A natural language is a language that is or was used by humans, e.g. the natural languages used here are <i>English</i> , <i>German</i> and <i>Russian</i> . Counterexamples of natural language are programming languages, e.g. the ones used here <i>Perl</i> , <i>Python</i> , and <i>Java</i> , and constructed languages, e.g. <i>Dothraki</i> , <i>Esperanto</i> , and <i>Klingon</i> .
nominals	In this thesis, nominals will be used as a term encompassing both <i>complex nominals</i> and <i>simple nouns</i> . Levi (1978) defines <i>complex nominals</i> as a term including <i>nominal compounds</i> , <i>noun compounds</i> , <i>nominalizations</i> and <i>noun phrases with nonpredicating adjectives</i> . The term <i>nominal</i> is chosen, because some definitions restrict noun to a single orthographic unit.
ontology	In computer and information science, an ontology is a conceptualization of domains of knowledge. In an ontology, entities are structured and among other representational terms, presented through relations to other entities (Gruber, 1993).
phonology	Phonology is the linguistic field which concerns itself with sound and their usage and meaning in language.
regular expression	<p>A regular expressions is defined as "An expression that describes a set of strings (= regular language) or a set of ordered pairs of strings (= a regular relation). [...] Also called a rational expression." (Mitkov, 2004, p. 754).</p> <p>For example, a regular expression used in this thesis is <i>adjective*noun+</i>. The first part, <i>adjective*</i>, denotes a sequence of adjectives, the '*' denoting an arbitrary length, including 0. The second part, <i>noun+</i>, denotes a sequence of nouns, the '+' denotes a length of at least 1. This means that the regular expression refers to any phrase that consists of at least one noun, including preceding adjectives and following nouns, like <i>important football match</i>.</p>

reflexivity	<p>Reflexivity is a property describing a relation that is turned back ("Reflexivity", 2015).</p> <p>This property is best exemplified with the synonymy relation. If a <i>handbag</i> is synonym of <i>purse</i>, than <i>purse</i> is also the synonym of <i>handbag</i>.</p>
semantics	Semantics is the linguistic field which concerns itself with meaning.
signified	<p>A signified is one of the two parts of de Saussure's theory of signs. The signified is the concept that the signifier, which is the phonetic component of the word, describes.</p> <p>For example, the signified of <i>orange tree</i> would be a mental concept of it that appears in the human mind, whereas the IPA represented phonetic transcription [ɔrændʒ tri] is the signifier.</p>
synset	A synset denotes a set of synonyms in WordNet (Miller, 1995).
tag	<p>In this thesis, tag is used similarly to label. A tag is additional information that is automatically added to a text item. This process is called tagging.</p> <p>For example, a part-of-speech tag of <i>bag</i> is <i>noun</i>.</p>
token	<p>A token is a meaningful element of text, in the case of this thesis it is a word. The automatic process of breaking a raw text into tokens is called tokenization.</p> <p>In contrast to type, token counts every occurrence of a word.</p> <p>For example, the sentence <i>I saw the dog chase the cat.</i> has 7 tokens, but 6 types, because there are 7 words and 6 different words.</p> <p>The type-token ratio is a measure for lexical diversity.</p>
transitivity	<p>In mathematics, transitivity describes the property of transfer of relations. If <i>a</i> relates to <i>b</i> in the same way that <i>b</i> relates to <i>c</i>, then <i>a</i> relates to <i>c</i> in the same way as to <i>b</i> ("Transitive", 2015).</p> <p>For example, if <i>bag</i> is a hypernym of <i>handbag</i> and <i>handbag</i> is a hypernym of <i>clutch</i>, then <i>bag</i> is also a hypernym of <i>clutch</i>.</p>

1. Introduction

In this thesis, classical semantic relations between noun compounds denoting types, parts and similar words will be annotated and analysed contextually in a multilingual and multigenre setting. The title of this thesis, *SemRelData*, is an abbreviation of *Semantic Relation Dataset*, which refers to the dataset of classical semantic relations that are created and analysed. This thesis aims at investigating the nature of such relations with respect to their impact on human knowledge representation in text. The variables of language and genre allow a universal analysis of the investigations. In this way, not only peculiarities of semantic relations in specific genres or languages, but also the nature and impact of classical semantic relations between nouns in general can be researched.

Semantic relations, present in texts of any genre and language, are relevant to the representation of information in text. They structure information in a human-understandable way, e.g. by establishing word hierarchies. The semantic relations considered in this thesis are restricted to nominals, and some of the observed classes are umbrella terms, containing several smaller subclasses of relations. *Synonyms* are mostly defined as different words with the same meaning, e.g. *handbag* and *purse*. *Hypernyms* are superordinate terms to their subordinate *hyponyms*, e.g. *bag* is a hypernym of *handbag*, and *handbag* is a hyponym of *bag*. *Co-hyponyms* are words with the same hypernym, e.g. *handbag* and *paper bag* having the mutual hypernym *bag*. *Holonyms* are terms referring to the whole, which consists of *meronyms*, its parts, e.g. *handbag* is a holonym of the meronym *handle*.

Due to their relevant role in information representation, the relations investigated in this thesis are important to information processing, both for humans and for computers. Thus, the improvement of techniques that automatically extract semantic relations can be expected to increase the performance of automatic information retrieval in general. Search engines such as Google can be viewed as the most common example of tasks that require information retrieval. Although automatic information retrieval systems already make use of classical semantic relations, the existing methods leave space for improvement, as they typically neglect context. Furthermore, contextual semantic relations have not been analysed with focus on different features such as genre or language. These different aspects may initiate new approaches towards classical semantic relations and may consequently not only improve the linguistic understanding of these, but also their automatic extraction, which would result in an improvement of information retrieval techniques.

More specifically, this thesis deals with classical semantic relations, such as synonyms, hypernyms, hyponyms, co-hyponyms, holonyms, and meronyms, between nominals in three languages – English, German and Russian. The relations were manually annotated and subsequently analysed. Noun phrases and their relations were annotated within paragraphs extracted from online freely available texts of different genres.

All of the investigated relations play a big role in both past and current research on semantic relations. It is assumed that semantic relations are important to the organization of the human mental lexicon. In text, they have a correlation with the notion of understanding written information.

The central question of this thesis is whether classical semantic relations between nominals have a crucial role in the linguistic representation of information. To answer this question, the following questions have to be addressed:

What role do semantic relations play in the representation of knowledge and information?

Is the use of semantic relations and semantic relation types universal, or rather dependent on language, culture or genre?

Can a uniform structure for the annotation of this task be found?

Can this contextual approach find relations other than those obtained by previous approaches?

Do terms with many relations have a special function in the text?

The analysis of the dataset will deal with the comparison of the distribution and types of classical semantic relations in different languages and genres. Moreover, terms having a high number of relations will be analysed with a special focus on their context.

The corpus consists of texts extracted from three different genres – encyclopaedic texts, extracted from Wikipedia; newspaper articles, extracted from Wikinews; and literary texts, which are out of copyright.

One of the main steps of the thesis, the annotation of the dataset, will be performed according to guidelines, which are a side-product of the thesis. To ensure the quality of this step, it will be performed in double annotation by a student annotation team, followed by a tool-supported curation step. The annotation and the development of the guidelines is a challenging step of this project because of the innovative approach of this thesis. The iterative improvement shown by the κ -metric and also the use of the κ -metric in this context will be likewise discussed.

To evaluate the annotated dataset, the result will also be compared with WordNet, GermaNet and RuTes which are the largest manually created or revised knowledge resources for the respective languages. Due to its contextual approach, the resulting dataset has the potential of detecting semantic relations which have so far not been listed in knowledge-based resources. Especially the non-encyclopaedic sources may show valid relations which would never occur in classical knowledge-based resources. As a result of this promising perspective, one may on the one hand detect, or rather mark new knowledge, on the other hand one may use it for information extraction tasks.

Afterward the occurrences and frequencies of different semantic relations within different languages, genres, Wikipedia categories and of more or less frequent terms of the same category in Wikipedia are examined in order to answer the described research questions. The comparisons will be performed using χ^2 .

This thesis has the aim to investigate whether semantic relations between nominals play a crucial role in the linguistic encoding of knowledge, but also to show that linguistic variation such as genre and language is reflected by the distribution and type of semantic relations between nominals. Moreover, the results may reflect the distance of the relations between genre types and languages. These results may help in the understanding of knowledge and knowledge creation in the context of language, reader community and genre.

1.1. Thesis Structure

This section briefly introduces the structure of this thesis by summarizing the content of the following chapters in order to provide a possibility of orientation.

Chapter 2 introduces the state of the art on classical semantic relations by presenting different approaches of various scientific fields towards semantic relations in Section 2.1. Section 2.2 discusses classical semantic relations by presenting different kinds of definitions and approaches towards each type of classical semantic relations that is of interest in this thesis. Based on the definitions and approaches presented in this section, Section 2.3 presents implementations of these relations by

demonstrating the use of patterns for the extraction of classical semantic relations on the example of Hearst Patterns in Section 2.4 and showing examples of well-known knowledge bases in 2.5, as well as semantic web ontologies in 2.6.

Chapter 3 presents the methods that will be applied in this thesis. Section 3.1 will present Cohen's κ -metric, which will be further used to calculate inter-annotator agreement, as well as the impact of different variables such as time, language, genre, and text size on the performance of the annotators. Section 3.2 presents the χ^2 -test, which will be used for the comparisons of distribution of semantic relations and their types.

The collection of the dataset as well as the titles of the texts annotated in this thesis will be discussed in Chapter 4. The important features limiting the choice of texts for the collection are presented in individual sections of this chapter.

Section 5 will describe the steps that were made in order to prepare for the annotation, mainly through POS-tagging and formatting.

Section 6 will deal with one of the two main tasks of this thesis, namely annotation. The annotation tool that was used for this thesis will be presented in Section 6.2, whereas the next section, 6.3, will explain the annotation process, which consists of three steps demonstrated in the three subsections. Section 6.4 presents the creation of the guidelines for the annotation. As the annotation is actually based on two annotation layers, one of which, noun compounds, has not been defined yet. This will be done in Subsection 6.4.1. Subsection 6.4.2 will explain the iterative approach that was employed here. Section 6.5 will present the inter-annotator agreement calculated with the κ -metric. The subsections of this section will demonstrate the possible impact of various factors on annotation.

Chapter 7 presents the steps taken in order to extract the relations that were previously annotated to SemRelData and the statistics of the dataset.

Chapter 8 presents the statistics and characteristics of the annotated dataset.

The results of the analysis are illustrated in Chapter 9. Each of the sections of this chapter describes a comparison of different subsets or entities. Sections 9.1 to 9.5 each show a separate comparison of both distribution of semantic relations overall as well as the distribution of the different types of subsets using χ^2 . Section 9.6 shows a comparison of the most frequent entities within the relations in SemRelData with respect to their function in context.

Chapter 10 summarizes and discusses the results of both annotation and analysis.

Chapter 11 presents possible applications of SemRelData as well as further research issues that could not be addressed in this thesis due to time and scope restrictions.

2. State of the Art

This chapter gives an overview of the past and current studies on classical semantic relations. This will demonstrate their importance in the variety of fields in which they are present. Subchapter 2.1 gives an introductory overview of the approaches of different scientific fields towards semantic relations. Subchapter 2.2 outlines the importance of classical semantic relations and presents several approaches towards these relations. The sections of this subchapter deal with each of the relations analysed in this thesis individually, presenting different approaches and definitions of each relation. Subchapter 2.3 briefly introduces computer scientific implementation approaches towards classical semantic relations, which most frequently use a pattern-based approach for the contextual extraction of relations. Some of the first and most frequently used patterns are Hearst Patterns, which are presented in Subchapter 2.4. Based on the implementations and patterns presented in these chapters, knowledge bases and semantic web ontologies were created. As distinctions between the definitions of these two terms are vague, the definition of the respective knowledge base was used in this thesis. Knowledge bases are presented in Subchapter 2.5. Individual knowledge bases are presented in the sections of this subchapter. Semantic web ontologies are presented in Subchapter 2.6, including presentations of individual ontologies in its sections.

2.1. Semantic Relations

Semantics, which studies meaning, is one of the most fundamental parts of linguistics. In particular, semantics is vital for the study of language acquisition or language change. As social context is likely to affect meaning, semantics is essential for understanding language varieties and style (Moore, 2000).

Semantic relations have been subject to many research fields, such as philosophy, cognitive psychology, linguistics, anthropology, early childhood and second language education, computer science, literary theory, cognitive neuroscience and psycholinguistics. The methods, definitions, perspectives and research questions vary from field to field and also within fields, but borrowing and transdisciplinary approaches exist. The consensus that can be found between most involved parties is that paradigmatic semantic relations¹ such as the classical semantic relations among words (Murphy, 2003):

... are somehow relevant to the structure of lexical or contextual information. Beyond this vague statement of “relevance,” however, opinions, assumptions, and models vary drastically. For some investigators (e.g. Katz 1972, Kempson 1977, Pustejovsky 1995) accounting for such relations is one of the purposes of lexical semantics [...]. For others (e.g., Deese 1965, Lehrer 1974, Mel’čuk 1996, Fellbaum 1998c) relations among words constrain or determine meaning, rather than vice versa. (Murphy, 2003, pp. 4–5).

As outlining all of these approaches would be out of scope, only those approaches that are relevant for this study will be briefly discussed.

In linguistics, many structural semantic approaches have closely dealt with paradigmatic relationships. According to the founder of structuralism, de Saussure, the study of relations is fundamental to the study of language, as words must be related to other words in order to have a meaning. However, de

¹ Paradigmatic relations are sets of words that form some sort of paradigm, that have some semantic characteristic in common (e.g. part of speech), but are incompatible in others (e.g. word form) as for example the synonymous relationship between *truck* and *lorry*. They are opposed to syntagmatic relations, which are sets of words that go together in a syntactic structure, as for example the relation between *drive* and *car* (Murphy, 2003).

Saussure did not distinguish between relation types or classifications (1996). In Neo-Humboldtian ethnolinguistics, lexicalization patterns were compared across languages with the aim to find lexical structures that represent individual culturally characteristic conceptualizations of the world. As reported by Murphy, this tradition was furthest developed by Trier. However, Weißenberger highlighted the ethnological perspective of language influencing thought. Lyons and Cruse, both focusing on paradigmatic relations "... have given linguistics its most exhaustive definitions and descriptions of semantic relations." (Murphy, 2003, p. 67). According to Lyons, a lexical item may be defined through its relations to other items in the same lexical system (as cited by Murphy, 2003). Cruse states that "... the meaning of a word is constituted by its contextual relations." (Cruse, 1986, p. 16).

Anthropological approaches use interviews as a source of semantic information. Studies of that kind often focus on folk taxonomies. According to Murphy (2003), early studies assumed that only advanced, literate languages had a taxonomic² organization of the world, based on the legends like Eskimos not having a term for snow, but 100 words for different kinds of snow. However, these assumption were proven wrong and it could be shown that lexicons of all languages have a taxonomical organization (Kay, as cited by Murphy, 2003). This triggered the question whether those taxonomies are universal or culture-specific. Moreover, such studies produced *dictionaries* which were not alphabetically structured, but semantically linked. This led to an increase of taxonomies, numbers of semantic relations and network representations of such, and computer scientific implementations which will be further described in the Chapter 2.3. To circumvent the problem of choosing non-representative or irrelevant relations, Casagrande and Hale introduced a new method of finding semantic relation types. In this research, Papago³ speakers were asked to write definitions for a set of words. In the next step, the relations communicated in the definitions were classified which resulted in 13 classes of relations, including some of the classical semantic relations (Casagrande and Hale, as cited by Murphy, 2003). According to Murphy, the difference between Casagrande and Hale's list of relations and that of classical semantic relations is rather a difference in the definition of relations and their boundaries than essential differences (2003).

The pragmatic and psycholinguistic approach has the main focus on finding a mental representation of semantic relations by assuming that words must be examined in context. Two basic points of view have developed in order to find these representations. Either semantic relations are facts that humans know or they are derived from other world knowledge. The first possibility would mean that learners acquire the knowledge of relations as facts in the same ways as they acquire other facts about words, like pronunciation or part of speech. More concretely this would mean that a learner knows that *life* and *death* are antonyms⁴ because he heard them being used in contrast and thus subconsciously saved this information in his mental lexicon. The second possibility would mean that knowing that *life* and *death* are antonyms involves a rule-generated representation for the generation of semantic relations among words, which is used every time the knowledge is needed. Although admitting that neither of these possibilities is a unique explanation to our linguistic performance concerning semantic relations, Murphy employs the second perspective, a meta-lexical approach to semantic relations, which defines relations of words not being represented in the lexicon, arguing that "(a) They are not relevant to linguistic competence; (b) they depend on the context in which they occur; and (c) they are predictable by means of single relation principle." (Murphy, 2003, p. 25).

² A taxonomy is a classification scheme which organizes objects hierarchically. The difference between scientific and folk taxonomies is not clearly defined. However, folk taxonomies generally are not bound to scientific definitions but rather to human judgement (Murphy, 2003). Thus, palm trees may be regarded as a kind of trees, although biologically they are rather a kind of grass.

³ Papago is an Uto-Aztecan language.

⁴ Antonymy is the relation describing contrast.

Murphy disagrees with the prevalent opinion represented in literature on lexical semantics that claim that semantic relations are not really relations among words but relations among word senses. This contradiction is supported by the idea that many relations between words hold between many of their senses, such as for example *keys* being part of a *keyboard*, whether one uses it in the sense of a *musical instrument* or an *electronic device*. Although some texts call them sense relations (Lyons as cited by Murphy, 2003) or meaning relations (Allan as cited by Murphy, 2003), further on in this text this distinction in terms will not be made.

A classical philosophical approach to semantic relations is that of using them in logical analytic statements in order to determine whether assertions are true or false⁵. A more current approach in philosophy is performed by Marconi, who assumes that the ability to use words in semantically appropriate ways requires knowledge of how those words relate to things in the world and how words relate to each other (as cited by Murphy, 2003). Like Murphy, Marconi regards semantic relations as relations between not words and word meanings (Marconi as cited by Murphy, 2003; Murphy, 2003).

Many kinds of approaches have been developed to find mental representations of semantic relations, such as speakers' judgements of semantic relatedness, corpus-based studies of semantically related words, descriptions of semantic relations in thesauri and dictionaries, tests of computational models of lexical knowledge or occurrences in natural language acquisition (Murphy, 2003). As Murphy states "Each of the above sources of information has its own limitations." (2003, p. 7).

2.2. Classical Semantic Relations

The study of semantic relations may help to improve the understanding of the structures reflecting language variation and genre. There exist many types of relations between words, but a selection of these needs to be made for the sake of clarity. According to Girju et al., semantic relation classes have been mostly "... designed to maximize coverage [...] and minimize overlap [...].The ideal class scheme would be exhaustive (include all relations) and mutually exclusive (no overlapping classes)." (2009, p. 107).

The relations that are referred to as classical semantic relations are those that are called *traditional 'nym relations* by Murphy and one of their subtypes (2003). An exact definition of such relations is necessary for a task such as presented in this thesis. However, such a definition is not trivial. According to Cruse,

To be worth singling out for special attention, a semantic relation needs to be at least systematic, in the same sense that it recurs in a number of pairs or sets of related lexical units.[...] There are innumerable 'low level' semantic relations restricted to specific notional areas. (1986, p. 84).

He continues his statement by saying that a relatively small number of semantic relations, such as *synonymy*, *antonymy* and *hyperonymy*, has achieved a central role in lexical semantics (Cruse, 1986).

Murphy admits that "... the traditional 'nym relations receive the most attention here because of the rich literature on them and hence the increased opportunity to test the current treatment against observations about semantic relations from a number of different perspectives." (Murphy, 2003, p. 25). Furthermore, Murphy describes properties of semantic relations: productivity, binarity, variability, prototypicality and canonicity, semi-semanticity, uncountability, predictability and universality. While

⁵ Murphy gives the following examples for such sentences:

- a. No **unmarried** man is **married**.
- b. If it is a **rose**, then it is a **flower**.
- c. A **circular** shape is **round**. (2003, p. 63)

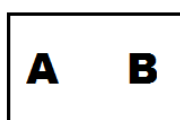
defining all these properties in detail may seem space-consuming, Murphy also offers the relational principle *relation by contrast*, which claims to be a paradigm for all semantic relations (Murphy, 2003). An example of the application of this principle to some of the classical semantic relations as shown by Murphy may be viewed in Table 2.1.

Relation	Relates	Similarity	Incompatibility	Example
Synonymy	words	meaning, syntactic category, register, etc.	word form	couch = sofa = divan = settee
Antonymy	words	semantic category, categorization, level, register, morphology, etc.	sense	rise/fall happy/sad life/death
Categorical Opposition	categories	semantic field, categorization level	categorization criterion	rise/go down happy/sad happy/angry
Hyponymy	categories or names of categories	semantic category	level of categorization	bird > [robin/swift/swan...]
Meronymy	categories or names of categories	same object	level of completeness	house > [wall/roof/floor/doors ...]
Grammatical Paradigm	words	lexeme, inflectional category type	inflection	drink - drank - drunk

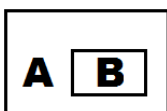
Table 2.1 Instantiations of relation by contrast (Murphy, 2003, p. 45)

Murphy admits that the level of completeness as contrastive difference in meronymy is not a satisfactory distinction, since a part can also be complete, as e.g. *tree*, which is also a meronym to *forest*. Moreover, Murphy claims that meronymy and hyponymy are not lexical relations, because they mostly refer to relations among concepts and things, whereas synonyms and antonyms refer to relations among words (2003).

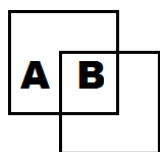
Cruse divides the basic lexical relations that are subject to this thesis in four relation variants: identity, inclusion, overlap and disjunction, which are demonstrated graphically below.



I. **identity:** class A and class B have the same members



II. **inclusion:** class B is wholly included in class A



III. **overlap:** class A and class B have members in common but each has members not found in the other



IV. **disjunction:** class A and class B have no members in common

(Cruse, 1986, p. 87)

The lexical relationship referring to identity is synonymy; the class reflected by inclusion is hyponymy. Co-hyponymy could be regarded as a relation with the relation variant overlap. Further descriptions of the individual relations types that are also used in this thesis are provided in the following subsections. Before continuing with these subsections, Cruse's notion of *unfull* relations shall be briefly discussed. As Cruse defines those relations for all semantic relations, the naming of those subdefinitions is performed in this section. All those kinds of relation that do not fulfil the requirements of full relations are applicable to one or two types of the further on described relations. Thus the more detailed

definition with examples of those relations will be conducted in the corresponding chapters. *Partial relations* “are relations which hold between lexical items whose syntactic distribution only partially coincide” (Cruse, 1986, p. 96). *Para-relations* are lexical relations defined in terms of expectation. Cruse describes the class of *quasi-relations*, which occur when a term fully fulfilling the requirements of the semantic relation is missing in the language, but an equivalent of the wrong syntactic category exists. The relation introduced as *pseudo-relation* by Cruse describes the relation between two lexical items being contextually restricted (Cruse, 1986).

2.2.1. Synonymy

Synonymy, or sometimes also referred to as *poecilonymy*, is regarded as the most significant relation in the WordNet model (Miller & Fellbaum, 1991). Murphy distinguishes between two different approaches to the definition of synonymy – through similarity or through contrast through similarity or through contrast (2003)⁶.

According to philosophic and psychological theories, relying on the definition of synonymy through similarity is meaningless (Goodman as cited by Murphy, 2003; Murphy and Medin as cited by Murphy, 2003), although it is regarded as an efficient device for the description process (Medin et al. as cited by Murphy, 2003). The definitions discussed in this thesis solely reflect the view of synonymy relating to just lexical entities, such as words and lexical units. However, it should be noted that synonymy may also relate to both morphological and syntactical entities. In the field of philosophy, synonymy mostly refers to relations among sentences or propositions (Quine, as cited by Murphy, 2003).

Murphy states that

Rather than defining synonymy on logical criteria, the RC-S⁷ definition reflects the types of sets that count as synonyms in real linguistic contexts (such as thesauri), since these rarely conform to definitions that require logical equivalence or mutual entailment. RC-S takes a pragmatic perspective on semantic relations [...], providing a means for identifying appropriate synonyms in situations where the context demands logical equivalence – and in those where it does not. (2003, p. 142).

Murphy defines synonymy as “A synonym set includes only word-concepts that have all the same contextually relevant properties, but differ in form.” (2003, p. 134). Murphy further states that the similarity of synonyms depends on their context, meaning that in this context the meaning of the words needs to be similar, having identical contextually relevant properties. For example, in the context of calculating available seats in the room, loveseat and sofa are not synonymous, as they by usual definition have a different number of seats. In any context where the number of seats is unimportant, they may be used as synonyms (Murphy, 2003). In Murphy’s definition synonymous relations between words such as end and ending are regarded as synonyms. Although Murphy also discusses the scale of similarity or better or worse synonyms, this will not be further discussed here (2003).

Moreover, Murphy regards synonymy as a not purely bi-directional relation, saying that sometimes synonymy can be hyponymous. To exemplify his point, Murphy gives the following example, stating that in (8) chair and seat are synonymous, in (9) they are not.

⁶ Werner, Apresjan, Kempson and Kreidler define synonymy purely through similarity, strictly speaking allowing the same word to be a synonym of itself (as cited by Murphy, 2003). In contrast to those notions, Katz, Harris, Cruse, Jackson, Chierchia and McConnel-Ginet and Hudson additionally define that the words in the synonymous relation must be two different words (as cited by Murphy, 2003).

⁷ Author’s note: By RC Murphy refers to Relation by Contrast and RC-S refers to Relation by Contrast Synonymy.

- (8) a. The receptionist indicated a chair where I should wait. →
 b. The receptionist indicated a seat where I should wait.
- (9) a. The receptionist indicated a seat where I should wait. →
 b. The receptionist indicated a chair where I should wait. (2003, p. 140).

Cruse defines terms in a synonymous relation in the following way:

X is a cognitive synonym of Y if (i) X and Y are syntactically identical, and (ii) any grammatical declarative sentence S containing X has equivalent truth-conditions to another sentence S¹, which is identical to S except that X is replaced by Y. (1986, p. 88)

Both Cruse and Murphy subclassify synonymy in several categories. *Full synonyms*, or *absolute synonyms*, as they are called by Lyons (as cited by Murphy, 2003), are words that are used equally in every context. Such synonyms “tend to be words with relatively limited numbers of conventionalized senses” (Murphy, 2003, p. 146). In natural language use, there is no need for terms that can be used completely interchangeably in all contexts. Mostly dialect, domain or linguistic register restrict the use of synonymous terms.

Thus Cruse talks of *partial relations*, or more specifically of *partial synonymy*. As described before, partial relations exist between only partly similar lexical items. To exemplify his point, Cruse names the partial synonyms *finish* and *complete*, which cannot be considered exchangeable in any context, e.g. *finish* being able to occur without a direct object⁸ (1986).

Murphy generally differentiates between *logical synonyms* and *context-dependent synonyms*, which is demonstrated in the table below.

	Identical senses (logical synonyms)	Similar senses (context-dependent synonyms)
All Senses	full synonyms	?
One (+) Sense	sense synonyms	near-synonyms (plesionyms)

Table 2.2 Dimensions of synonymy (Murphy, 2003, p. 146)

Logical synonyms share the same lexical or semantic representation. The subcategory of *full synonymy* was already discussed in the previous paragraph. Murphy’s example of full synonyms is *groundhog* and *woodchuck*. *Sense synonyms* are equivalent to Cruse’s *partial synonyms*. Murphy names *sofa* and *couch* as an example. *Context-dependent synonyms* are words that share the same meaning in some context. Near synonyms share no senses that are exactly the same, but one term in this kind of relation can be used to define the other, like e.g. *mob* and *crowd*. Near-synonyms are often found in thesauri (Murphy, 2003).

2.2.2. Hyperonymy and Hyponymy

According to Cruse, Lyons and Pustejovsky *hyperonymy*⁹ is one of the major structural relations (as cited by Murphy, 2003). Generally it is often paraphrased as the *kind-of* relation or as *set inclusion* in

⁸ *Finish* can be used in e.g. *Have you finished?* *Complete* however, needs a direct object e.g. *Have you completed X?* (Cruse, 1986). In Ishiguro’s *Never Let Me Go*, the *donors complete*, meaning that they die. Using *complete* without a direct object is a rhetorical device to demonstrate the unnatural action described in the novel.

⁹ *Hyperonymy* is the *token>type* relation, whereas *hyponymy* is the *type<token* relation (Murphy, 2003). In this thesis, the term *hyperonymy* is used preferably. However, if *hyponymy* occurs in quotations, it is left unaltered.

logical definitions. Hyperonymy is mostly defined as a unidirectional, non-reflexive and transitive¹⁰ (Murphy, 2003). Cruse gives counterexamples to the transitivity claim, with the example of *airplane* being a hypernym of *glider* and *glider* being a hypernym of *hang-glider*, but *airplane* not being a hypernym of *hang-glider*. However, the transitivity claim holds for taxonomic hyperonymy (Cruse as cited by Murphy, 2003). Furthermore, Murphy states that “Hyponymy is a central notion in many models of the lexicon due to its inference invoking nature, its importance in definition, and its relevance to selectional restrictions in grammar.” (Murphy, 2003, p. 217).

Murphy also declares that hyperonymy is important in our conscious reflection on word meaning (2003). Wierzbicka admits this notion, but nonetheless adds that the role of hyperonymy in human thinking is overestimated (1984). Murphy states, as already briefly discussed above, that hyperonymy is not a lexical-semantic relation as it relates concepts of things that words denote and not words (2003).

Further on, Murphy says that hyperonymy, like other relations, can be subdivided into several subtypes. The number and relevance of a full taxonomy is arguable and varies from definition to definition. The most common subcategorization, however, is between *taxonomic* and *functional hyperonymy* (Miller as cited by Murphy, 2003).

As mentioned before, Cruse describes the class of *quasi-relations*, which appear when “an exactly appropriate lexical partner that would complete a paradigmatic relationship is missing, but a lexical item exists, with virtually the required meaning, but of the wrong syntactic category.” (1986, p. 97). An example of a *quasi-hyperonymy* is there being no superordinate for *fork* and *spoon*. However, there is the mass noun *cutlery*, which could be considered as their hypernym in this thesis (Cruse, 1986). According to Murphy, it is dubitative whether paradigmatic relations may be characterized through sameness of syntactic category. Some definitions propose to allow members of different syntactic categories to be related on purpose (2003). To avoid such problems, Cruse proposes to treat hyperonymy as a prototypical relation in which taxonomy is treated as a fundamental subcategory (Cruse as cited by Murphy, 2003). Next to taxonomy and quasi-hyperonymy, Cruse defines para-hyperonymy. He states that “Whereas linguists normally frame definitions of lexical relations in terms of critical or canonical traits, natural language is very often satisfied with expected traits. A lexical relation defined in terms of expectation will be called a para-relation“ (1986, p. 99). He presents *para-hyperonymy* by the example of *dog* (*hyponym*) and *pet* (*hypernym*) (not all *dogs* being *pets*) (Cruse, 1986).

In *Apples are not a kind of fruit* Wierzbicka discusses the fallacy of considering functional concepts as a supercategory for the categorization of the language-encoded world. In her work, she discusses the categorization of concepts into unique taxonomies (one concept being part of only one other concept). Wierzbicka argues that the conclusion of apples being fruit is due to the assumption that all apples are fruit, but not all *fruit* are *apples*, which is logically correct, but does not imply a semantic relation (1984). The structures that she defines as non-taxonomic categories are of interest in this thesis, as this structures are similar to the subcategories of hyperonymy of Cruse. Further on she states that

Meaning is not a sum of shared properties of denotata — it is a conceptual structure. Not all the shared properties are conceptually relevant, and some conceptually relevant properties may be fictitious, based on prejudice, error, myth, symbolic associations, and so on. Thus, the fact that all apples are fruit and that all carrots are vegetables, and not vice versa, does not mean that conceptually apples are a kind of fruit or that carrots are a kind of vegetable. The conceptual relation “kind of” must be clearly distinguished from the referential relation of set inclusion. (Wierzbicka, 1984, p. 315).

¹⁰ There are *autohyponyms*, which are reflexive. Autohyponyms are words that have both a general and a specific sense, such as *dog*, referring to both the *animal in general*, but also to *male dog* as opposed to *bitch* (Cruse, 1986).

Wierzbicka argues that in a folk taxonomical classification, language users would not use hypernyms such as *animal* to refer to a *kangaroo*, but rather *creature*, as in natural language use not everything that biologically is an animal is referred to as one. She states that a *kangaroo* could as well be described as *hopper*. Wierzbicka argues that the crucial difference between functional concepts such as *animals* or *fruit* and taxonomic concepts in her definition is the possibility of picturing taxonomic concepts. One can draw a tree in general, without drawing an explicit tree, but not a fruit in general. It should be noted that Wierzbicka uses the device of imaginability to explain the difference between functional concepts, standing for a kind of function or a kind of thing, and concrete concepts, but she does not restrict taxonomy to picturable concepts. The concept *fruit* contains the notion of *any kind of*, whereas *apple* stands for a specific particular kind. Wierzbicka claims that “The failure to distinguish between taxonomic concepts and purely functional concepts leads to great arbitrariness in semantic description.” (1984, p. 318), as purely functional concepts and other non-taxonomic structures are fuzzy.

Another non-taxonomic structure as defined by Wierzbicka are *collective supercategories based on contiguity*. She argues that so-called partonomies are also present at the level of supercategories and are mistaken for taxonomies. She subdivides this category in *singularia tantum* and *pluralia tantum*. The category of *singularia tantum* is what Cruse defines as *quasi-hyperonymy* – class nouns relating to singular entities, e.g. *cutlery* referring to *fork* and *knife*. Wierzbicka argues that collective concepts cannot be included in countable concepts, by stating that

Of course, there is nothing wrong in saying that tables are a kind of furniture or that shirts are a kind of clothing, but statements of this kind must not be regarded as reflecting semantic structure. Semantically, tables are not a kind of furniture, shirts are not a kind of clothing, cups are not a kind of kitchenware, and so on. (Wierzbicka, 1984, p. 320).

*Pluralia tantum*s label heterogenous collections of things such as “goods, goodies, clothes, groceries, refreshments, odds-and-ends, bits-and-pieces, remains, belongings, supplies, trappings, trimmings, spoils, valuables, nuts-and-bolts (in the sense of party snacks), covers (bedcovers), dishes (as in “wash the dishes”).” (Wierzbicka, 1984, p. 321). According to Wierzbicka, all members of these collections are located together for some reason, which may be, but are not necessarily functional (1984).

Wierzbicka regards taxonomy as a hierarchy in which “all taxonomic concepts must be defined in terms of other taxonomic concepts.” (1984, p. 323), except for what Berlin et al. and Brown call *unique beginners* (as cited by Wierzbicka, 1984).

2.2.1. GermaNet

According to its official homepage¹¹, GermaNet is much similar to WordNet, consisting of subnets of nouns, adjectives and verbs linked by semantic relations. It has been developed since 1997 and is free for academic use. The license used for this work is that of the Language Technology Group of the Computer Science Department of the Technische Universität Darmstadt. The current version, 9.0, consists of 121,810 lexical units, 93,246 synsets and 105,912 conceptual relations (Henrich & Hinrichs, 2011). A similar German database is OpenThesaurus¹², which is available under the GNU license. However, it only provides relations such as synonyms and associations (Naber, 2004).

¹¹ <http://www.sfs.uni-tuebingen.de/GermaNet/>

¹² <https://www.openthesaurus.de/>

2.2.2. Holonymy and Meronymy

Holonymy¹³ describes the relation of the part-whole type. Cruse declares that holonymy is a relation that is more difficult to define than taxonomy, as there is no single clearly distinguished relation, but many similar relations (1986), which will be discussed below. Winston et al. state that meronymy has often been confused or not clearly distinguished from other semantic relations such as possession, attribution and class inclusion (1987). The consensus on the characteristics of holonymy is that it is an irreflexive and antisymmetric relation (Cruse, 1986; Winston et al., 1987). According to Murphy, holonymy has even fewer properties of a lexical semantic relation than hyperonymy and was not one of the relations identified in Casagrande and Hale's study that was discussed earlier (as cited by Murphy, 2003). Cruse's (in his own words too restrictive) general definition of meronymy is the following:

X is a meronym of Y if and only if sentences of the form *A Y has Xs/ an X* and *An X is a part of Y* are normal when the noun phrases *an X*, *a Y* are interpreted generically. (Cruse, 1986, p. 160).

To his definitions he adds that in meronymy all parts have to be of the same class, e.g. if the holonym is an abstract noun, so must be all its meronyms. Cruse gives the following more open definition: "The parts of a Y includes the X/Xs, the Z/Zs, etc." (1986, p. 161). Another crucial distinction that Cruse makes in order to define holonymy is the distinction between parts and pieces. The illustrative example clarifies this difference:

- a) hacking a typewriter into pieces
- b) unscrewing it into its parts.

The portions in a) are not considered meronyms of *typewriter*, whereas the ones in b) are considered such. Cruse argues that pieces do not fulfil sufficient requirements, such as stability, continuity and recreatability, and therefore do not qualify for lexical labels. Hence, further on only the notion of parts will be regarded. Cruse names the following characteristics that need to be fulfilled by a part:

- 1) It needs to theoretically belong to a denotable whole.
- 2) It needs to be limitable from other parts of the whole¹⁴.
- 3) The possession of a definite function relative to the whole.

According to Cruse, meronymy can be subclassified according to optionality and necessity of the relation, defining *canonical holonyms*, such as *body* is to *ear*, and facultative relations such as *door* to *handle*. Moreover, Cruse states that "A well-formed part-whole hierarchy should consist of elements of the same general type" (1986, p. 168). To do so, he differentiates between segmental parts, e.g. *trunk*, *head* and *limbs* in the *human body*, and systemic parts, e.g. *skeleton*, *muscles* and *nerves* in the *human body* (Cruse, 1986). Lyons distinguishes between several sub-classes of holonymy, such as *singular collections*, *plural collections* and *optional* and *necessary meronyms* (as cited by Winston et al., 1987). Nonetheless, Cruse admits that, unlike taxonomy, a holonymic relation is not a guaranteed well-formed hierarchy, because convergence cannot be excluded, as some meronyms may be parts of several hyponyms. One could try to avoid the problem of convergence by confining the elements to congruent pairs, accepting that with this restriction many relationships of interest would be excluded.

¹³ Holonymy is the *has-a* relation, whereas its opposite meronymy is the *is-part-of* relation. In this thesis the term holonymy is preferred (Murphy, 2003). However, if quotations contained the term meronymy, they were left unaltered.

¹⁴ Some parts, such as *wheels* of a *car* are more clearly detachable from the whole than others, such as *hip* from *thigh*.

Cruse believes that meronymy is applicable to three of the four classes of congruence that were discussed earlier in Section 2.1 – *identity*, *inclusion* and *overlap*. He addresses the problem of inclusion due to word ambiguity.

Further on, Cruse describes a subclass of meronymy that he calls *holo-meronymy*, where the term for the meronym may also describe the holonym. An example of this relation is the relation between *leaf* and *blade* (*blade* can describe the whole *leaf* or only a part of it, depending on whether there is a *stalk*).

The complications that exist in the holonymy relation are partly due to the question of transitivity – although holonymy is transitive, not all transitive relations are seen as sensible. The classical example of this is the relation between *house* and *handle*¹⁵. Cruse states that these transitivity failures are due to the difference between *attachments* (e.g. *handle* being an attachment of *door*) and *integral parts* (*palm* being an integral part of *hand*). The whole is destroyed as an entity if an integral part is missing, but this is not the case with attachments. As attachments can be integral parts of the whole, it is not trivial to determine when transitivity is semantically correct, but still noteworthy when discussing the problem.

Six Types of Meronymic Relations with Relation Elements

Relation	Examples	Relation Elements		
		Functional	Homeomeric	Separable
Component/ Integral Object	handle-cup punchline-joke	+	-	+
Member/ Collection	tree-forest card-deck	-	-	+
Portion/Mass	slice-pie grain-salt	-	+	+
Stuff/Object	gin-martini steel-bike	-	-	-
Feature/Activity	paying-shopping dating-adolescence	+	-	-
Place/Area	Everglades-Florida oasis-desert	-	+	-

Functional (+)/Nonfunctional (-): Parts are/are not in a specific spatial/temporal position with respect to each other which supports their functional role with respect to the whole.

Homeomeric (+)/Nonhomeomeric (-): Parts are similar/dissimilar to each other and to the whole to which they belong.

Separable (+)/Inseparable (-): Parts can/cannot be physically disconnected, in principle, from the whole to which they are connected.

Table 2.3 Subclasses of holonymy expressed through part-of (Winston et al., 1987, p. 421)

Cruse notes the existence of gaps in hierarchical relations, saying that terms for some elements in the hierarchy are missing. According to Cruse, in the case of meronyms, sometimes the most inclusive part lacks a term, e.g. the part of the *spoon* or *fork* that is called *blade* in a *knife*¹⁶.

¹⁵ Although *house* is a holonym of *door* and *door* is a holonym of *handle*, the functional meaning of *handle* is not applicable to higher points of the holonymic hierarchy.

¹⁶ It shall be mentioned that according to Merriam Webster, the discussed part of the *spoon* is called bowl (<http://visual.merriam-webster.com/food-kitchen/kitchen/silverware/spoon.php>) and the discussed part for *fork* consists of a *root* and *tines* (<http://visual.merriam-webster.com/food-kitchen/kitchen/silverware/fork.php>).

Winston et al. subclassify holonymic relations that are expressed through the English phrase *part-of* (not denying that there are also other ways to express holonymy linguistically). The result of their subclassification is presented in Table 2.3.

Moreover, Winston et al. state which relations are often misclassified as holonymy. These are *topological inclusion*, e.g. *wine* and *cooler* or *prisoner* and *cell*, *class inclusion*, e.g. *cars* and *vehicle* or *roses* and *flowers*, *attribution*, e.g. *tower* and *height* or *hair* and *colour*, *attachment*, e.g. *earrings* and *ears* or *picture* and *wall*, and *ownership*, e.g. *millionaire* and *money* or *author* and *ownership* (1987).

In contradiction to Cruse (1986), Winston et al. (1987) support Halmos and Moore in regarding holonymy as a transitive relation (as cited by Winston et al., 1987), with the restriction that the holonymy is within one subclass. Consequently, they conclude that transitivity and the other characteristics make holonymy “particularly important to our understanding of the structure of the lexicon since, as a partial ordering relation, like class inclusion, meronymic relationships structure semantic space in a hierarchical fashion.” (Winston et al., 1987).

2.3. Implementations of Semantic Relation Classification

In linguistics, the task of recognition and classification of semantic relationships between nouns has been conducted in different forms, their results being used for further natural language processing tasks or knowledge base creation.

Rosario and Hearst (2001), Rosario et al. (2002), Nastase and Szpakowicz (2003), Girju et al. (2007), and Davidov and Rappoport (2008) performed a classification of relations between nouns in compounds. Turney and Littman (2005) and Hendrickx et al. (2009) performed a recognition and classification between pairs of nominals. Most of the above listed works used patterns to automatically extract the relations. As stated by Davidov and Rappoport, “a leading method for utilizing context information for classification and extraction of relationships is that of patterns (Hearst, 1992; Pantel and Pennacchiotti, 2006)” (2008, p. 227). The so-called Hearst Patterns will be presented in detail in the next chapter. Another contextual approach is presented by Biemann et al. (2004), who introduced a machine-learning approach that learns semantic relations on the basis of collocations.

Using these automatic extractions, knowledge bases such as BabelNet (Navigli & Ponzetto, 2012), Mimida Project (Gittens, 2005) and NELL (Zimmermann, Gravier, Subercaze, & Cruzille, 2013) were created. The first two projects integrate WordNet (Miller, 1995; Fellbaum, 1998), a large manually created lexical database of English.

2.4. Hearst Patterns

Many of the below listed knowledge bases and ontologies make use of patterns to automatically extract semantic relations from continuous text. Based on the previously described assumption of semantic relations involving rule-generated representation, Hearst (1992) was one of the first to create such patterns for the automatic detection of hypernym relations between nouns. The patterns were created by thorough observation of texts and the setting of the contained relations. Attempts to build analogous patterns for holonymy were barren of results (Hearst, 1992). The five relations that are known as *Hearst Patterns* are listed below:

- (1) ... *such NP as {NP ,}* {(or [and])} NP*
... works by such authors as Herrick, Goldsmith, and Shakespeare.
→ *hyponym* ("author", "Herrick"), *hyponym* ("author", "Goldsmith "), *hyponym* ("author", "Shakespeare")

(2) *NP {, NP} * {,}* or *other NP*

Bruises, wounds, broken bones or other injuries...

→ *hyponym* ("bruise", "injury"), *hyponym* ("wound", "injury"), *hyponym* ("broken bone", "injury")

(3) *NP {, NP}* {,}* and *other NP*

... temples, treasuries, and other important civic buildings.

→ *hyponym* ("temple", "civic building"), *hyponym* ("treasury", "civic building")

(4) *m, {,}* including *{NP ,}* {or | and} NP*

All common-law countries, including Canada and England...

→ *hyponym* ("Canada", "common-law country"), *hyponym* ("England", "common-law country")

(5) *NP {,}* especially *{NP ,}* {or | and} NP*

... most: European countries, especially France, England, and Spain.

→ *hyponym* ("France", "European country"), *hyponym* ("England", "European country"), *hyponym*("Spain", "European country")

(Hearst, 1992, p. 541), numbering changed by the author of the thesis.

These patterns have been enhanced by Mititelu (as cited by Klaussner & Zhekova, 2011). Klaussner and Zhekova have used the best-rated enhanced patternset in order to create an ontology of Wikipedia articles. In this study they concluded that the applied patterns are often ambiguous, insufficient and not hyperonymy-specific (Klaussner & Zhekova, 2011).

2.5. Knowledge Bases containing Semantic Relations

As already described in the previous subchapter, knowledge bases containing semantic relations were created in various ways. In the following, both manually created databases such as WordNet and its German and Russian counterparts GermaNet and RuTes, as well as automatically created bases, such as BabelNet and NELL, are presented. Table 2.4 gives a size comparison of those databases. The sizes were retrieved from the respective webpages.

Additionally to the databases presented in detail, the notion of computer scientific ontologies will be discussed. The following subdisciplines of computer and information science built ontologies to efficiently organize information and reduce complexity: artificial intelligence, Semantic Web, systems and software engineering, biomedical informatics, library science, and information architecture (Noy & McGuinness, 2001). The term ontology describes a structure that organizes types, properties and relations among entities that are subject to a specific domain. For better understanding of the structure and content of those resources, exemplary entries are shown in English. However, some resources are available in other languages, as will be described below.

Knowledge Base Type	Knowledge Base	#words (lemmas)	#relations/#facts
Manually created Knowledge Base	WordNet 3.0	155,287	206,941
	GermaNet 9.0	121,810	105,912
	RuTes	153,561	219,576
Automatically / Semi automatically created Knowledge Base	Freebase (retrieved 08.02.2015)	47,000,000	2,696,000,000
	BabelNet 3.0 (English version)	11,000,000	354,000,000
	YAGO (3)	10,000,000	120,000,000
	DBpedia(English 2014 version)	4,580,000	583,000,000
	NELL (02.2015)	unk	2,000,000

Table 2.4 Size comparison between different databases

2.5.1. WordNet

The collection of the manually created database started in 1985. It consists of so-called synsets, which are collections of cognitive synonyms. These synsets are linked to other synsets in the database through semantic relations. It is the largest freely available database of this kind and is widely used in linguistic and natural language processing tasks, e.g. in the creation of other knowledge bases such as BabelNet or Mimida, or in tasks such as word sense disambiguation, information retrieval, automatic text classification, automatic text summarization, machine translation, semantic relatedness and similarity between words and documents. As WordNet is widely used, a Java API as well as an access through the Python NLTK (Bird, 2006) platform have been made freely available. The 3.0 version of WordNet consists of 155,287 words, 117,659 synsets and 206,941 relations.

The knowledge base can be accessed online through a graphical user interface¹⁷, but can also be downloaded for further processing, both in a user interface and an XML database (Miller, 1995; Fellbaum, 1998).

The words in WordNet are part-of-speech (POS) tagged. The majority of relations are between words belonging to the same POS. The database mainly consists of four subnets, those of nouns, verbs, adjectives and adverbs. Some of the relations linking nouns in WordNet are hyperonymy, holonymy, synonymy and antonymy. The creators of WordNet responded to the criticism of not differentiating between proper nouns and nouns in relations (Gangemi et al., and Oltamari et al., as cited by Miller & Hristea, 2006) by introducing this distinction in Version 2.1 (Miller & Hristea, 2006). The reasons for this criticism will be further described in Section 6.4.

The concept of WordNet was also used in the creation of similar databases in other languages, which can be found in OpenMultilingual WordNet¹⁸. Two of these will be described in the following sections.

The default output of WordNet, which is offered in the online application when no other restrictions were chosen by the user, returns all senses of the searched word. The output for *trousers* is shown in Figure 1.

The noun *trouser* has 2 senses (first 1 from tagged texts)

1. (3) *trouser*, *pant* -- ((usually in the plural) a garment extending from the waist to the knee or ankle, covering each leg separately; "he had a sharp crease in his trousers")
2. *trouser* -- (a garment (or part of a garment) designed for or relating to trousers; "in his trouser's pocket"; "he ripped his left trouser on the fence")

Figure 1 Example of default output of *trousers* in WordNet

However, WordNet holds more information pertaining semantic relations of words. The internal representation of the data is not trivial to understand, thus a more intuitive representation of some of the knowledge on the first sense of *trousers* will be presented below in order to provide an idea of WordNet's structure. The lists of hyponyms and meronyms presented in Table 2.5 were cut due to space limitations.

¹⁷ <http://wordnetweb.princeton.edu/perl/webwn>

¹⁸ <http://compling.hss.ntu.edu.sg/omw/>

Relation class	Related Word							
Synonyms	pant							
Hypernyms	garment <	clothing (and synset) <	Covering <	artifact < (and synset)	whole < (and synset)	object < (and synset)	physical entity < (and synset)	
			consumer goods <	commodity < (and synset)	artifact < (and synset)	whole < (and synset)	object < (and synset)	physical entity (and synset)
Hyponyms	bellbottom trousers							
	breeches > (and synset)	britches						
		buckskins						
		plus fours						
		trunk hose						
	chino							
Meronyms	hip pocket							
	lap covering							
	trouser leg							

Table 2.5 Exemplary extract of the relations of *trousers* in WordNet, with hyperonymic relations of all terms

2.5.2. RuTes

RuTes is an on-going project since 1994 aimed at creating a hierarchical linguistic resource, which in contrast to WordNet was not created in order to represent human knowledge, but as a natural language processing resource. The current version holds 158,000 terms, organized in 55,000 subsets and more than 210,000 relations. The version that is used in this thesis, RuTes-light is a subset of the full thesaurus, holding over 107,000 relations, 97,000 terms and 26,000 subsets.

It was created through an automatic extraction and a subsequent manual correction of terms and relations retrieved from the normative documents of the Russian Federation. The data is further enhanced through disambiguation tasks, lemmatization of the terms, further relations and words that are found through works based on RuTes (Loukashevich, 2011). RuTes is available under the Attribution-Non-Commercial-Share-Alike 3.0 licence¹⁹.

There is a publicly available Russian version of WordNet, but it was not manually created like WordNet and GermaNet which are used for the comparison with SemRelData. The creators wrote an algorithm, which automatically translated the original version and cleaned the result of concepts which do not exist in the Russian language (Gel'venbeyn, Goncharuk, Lehel't, Lipatov, & Shilo, Viktor V. A., 2011). It contains 111,749 words and 144,980 synsets (Balkanova, Sukhonogov, & Yablonskij Sergey, 2004), which were neither reviewed nor evaluated.

There is also a manually created version of a Russian WordNet, called RussNet, but only a prototype version of the project is publicly available. Moreover, there are commercial projects by the enterprises UIS Rossiya and Novosoft (Suhonov & Yablonskij, 2004).

The Yet Another RussNet (YARN) is a Russian ontology crowdsourcing project with CC BY-SA licence. However, it is still under development and so far consists only of unrevised synsets (Braslavski, Ustalov, & Mukhin, 2014).

¹⁹ <http://creativecommons.org/licenses/by-nc-sa/3.0/deed.ru>

2.5.3. BabelNet

BabelNet is an extensive multilingual knowledge repository, which automatically aligns WordNet to the English Wikipedia by using a set of rules concerning the characteristics of the existing semantic relations. The multilingualism is achieved on the basis of Wikipedia cross-language links and the output of a machine translation system (Navigli & Ponzetto, 2012). The database is available under CC-license and is provided both as an online interface and an API. It contains a network of over 3 million synsets and 70 million semantic relations²⁰.

Figure 1 demonstrates a snippet of the output to the search term *trousers*. Like WordNet, BabelNet presents different meanings of *trousers* to the user. For better comparability, the same sense (or an equivalent to the WordNet synset) was chosen – *trouser, pants*.

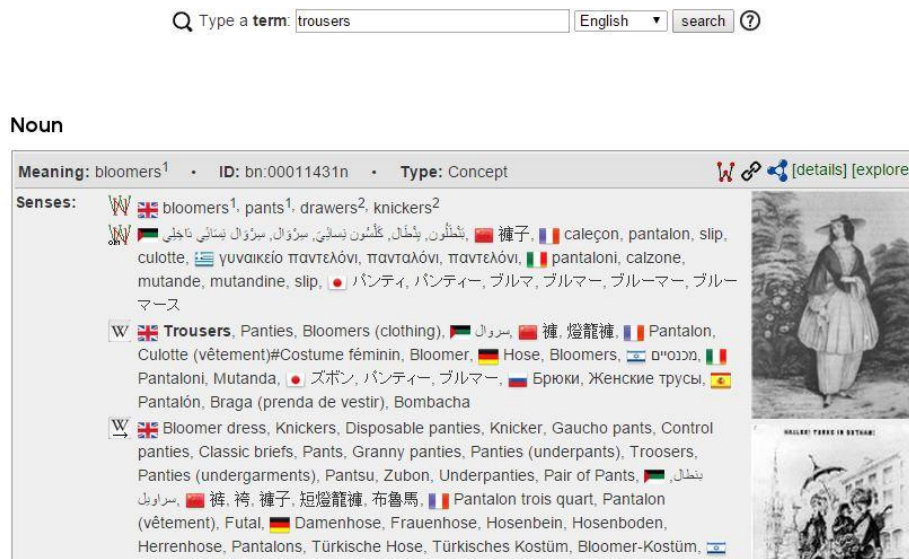


Figure 2 Image of BabelNet output to the search term *trousers*

Moreover, BabelNet provides information on classical semantic relations of the terms. An exemplary aggregated snippet of the contained information is presented in Table 2.6. However, due to space limitations, only the first-order relations of *trousers* are shown.

Category	Word	Category	Word
synonyms	pant	meronyms	lap
hypernyms			trouser cuff
hyponyms	strech pants		hip pocket
	jean		pant leg
	chino		slide fastener
	bellbottom trousers		trouser
	trews		seat
holonyms		DBpedia category	history of clothing
			history of fashion
			trousers and shorts

Table 2.6 Exemplary extract of the relations of *trousers* in BabelNet

²⁰ <http://babelnet.org/stats>

Categories	clothing				
Co-Hyponyms ²²	blouse	undershirt	top	vest	sleeves
	blouses	vest	white_shirt	coat	sleeves
	blouses	white_shirt	shoes	belt	white_shirts
	jacket	work_shirt	boots	skirt	skirt
	Seed	shirt	boots	sweater	blouse
	long_sleeves	shirt	t_shirt	dress	jackets
	shirts	shoes	tops	socks	skirts
	sweater	jacket	socks	white_shirts	
	tie	shirts	tops	tunic	
	tunic	waistcoat	blue_shirt	ties	
	jumper	waistcoat	blue_shirt	tunics	
	jumper	blazer	coats	tunics	
	suit	blazer	pants	work_shirt	
	suit	cap	pants	t_shirt	
	coat	cotton_shirt	polo_shirt	top	
	cotton_shirt	dresses	suits	hat	
	dress	jeans	suits	tie	
	dresses	jeans	sweaters	long_sleeves	
	hat	polo_shirt	ties	jackets	

Table 2.7 Relations of *trousers* in category *clothing* in NELL

2.6. Semantic Web Ontologies

The Semantic Web community aims at structuring the information contained in web pages to a standardized *web of data*, which would make the semantic information in the web reusable. The domain structured in an ontology could be seen as world knowledge. Examples of huge semantic web ontologies under GNU Public license²³ are DBpedia²⁴ (Lehmann et al., 2014), Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008), Yet Another Great Ontology (YAGO) (Suchanek, Kasneci, & Weikum, 2007) and Suggested Upper Merged Ontology (SUMO)²⁵. All listed ontologies, except Freebase, are available in several languages. DBpedia, Freebase and YAGO use Wikipedia as a source for the extraction of knowledge. These ontologies do not differentiate between proper and common nouns. However, they are better suited for ontologies of proper nouns. Thus, the examples shown for these databases will be that of a proper noun – *Paul McCartney*. Relations of proper nouns and the relations treated in the presented Semantic Web Ontologies are different from classical semantic relations, not necessarily combining nouns with other nouns. Moreover, the databases provide a mass of different relations that cannot be fully reflected here²⁶. Thus only some exemplary relations that are similar to classical semantic relations, such as *alias* or *alternative Names* being similar to synonym and *type* being similar to *hypernym*, are shown.

²² In NELL the relation is called *clothingtogowithclothing*, but it was named *co-hyponym* here, because it fulfils the requirements of this relation.

²³ <https://www.gnu.org/copyleft/gpl.html>

²⁴ “DBpedia data from version 3.4 on is licensed under the terms of the Creative Commons Attribution-ShareAlike 3.0 license and the GNU Free Documentation License. All DBpedia releases up to and including release 3.3 are licensed under the terms of the GNU Free Documentation License only.” (Lehmann et al., 2014)

²⁵ <http://www.adampease.org/OP/>

²⁶ Some of those relations are very specific and can in some cases be more correctly described as facts, as is done by YAGO. Examples of such relations are *wasBornOnDate*, *hasWikipediaURL*, and *hasFotoCollection*.

2.6.1. DBpedia

DBpedia entries are classified in consistent ontologies, the information of which is mostly extracted from Wikipedia infoboxes. It is updated once a year. Every DBpedia source has a label, two English abstracts and a link to the corresponding Wikipedia page. Moreover, it has optional links to images, external Webpages, Wikipedia and YAGO categories. It provides three different classification schemas for entities – Wikipedia Categories, YAGO classifications and WordNet synset links, which were created by manually relating knowledge contained in Wikipedia infoboxes, to WordNet synsets. The data from the infoboxes is extracted to three different datasets – *types*, *properties* and *special properties*, which specify concrete units for the property. A mechanically generated linkage of Freebase topics and DBpedia resources was implemented in 2012²⁷. Exemplary relations of *Paul McCartney* in DBpedia are presented below. Some of the *Related Entities* are linked to their own DBpedia entries or other web pages.

Relation	Related Entity
alias	Sir James Paul McCartney
type	Hard rock Artist Musical Artist Broadcast Artist Lyricist Celebrity Film writer Film director Award Winner Film producer Influence Node
genre	Pop_music Rock_music Electronica Classical_music
associatedBand	The_Beatles The_Quarrymen The_Fireman_(band) Wings_(band)

Table 2.8 Exemplary semantic relations of *Paul McCartney* in DBpedia

2.6.2. Freebase

Freebase, which is run by Google, is a graph-structured ontology whose information is extracted from various sources, the Wikipedia data being renewed every two weeks. Terms (*topics* in Freebase) are assigned to hypernyms or so-called *types*, which may have several properties. The types are parts of domains and thus path-like IDs are formed for terms contained in Freebase. Every ID is unique, but one term may have several hypernyms.

Relation	Related Entity
alias	Paul Bernard Webb Wings
type	Hard rock Artist Musical Artist Broadcast Artist Lyricist Celebrity

²⁷ <http://wiki.freebase.com/wiki/DBPedia>

Relation	Related Entity
type	Film writer Film director Award Winner Film producer Influence Node
genre	Rock music Pop music Classical music

Table 2.9 Exemplary semantic relations of *Paul McCartney* in Freebase

2.6.3. YAGO

YAGO automatically extracts terms and relations (or so-called *facts*) from Wikipedia and other sources. The manual evaluation of an extract of the relations gave an average of 95% accuracy. Additionally to the linkage to the DBpedia ontology, YAGO is also linked to Freebase and SUMO. The table below shows some of the facts about *Paul McCartney* that are stored in YAGO. The section *type* was shortened due to space reasons.

Relation	Related Entity
hasGivenName	Paul
hasFamilyName	McCartney
type	20th-century_English_singers British_drummers Transcendental_Meditation_practitioners British_rock_musicians British_people_convicted_of_drug_offences Rock_musicians English_rock_bass_guitarists
isMarriedTo	Linda_McCartney Heather_Mills Jane_Asher

Table 2.10 Exemplary semantic relations of *Paul McCartney* in YAGO

2.7. Concluding Remarks on Existing Resources

As shown in this chapter, many big and high-quality resources containing classical semantic relations already exist. However, apart from the issue of coverage of that knowledge, which will not be solved in the near future, most of these resources take little or no consideration of context. Especially from the pragmatic and semantic point of view, context is an important aspect in tasks that seek to understand or extract knowledge from natural language text. Some relations may only exist in the context of a given text, but are nonetheless crucial for its understanding. As this thesis seeks to research the impact of semantic relations in linguistic knowledge representation, the aspect of context may be important here. Thus, a novel approach to the extraction of semantic relations is chosen in this thesis. To analyse the impact of relations in context, they have to be compared with other knowledge resources such as presented in this section. Also, their influence can be measured through statistical analysis of the contained entities. Moreover, it must be proven that the semantic relations annotated in this thesis are common assumptions and not theoretical constructions. All of these issues are intended to be solved with the help of the methods and approaches presented in the next chapter.

3. Methods and Approaches

As all previously discussed knowledge bases do not take context into consideration, or only do it partly, a new dataset of semantic relations in context had to be built for the research pursued in this thesis. The details of corpus collection and the description of the annotation process are provided in Chapters 4 and 6. This chapter deals with the methods applied to the dataset on the one hand to measure the agreement on the annotations and on the other hand to analyse semantic relations from the different aspects such as language and genre.

After the collection and annotation of the dataset, the results were presented in a knowledge base, containing all nominals and their relations, together with a reference to the context in which they occurred. The results of the language subsets were compared with WordNet and its counterparts in the other languages. Afterwards differences between the dataset created in this thesis and the existing datasets, as well as peculiarities in the newly created dataset, are discussed.

To analyse the question of whether the use of semantic relations and certain types of relations is universal or rather dependent on language or genre, texts of different languages and genres were collected. All texts of one genre, either encyclopaedic, news or literary, are available parallel in all three languages, English, German and Russian, in order to be compared in the analysis. A more detailed description of the process of data collection and the resulting dataset can be found in Chapter 4.

The annotation of classical semantic relations between nominals was performed in a double annotation process according to guidelines (see A.2). The detailed description of the annotation process as well as the iterative development process of the guidelines containing a definition of semantic relations and nominals that was used in this thesis is presented in Chapter 6. To answer the question of whether a uniform structure for the annotation of this task can be found, the annotator agreement was calculated using Cohen's κ . These measures will also be used in order to show the improvement of the guidelines produced in this thesis by calculating it at different stages of the annotation process. The calculation of κ is presented in Cohen's κ in further detail.

After the completion of the annotated dataset, the results of the annotation will be analysed. To investigate the question of universality of semantic relations and the density of semantic relations in general as well as the distribution of individual relation types, the comparisons of those will be calculated using the nominal χ^2 -test (see Subchapter 3.2 for further detail).

In order to examine the question of whether the contextual approach finds other relations than previous approaches, the annotated dataset will be compared with WordNet for the English subset and its counterparts for the other languages. For the comparison with WordNet, the NLTK platform will be used. For the comparison with the German counterpart, GermaNet (Henrich & Hinrichs, 2011) the Java API will be used. The comparison will be taken between words that are in a relation in the created dataset and are also both present in the other dataset. Both the presence and the type of the relation between two words will be compared. Furthermore, in the transitive or partly transitive semantic relations hyperonymy and holonymy, the relations will be observed at all levels, meaning that not only the lowest hypernym, but all hypernyms will be observed.

To study whether terms having many semantic relations play an important role in their semantic context, such terms will be examined with reference to their source texts. In order to restrict the research of the influence factor to relations only, these will be compared with the most frequent nominals overall. Semantic relations will be categorized according to their function in text so as to investigate the role of the entities in these relations.

For the purpose of studying whether terms of different categories have different relation types, the subset of texts will be observed in further detail in order to analyse the use of relation types in distinct categories.

For all computational steps for which no applicable API or program was available, Perl, Java or Python programs have been implemented in order to verify the scientific hypotheses of this task. More specifically, the implementation of the comparisons of SemRelData with the three other databases have been performed in one of the three programming languages. The comparison with WordNet was implemented in Python, as it provides an API for WordNet through the NLTK platform. The comparison with GermaNet was implemented in Java, as an API for it was available in this programming language. In contrast to the other two languages, there was no applicable API for extracting the relations from RuTes. Thus, both the relation extraction as well as the comparison with SemRelData were implemented in Perl. The implementation of the relation extraction from SemRelData, the calculation of entities with the highest number of relations, the computation of the most frequent nominals, as well as the calculation of κ was performed with Perl.

The error classification for the comparisons, the macro-averaging of κ and the calculation of χ^2 is performed using Microsoft Excel.

3.1. Cohen's κ

According to Carletta Cohen's κ that was introduced in 1960 "measures pairwise agreement among a set of coders making category judgments, correcting for expected chance agreement." (1996). In annotation tasks Cohen's κ is used to measure inter-annotator agreement. It can be used for the calculation of agreement in *nominal annotation tasks*, e.g. the classical semantic relation labelling used in this thesis. In 1968 Cohen also proposed a calculation for weighted annotation, e.g. a measurement scale such as grades for pupils. The nominal κ coefficient that will also be applied in this thesis is calculated using the following formula²⁸:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad (1)$$

$P(A)$ is the proportion of annotator agreement, whereas $P(E)$ is the proportion of stochastic agreement. To calculate these measures, a contingency table, also known as confusion matrix, needs to be calculated. The table shows the counts of all agreements and disagreements of annotators for all classes. The following table exemplifies a contingency table and shall be used to demonstrate the construction of contingency tables in this thesis.

Each field in the calculation is depicted as h , with the identifiers

		Annotator 1		
		Label A	Label B	Sums 1
Annotator 2	Label A	$h_{AA} i anno1(i) = anno2(i) = "A" $	$h_{BA} i anno1(i) = "B", anno2(i) = "A" $	$h_{A=} = \sum h_{AA}, h_{BA}$
	Label B	$h_{AB} i anno1(i) = "A", anno2(i) = "B" $	$h_{BB} i anno1(i) = anno2(i) = "B" $	$h_{B=} = \sum h_{AB}, h_{BB}$
	Sums 2	$h_{A=} = \sum h_{AA}, h_{AB}$	$h_{B=} = \sum h_{BA}, h_{BB}$	$N = \sum h_{AA}, h_{BA}, h_{AB}, h_{BB}$

Table 3.1 Exemplary contingency table

To calculate $P(A)$, the proportion of all agreed labels, the following calculation is performed:

$$P(A) = \frac{\sum h_{AA}, h_{BB}}{N} \quad (2)$$

²⁸ Cohen's κ calculation Carletta, 1996, p. 4

This means that the diagonal fields, denoting the counts of all labels the annotators agreed on, are summarised.

To calculate $P(E)$, the proportion of random agreement, the following calculation is conducted:

$$P(E) = \frac{\sum_{i=1}^{\text{Number of labels}} \sum h_{i,i}}{N^2} \quad (3)$$

The κ calculation results in a value between -1 and 1. A κ value of 1 signifies full agreement; a κ value of 0 signifies chance agreement. According to Umesh et al., the annotator agreement cannot reach 1 due to observer bias (as cited by Bakeman & Quera, 2011). One of the first scales to appear in order to measure the significance of κ were Landis and Koch (1977). The scale is presented in the table below.

κ	Level of Agreement
<0	No agreement
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1	Almost Perfect

Table 3.2 Landis and Koch's scale of κ agreement

However, Bakeman et al. state that there is no universal guideline for the measurement of κ . Thus they implemented an approach that calculates the expected values of κ given various circumstances, such as number of labels and their prevalence (Bakeman & Quera, 2011).

The annotation task in this project does not only have several labels which are applied to relations between nominals, but there is also the possibility of not assigning any relation. Moreover, the calculation of the inter-annotator agreement has to deal with cases of annotations of multiple annotations of the same relation. Thus, in the calculation of the contingency table, another label, namely *No Annotation*, was added. Hence, it is possible to compare labelling and detect regularities in the disagreements.

Due to the difficulty of dealing with two layers of annotation, namely the annotation of compound nouns, and the semantic relations between them, the agreement of the annotation is expected to be lower than that of a one layer annotation task.

3.2. χ^2 -Test

As stated by McEnery and Wilson,

The Chi² test is probably the most commonly used significance test in corpus linguistics and also has the advantages that (1) it is more sensitive than, for example, the t-test; (2) it does not assume that the data are 'normally distributed' [...] and (3) [...] it is very easy to calculate. (2004, p. 84)

The χ^2 -test calculates the probability of differences in observed frequencies being chance by comparing the observed frequencies (of) with the expected frequencies (ef) with the following formula:

$$\chi^2 = \sum \frac{(of-ef)^2}{ef} \quad (4)$$

The bigger the difference between those values, the higher is the probability of the differences being not coincidental. To calculate these values, a contingency table, similar to the one presented in Cohen's κ , is built. As a first step, the observed frequencies are entered into the table. As a second step, the expected frequency for every cell is calculated with the following formula:

$$ef_{cell_{ij}} = \frac{\Sigma of\ of\ Class\ i * \Sigma of\ in\ Corpus\ j}{\Sigma of} \quad (5)$$

As a next step, χ^2 is calculated for every cell with the following formula:

$$\chi^2 = \frac{(of-ef)^2}{ef} \quad (6)$$

To further interpret the result of χ^2 , the degree of freedom (df) needs to be calculated as shown in the following formula.

$$df = (number\ of\ columns\ in\ table - 1) * (number\ of\ rows\ in\ frequency\ table - 1) \quad (7)$$

As a last step, the χ^2 value of the df can be looked up in a χ^2 distribution table in order to determine the p-value. The smaller the p-value, the higher is the probability of denying the hypothesis. If the p-value lies within the significance level α , the null hypothesis of independence can be rejected. The significance level is mostly set at 5%. Bortz and Weber (2005) categorized the interpretation of the p-value in the following way:

p-value	Significance level of result
≤ 5%	significant
≤ 1 %	very significant
≤ 0,1 %	highly significant

Table 3.3 Significance level and p-value correlation as presented by Bortz and Weber (2005)

4. Collection of Dataset

Although there is consensus on the fact that context is an important factor in the detection and analysis of semantic relations (Cruse, 1986; Murphy, 2003), the presentation of the different knowledge bases in Subchapter 2.5 and semantic web ontologies in 2.6 showed that context is not, or only slightly, considered in these projects. Thus, for an analysis of classical semantic relations in context and also for an analysis of the impact of context in such relations, a new dataset needs to be created.

The collection of the dataset proved to be arduous due to several criteria which the included texts needed to fulfil. These factors were representativeness, quality, comparability and copyright. Each of the three different genres, namely encyclopaedia, news, and literature had its particular issues that had to be dealt with in order to fulfil the criteria.

In the following, issues that were solved during the collection of the data set are discussed. The overall dataset consists of 20 files per genre, parallel available in the three languages. The overall set consists of nearly 60.000 tokens. The distribution of tokens and also nominals, which were the target of relation annotation, between the languages and genres can be viewed in the tables below.

	Encyclopaedic	Literary	News	Sum
Noun Compounds	2,301	6,519	6,028	14,848
Tokens	7,694	32,727	19,465	59,886

Table 4.1 Number of tokens and noun compound in the individual genres

	German	English	Russian	Sum
Noun Compounds	4,766	5510	4,572	14,848
Tokens	20,546	22559	16,781	59,886

Table 4.2 Number of tokens and noun compound in the individual languages

The sources of all texts in the dataset are presented in Table A.1, Table A.2 and Table A.3 in the appendix. The tables represent texts of different genres, the three parallel titles in the respective languages are shown in successive lines. In the following subchapters tables an aggregated view of the selection are given.

4.1. Representativeness

The representativeness of the data collection is ensured through a limitation of the corpus size and also by copyright, the manually selected text had to fulfil several criteria. To ensure the extensiveness of the subsets, only Wikipedia and Wikinews articles of at least three sentences were chosen. The threshold of three was chosen because less content would not be representative for the purpose of this thesis and more content appeared to be difficult to provide facing the parallelism issue. The titles of the news articles included in the dataset are presented below.

English title	German title	Russian title
Daisuke Enomoto will be the fourth space tourist at the ISS	Daisuke Enomoto fliegt als vierter Weltraum-Tourist zur ISS	Четвёртый космический турист
South Sudan gains independence	Südsudan ist unabhängig	Южный Судан стал независимым государством
Bush signs law to build fence at US-Mexico border	George Bush unterzeichnete Gesetz zum Bau eines Zauns an der Grenze USA-Mexiko	Буш подписал закон о строительстве забора
United States spies accused of illegally bugging the United Nations headquarters	Abhörmaßnahmen der NSA sorgen für Irritationen in Deutschland und Europa	Spiegel: АНБ США установило «жучки» в представительствах ЕС
Evo Morales wins presidential elections in Bolivia	Bolivien: Evo Morales siegt bei der Präsidentenwahl	Президентские выборы в Боливии
North Korea claims it has conducted a nuclear test	Fußballweltmeisterschaft 2018 in Russland, 2022 in Katar	Россия примет у себя Чемпионат мира по футболу 2018 года
Earthquake-damaged Fukushima nuclear power plant triggers evacuation	Atomalarm in Japan – Explosionen im Kernkraftwerk Fukushima I	Японский Чернобыль

English title	German title	Russian title
Kimi Räikkönen wins 2007 Australian Grand Prix	Kimi Räikkönen gewann im März 2007 den Großen Preis von Australien	Кими Райконен выиграл Гран-при Австралии 2007 года
100 icebergs heading for New Zealand	100 Eisberge auf dem Weg nach Neuseeland	100 айсбергов движутся к Новой Зеландии
European airspace closed by volcanic ash	Ausbruch des Vulkans Eyjafjallajökull behindert Luftverkehr	Из-за извержения исландского вулкана отменяются авиарейсы на севере Европы
NASA: Arctic Sea's icecap is melting	NASA: Rasanter Rückgang des „Ewigen Eises“ in der Arktis	Льды Арктики тают
America's atomic bombing commemoration held in Hiroshima	60. Jahrestag des Atombombenabwurfes über Hiroshima	Всемирный день борьбы за запрещение ядерного оружия
Polish President Lech Kaczyński dies as his plane crashes in Russia	Polnischer Präsident bei Flugzeugabsturz gestorben	Трагедия под Смоленском
Asiana Boeing 777 crashes upon landing at San Francisco International Airport	Bruchlandung eines südkoreanischen Verkehrsflugzeuges in San Francisco	Авиакатастрофа Boeing 777 в Сан-Франциско
Ratko Mladić arrested for war crimes	Serbien: Mutmaßlicher Kriegsverbrecher Ratko Mladić verhaftet	Арестован Ратко Младич
Spain defeat the Netherlands 1-0 in extra time to win 2010 FIFA World Cup	Fußball-WM: Tintenfisch Paul sagt Sieg Deutschlands im kleinen Finale gegen Uruguay voraus	Испания выиграла чемпионат мира по футболу
Rioting develops throughout England	Unruhen in Großbritannien: Lage eskaliert	Масштабные беспорядки вспыхнули ещё в нескольких городах Англии
FIFA announce Russia to host 2018 World Cup, Qatar to host 2022 World Cup	Fußballweltmeisterschaft 2018 in Russland, 2022 in Katar	Россия примет у себя Чемпионат мира по футболу 2018 года
Passenger airplane crashes in Siberia	Flugzeugunglück_in_Irkutsk	Крушение пассажирского самолёта в Иркутске
Mitt Romney wins 2012 Florida primary	Republikanische Vorwahlen: Florida geht an Mitt Romney	Митт Ромни одержал победу во Флориде

Table 4.3 Table of all news article titles that were used for this dataset

Moreover, in the case of encyclopaedic articles, one of the research questions in this work was whether nominals from the same category have similar classical semantic relations. Thus, three categories, namely *fruit*, *items of clothing* and *parts of the body*, are represented in the dataset. Those categories were chosen because they are often used as examples in the context of classical semantic relations. However, not the Wikipedia categories were used, as they are not equal for the three languages, but articles that fitted the criteria described in this chapter. The following table shows the article titles sorted by category.

Category	English title	German title	Russian title
Fruits	Durian	Durio zibethinus	Дуриан цибетиновый
	Orange	Orange	Апельсин
	Apple	Äpfel	Яблоня
	Melon	Zuckermelone	Дыня
	Clementine	Clementine	Клементин
	Prickly Pear	Opuntia ficus-indica	Опунция индийская
	Physalis	Blasenkirichen	Физалис
Clothing items	Catsuit	Catsuit	Кэтсьют
	Hat	Hut	Шляпа
	Trousers	Hosen	Брюки
	Boxer shorts	Boxershorts	Боксёры
	Waistcoat	Weste	Жилет
	Kilt	Kilt	Килт
Body parts	Finger	Finger	Палец
	Hair	Haar	Волосы
	Tongue	Zunge	Язык
	Eye	Auge	Глаз
	Thorax	Brust	Грудная клетка
	Vertebral column	Wirbelsäule	Позвоночник
	Ear	Ohr	Ухо

Table 4.4 Table of all encyclopaedic articles that were used for this dataset

Although two author lists (Gvishani-Kosygina, 1980; Smith, 2000) were searched, the representativeness of the literary subset is limited due to availability and copyright. The first table shows all authors and the titles of the works that were used in this dataset in the English translation.

Author	Work
G. Flaubert	Madame Bovary
L.N. Tolstoi	War and Piece
F. M. Dostoyevsky	The Idiot
E.T.A. Hoffmann	The Sandman
A. France	The Gods are Athirst
R. Kipling	The Jungle Book
H.G. Wells	War of the Worlds
A. P. Chehov	Kashtanka
A. M. Gorky	One Autumn Night

Table 4.5 Aggregated table of all literary works that were used for this dataset in the English translation

To prevent false conclusions due to translated texts varying from the original version, originals and translations of all three languages, as well as translations of all three texts from French were chosen. Although other factors concerning the author, like social background, age and gender are also important factors for the linguistic analysis of texts, those could not be considered in this thesis. In the following paragraph the reasons for this will be briefly discussed using the example of one factor.

In order to be representative of literary language, the texts need to be produced by both sexes. However, in all three genres, female authors are underrepresented. A study of the Wikimedia Foundation in 2010 concluded that only 13% of the Wikimedia articles were contributed by women (Cohen, 2011). Although the underrepresentation of female writers in literary text may also be due to there being fewer female writers, it should be mentioned that their number may also be lessened by another fact, addressed by Gleick. In the Wikipedia category *American novelists*, female writers are systematically removed to the sublist of *American Women Novelists* (Gleick, 2013). Although this notion concerns Wikipedia only, the issue it addresses may be applicable to other lists of writers: *female writers* may not be listed in author lists, because they are not considered *writers*. After the first searches for the literary subset, no texts of female authors were found. To circumvent the issue of female authorship underrepresentation, a list consisting of female authors only was used (Smith, 2000), which was barren of results, because the translators have only recently translated the texts and thus they still have copyright.

4.2. Quality

In this thesis, quality of texts was understood as the correct use of grammar and vocabulary as well as the reliability of the source. These factors posed different further factors upon the different genres.

The quality of the Wikipedia and Wikinews articles had to be secured by thorough reading, as the free production of the texts yields the problem of poor quality. According to a study of Giles, however, the text quality of Wikipedia rivals that of the Encyclopaedia Britannica (Giles, 2005).

In the literary text the reliability of the source was guaranteed by ensuring correct OCR and sufficient metadata, meaning information on author, translator and edition of the text. There were few OCR or edition mistakes, but those were not corrected in the source texts. The annotators were given the instruction to mark those terms, if they were of importance for the task.

4.3. Comparability

In order to make the multilingual texts comparable to each other, they had to be available in parallel in the three analysed languages. In the case of the Wikipedia articles this was done by choosing only those texts which were linked to each other in at least these languages. Though the texts are not the same in the different languages, they all have the same subject.

The choice of the Wikinews articles was more complex due to the fact that it does not contain as many articles as its encyclopaedic counterpart. The German version of Wikinews is sorted by continents that

the articles have been tagged to. All articles of all continents were manually searched and all those fulfilling the criteria described in this section were chosen. As the number of these articles was still smaller than those of the desired number in the final set, all articles tagged with the label *Umwelt* (engl.: environment) were also searched. Moreover, not only articles linked to each other were chosen, but also those which concern themselves with the same subject. Those were found by searching for articles which were of global importance. Some articles are not exactly on the same subject, e.g. the English and German articles *hiroshima_en.txt* and *hiroshima_de.txt* are about its 60th anniversary, the Russian article is about its 61st anniversary, but from the semantic point of view this circumstance should not make a decisive difference.

This should be enough to fulfil the criteria of parallelism, as semantic relations between nominals and not the content or overall language use is the focus of this thesis. Moreover, it was ensured that the texts were of comparable length in the three languages, so as to prevent analysis errors motivated by quantity and also to ensure an overall comparability of the subcorpora.

In the case of literary texts either the original or two of its translations or three translations of the same source text in a fourth language were chosen. When choosing the snippets for the corpora, it was taken care that parallel snippets were chosen, regardless of difference in length. As only few parallel literary texts could be found, several snippets of the found texts were used for the creation of the dataset.

The aim of this work is not only to compare semantic relations in various languages, but also in different genres. Thus, the subcorpora of different genres were made of approximately the same size. Another issue which was addressed in the course of corpus collection is that of diachrony, as it is one of the three main variations in linguistics next to location and genre. To be genuinely comparable, all texts needed to have been written in the same language period. This is difficult to accomplish, as Wikipedia and Wikinews articles, both easily available parallel in the three analysed languages under CC-BY licence, were written since 2001 or 2003 respectively, and parallel multilingual literary articles of this time are secured by copyright, which will be dealt with in the next section.

4.4. Copyright

The dataset and the results of its analysis described are available under CC-BY copyright²⁹, which makes the analysis replicable for anyone and the effort put in this dataset reusable. This is the main reason for choosing both Wikipedia and Wikinews articles, which are already distributed under this licence.

Full texts and sensible snippets are under copyright for 70 years after the author's death. The copyright law is not restricted to the author of the source texts, but also applies to the translator of these texts. Further on, both the author and the translator will be referred to as text creator. To be comparable to the other genres, the texts have to be as new as possible to ensure the use of current modern English. As demonstrated in this paragraph, the text creator has to have died no later than 1946, which is about 50 years from the first articles written in the other genres. This time period is already significant in terms of language variation. In order to keep the time variance as low as possible and also to prevent annotation difficulties due to the use of old language, only texts by text creators who died between 1900–1950 were chosen. Texts older than 1944 were taken from the Gutenberg Project and therefore are subject to the Gutenberg licence.

²⁹ License details: <http://creativecommons.org/licenses/by/4.0/>

Besides the generally tight restrictions, it has to be considered that during the time of the Iron Curtain, many alien writers were forbidden to be published and consequently also translated into Russian, although in the end of the Cold War those restrictions were loosened (Medushevskij, 2011). Moreover, during a large period of the Tsar era Russian was considered to be the language of the simple people and thus the educated literate social class, capable of reading in several European languages, did not need translations (Surina, 2009). Those factors confined the range of authors to only a few. In order to find works fulfilling these criteria, several lists of world-famous authors were examined in order to find their texts and those of their translators.

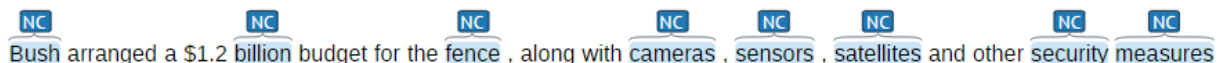
5. Preprocessing

The described texts have been first divided into the paragraphs as indicated in the edition they were taken from. This was not self-evident, as all texts from Gutenberg as well as other texts were formatted so as to fit the process of reading. Afterwards the texts were POS-tagged using the TreeTagger (Schmid, 1994; Schmid, 1995) to simplify the task of annotation. In the .tsv file that was uploaded to the annotation tool WebAnno (Yimam, Eckart de Castilho, Gurevych, & Biemann, 2014), only the nouns were annotated.

Not only simple nouns, but also noun compounds were of interest for the task at hand. Simplifications to find those were experimented with. However, this task was not conducted with German, as this language is known for its lexicalization of noun compounds. Two regular expressions were tested to automatically mark noun compounds in Russian and English. The spans were marked using the BIO-scheme³⁰.

Because adjective-noun and noun-noun compounds are the most productive in English, first the regular expression *adjective* noun+* was tested. This produced too many false positives, especially for Russian.

Since these results were poor, the first part of the annotations was pre-annotated by single nouns and annotators were asked to mark noun compounds manually. Using single nouns proved to be problematic especially in English, as annotators did not agree on the span of noun compounds frequently. The following figure shows the pre-annotated nouns in the first part of the annotation.

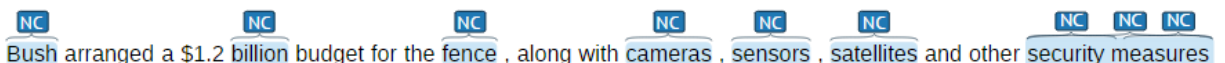


NC Bush arranged a \$1.2 billion budget for the fence, along with cameras, sensors, satellites and other security measures

Figure 4 Example of a pre-annotated sentence in the first part of the annotation

In this sentence, one annotator could have recognized *security measures* as a noun compound, whereas the other would have annotated *measures* only, which would have resulted in a conflict in the annotation of the entity as a hypernym of *fence*, *cameras*, *sensors* and *satellites*.

Thus, in the second part of the dataset, all spans matching the regular expression *noun+*, which is equivalent to a sequence of noun tags, were marked as noun compounds, which facilitates the search of noun compounds for the annotators³¹. Moreover, the improved guidelines determined that all noun sequences not containing a genitive are considered noun compounds. By this definition and pre-annotation, the annotators had a clearer guidance on noun compound annotation.



NC Bush arranged a \$1.2 billion budget for the fence, along with cameras, sensors, satellites and other security measures

Figure 5 Example of pre-annotated sentence in the second part of the annotation

Afterwards the files were uploaded as .tsv files internally separated by paragraphs, as demonstrated in the first three columns of Table 6.1.

³⁰ The BIO scheme suggests learning classifiers that identify the **B**eginning, the **I**nside and the **O**utside of the text segments (Ratinov & Roth, 2009).

³¹ Although the noun compound spans were marked, the nouns contained in these spans were also marked as noun compounds, as may be seen in the three NC tags above security measures.

6. Annotation

6.1. Introduction

Bird and Liberman define annotation in the following way:

‘Linguistic annotation’ covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions — audio, video and/or physiological recordings — or it may be textual. The added notations may include transcriptions of all sorts (from phonetic features to discourse structures), part-of-speech and sense tagging, syntactic analysis, ‘named entity’ identification, co-reference annotation, and so on. (2000, p. 23).

The annotation task in this work consists of two steps, which was explained separately in the following. The first step of annotation was identifying noun compounds, which are in a relation that was relevant to this thesis. The second step was marking and classifying these relations.

The annotation is performed with WebAnno (Yimam et al., 2014), a web-based annotation tool, which is described in more detail below. The annotation team consists of four annotators, two of which annotate for two languages. The German and the Russian annotators are bilingual or monolingual native speakers, the English annotators have a fluent knowledge of the language. Three of the annotators have linguistic background. The performance of the annotators was tested in a specialized task, which are presented in the appendix (see A.1).

Every document is annotated by two at least fluent speakers of the respective language. After this step, the two annotations are merged into a single final version through a curator by comparing, correcting and enhancing the two versions.

The annotations were made according to previously developed guidelines. As the guidelines were developed in a smaller setting and with one annotator only, they had to be iteratively improved. This was performed by both analysing mistakes in the annotations due to the lack of explicit rules in the guidelines and regular meetings of the annotation team, where problems and gaps in the guidelines were discussed.

The final version of the guidelines can be found in the appendix (see A.2). The development of inter-annotator agreement with the iterative improvement of the guidelines is shown through a time-dependent κ in Section 6.5.3.

6.2. WebAnno

WebAnno is a web-based tool for many different kinds of annotation. It supports the process of annotation starting from the upload of corpora and the creation of annotation levels suiting the need of the task. The process of annotation is offered in a graphical online user interface. The download in different file formats enables the further processing of the data. Moreover, the current version provides the possibility of automatically training annotated datasets (Yimam et al., 2014). The process of annotation as provided by WebAnno can be seen in Figure 6. The consecutive steps of the project are described in further detail in the next section.

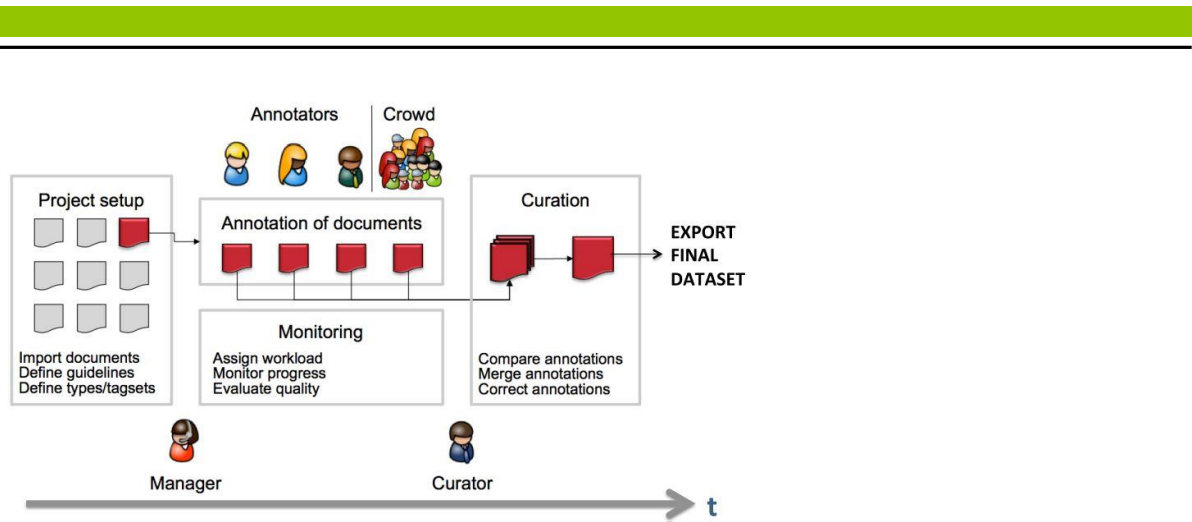


Figure 6 Prototypical workflow as implemented in WebAnno (Yimam et al., 2014)

6.3. Annotation Process

6.3.1. Project Upset Description

In this project, two custom annotation layers were created to fit the two steps of annotation that were previously described. Although the first layer, annotating noun compounds, was annotated in the pre-processing step, annotators were asked to correct wrongly or only partly marked noun compounds that were in a semantic relation to other noun compounds. Furthermore, the noun-compound layer contained the tags *NCpart* (denoting a part of a noun compound, which was cut off of its second part) and *Textmistake* (denoting spelling or tagging mistakes in the texts). The second layer annotated the classical semantic relations that are of main interest in this thesis. The layer contained the tags *Hypernym*, *Holonym*, *Synonym* and *Co-hyponym*. Furthermore, an uncertain relation could be tagged with ****UNCLEAR****.

6.3.2. Annotation of Documents

For each text in every language and every genre, two annotators were assigned one document that both of them annotated separately according to the guidelines.

The figure below shows a snippet of an exemplary annotation:

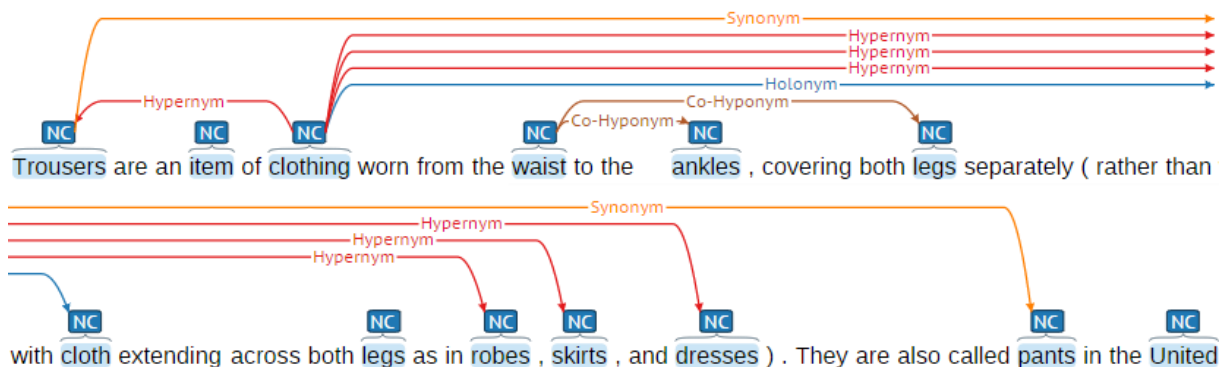


Figure 7 Annotation of hose_en.tsv for SemRelData, showing all four possible semantic relation tags

6.3.3. Curation and Export

After two annotators annotated one text, the text could be curated. In the Curation Page of the Tool, both annotation versions are shown. Congruent annotations are displayed in a third frame. However, the curator has the possibility to add or delete further annotations. At the stage of curation, many systematic inaccuracies in the guidelines could be detected.

The curated documents were used in the final dataset in the .tsv dataformat. The table below shows an exemplary snippet of a curated file corresponding to the annotated snippet shown in Figure 7 Annotation of hose_en.tsv for SemRelData, showing all four possible semantic relation tags.

ID	Token	NC-label	Relation	Related Token ID
1-1	Trousers	B-NC		
1-2	are	O		
1-3	an	O		
1-4	item	B-NC		
1-5	of	O		
1-6	clothing	B-NC		
1-7	worn	O		
1-8	from	O		
1-9	the	O		
1-10	waist	B-NC		
1-11	to	O		
1-12	the	O		
1-13	ankles	B-NC	Co-Hyponym	1-10
1-14	,	O		
1-15	covering	O		
1-16	both	O		
1-17	legs	B-NC	Co-Hyponym	1-10
1-18	separately	O		
1-19	(O		
1-20	rather	O		
1-21	than	O		
1-22	with	O		
1-23	cloth	B-NC		
1-24	extending	O		
1-25	across	O		
1-26	both	O		
1-27	legs	B-NC		
1-28	as	O		
1-29	in	O		
1-30	robes	B-NC	Hypernym	1-6
1-31	,	O		
1-32	skirts	B-NC	Hypernym Holonym	1-6 1-124
1-33	,	O		
1-34	and	O		
1-35	dresses	B-NC	Hypernym	1-6
1-36)	O		
1-37	.	O		
1-38	They	O		
1-39	are	O		
1-40	also	O		
1-41	called	O		
1-42	pants	B-NC	Synonym	1-1
1-43	in	O		
1-44	the	O		
1-45	United	B-NC		

Table 6.1 Exemplary snippet of a curated .tsv file

The first column gives the file-internal index of the token shown in the second column. The third column indicates whether the token is part of a nominal according to the BIO-scheme. B-NC marks the beginning of a noun compound, I-NC marks the continuation of a noun compound and O marks that the term is not a noun compound that is in a classical semantic relation. The 4th column shows whether there is a relationship to this noun compound. Several relations to the same token are separated by “|”. The next column gives the file-internal index of the token that the relation was annotated to.

6.4. Creation of Guidelines

Guidelines are manuals for annotation, which evolve in the process of corpus annotation. The process can be compared to the creation of a legal system, the ‘case law’ evolving through earlier cases and the setting of new leading cases when unfamiliar cases emerge (Leech, 2005).

In this thesis the creation of the guidelines was also performed iteratively. The first version of the guidelines was created prior to the formation of the annotation team. Several texts of all genres and languages were annotated with the goal to consistently annotate all classical semantic relations between nominals according to the definitions of the individual semantic relations. The definitions had the aim to be understandable without deep linguistic knowledge and both detailed enough to cover all relevant occurrences of the semantic relations and exclude all relations that were not of interest for the task. Parts of the definitions and subclassifications described in Chapter 2 were used in order to define the relations. Although no subclasses were defined in the guidelines, they were used so as to show which relations were and which relations were not included in the overall class. The full guidelines can be found in the appendix (see A.2). However, a brief definition and an English example per relation will be provided in SemRelData’s Iterative Relation Definition.

Relations with or among proper nouns were not annotated in this thesis. Like in WordNet, proper names are regarded as instances, not as types in a hierarchy, e.g. Paul McCartney is an instance of a singer, not a kind of singer (Miller & Fellbaum, 1991; Fellbaum, 1998; Fellbaum, 2013). Although semantic relations similar to hyperonymy exist between proper nouns, they should be regarded as a separate issue and cannot be addressed in this thesis.

6.4.1. Noun Compound Definition

Dealing with a multilingual text corpus, some measures had to be taken in order to provide comparability between relations of nominals. Nominals are differently realized in the three languages. Besides the general difference of the use of nominals in different languages, German provides a special type of nominals – a great number of lexicalized compound nouns. Comparing only relations between lexicalized nominals in the other languages to relations of nouns and compound nouns in German would be inefficient. Moreover, English and Russian do not lack the semantics of those nominals, so ignoring the fact that both languages also have noun compounds, which are, however, not realized in a lexicalized way as in German, would impede the study of semantic relations between nominals. The issue of noun compounding is not central in this thesis, thus it is discussed in this section and not in the State of the Art, where the main focus is semantic relations.

Grodal et al. address the linguistic debate of whether two orthographic units can be referred to as compounds (2014). In this thesis, this debate will not be discussed. The following definitions do not refer to this distinction and regard lexicalized and not lexicalized compounds as equals.

There are different kinds of nominal compounds concerning the POS being combined with at least one noun, e.g. noun-noun, noun-verb, noun-adjective and noun-preposition (Plag, 2003). However, the POS of the modifier is not a focus of this thesis. Plag states the issue of recognizing noun compounds in the following way:

Although compounding is the most productive type of word-formation process in English, it is perhaps also the most controversial one in terms of its linguistic analysis and I must forwarn readers seeking clear answers to their questions that compounding is a field of study where intricate problems abound, numerous issues remain unresolved, and convincing solutions are generally not so easy to find. (2003, p. 132).

Tokar defines the process of compounding as the word formation of a new compound lexeme through the combination of at least two input roots (2012). Plag defines a compound in the following way: “[...] a compound is a word that consists of two elements, the first of which is either a root, a word or a

phrase, the second of which is either a root or a word.” (2003, p. 135). Furthermore, Tokar states that “[...] compounding is an anisomorphic lexeme-building mechanism, i.e. a mechanism that produces output lexemes whose signifieds are not (or not entirely) representable in terms of their components’ signifieds.” (2012, p. 146).

Tokar divides the recently built compounds in *quasi-idiomatic* (also called *bahuvrihi*), *semi-idiomatic* and *fully-idiomatic*. He divides the *quasi-idiomatic* compounds in two main categories: *information fatigue-type* and *drum and bass-type*. The first type makes the additional idiomatic meaning of these compounds signifieds narrower than the sum of their components. The second’s signifieds describe some important characteristics of the compound. The *fully-idiomatic* compounds may come into existence via metaphorization of all components signifieds e.g. *carpet muncher* – “lesbian“, metonymization of all components signifieds e.g. *green accounting* – “a system in which economic measurements take into account the effects of production and consumption on the environment“ (Tokar, 2012, p. 149) and a combination of the two e.g. *grey nomad* – “a retired person who travels extensively” (Tokar, 2012, p. 149). There is a linguist view making a difference between *pseudo-compounds*, which describe derivations of compounds, e.g. *babysit* being a back-formation of *babysitter*, and *genuine compounds* (Tokar, 2012). However, this difference will be neglected in this thesis.

Despite the fact that many quasi-idiomatic compounds do indeed come to signify these ‘basic’ meanings, the same semantic outcome of compounding is to a very large extent unpredictable and unexplainable. That is, we cannot really explain why a particular quasi-idiomatic compound came to be associated with a particular idiomatic meaning. (Tokar, 2012, p. 152)

There is another kind of classification for compounds – *endocentric* and *exocentric*. *Endocentric compounds* have their semantic head inside the compound e.g. *laser printer* is a *kind of printer*. However, in the case of endocentric compounds, the meaning of the compound is not necessarily fully compositional, e.g. a *blackbird* is not just a *black bird* (Plag, 2003). *Exocentric compounds* do not have their semantic head inside the compound e.g. *redneck* is a *person*, not a *kind of neck* (Tokar, 2012; Plag, 2003). According to Tokar, the notion of endo- and exocentric compounds corresponds to the distinction between the two *quasi-idiomatic* compounds, the *information fatigue type* representing endocentric compounds, *drum and bass-type* representing exocentric compounds. Furthermore Tokar states that the distinction between endo- and exocentric compounds is broader than that of between the quasi-idiomatic compound types (2012).

Besides the distinction between *endocentric* and *exocentric compounds*, there are also linguists who describe an additional type – the *copulative compound* also known as *dvanda compound* e.g. *fighter-bomber*, which consists of two equally important signifieds from a semantic point of view (Tokar, 2012; Plag, 2003). Tokar argues that semantically seen the compound describes a lexical entity that is out of the scope of all signifieds and is thus exocentric (a *fighter-bomber* being an *aircraft*). Tokar concludes that from a semantic point of view, compounds can be divided into endocentric and exocentric compounds only (2012). Plag, on the other hand, proposes further subclassifications within copulative compounds – *appositional compounds*, where the components characterize the compound e.g. *scientist-explorer*, and *coordinative compounds*, where the relationship of the entities describing the nominal head is determined by this head e.g. *modifier-head structure*. There is a debate of whether there really exists a head-modifier structure, the arguments of which are mostly based on grammar, arguing that inflection affects only or not only the head according to one or the other side of the argument. Plag agrees with English compounds being mostly right-headed, meaning that the right side of the compound is the semantic head, which is modified by the left side. This is called the *modifier-head structure*. *Head* refers to “the most important unit in complex linguistic structures“ (Tokar, 2012, p. 135). Moreover, Plag defines an additional kind of compounds, so called *possessive compounds*, denoting a property of the semantic head, e.g. *loudmouth* – a *person having a loud mouth* (2003). Due to the prevalent right-headedness of English compounds, both Plag and Tokar also propose to discard the differentiation between endo- and exocentric nominal compounds from the formal perspective (Tokar, 2012; Plag, 2003). Furthermore, from the formal

perspective, there are endocentric and copulative compounds. The first are formally headed by their right-hand component (the plural of *fighter-bomber* being *fighter-bombers*), the second ones are two-headed (the past tense of *drag and drop* being *dragged and dropped*) (2012).

As seen in the definitions of Tokar (2012) and Plag (2003) the exocentric compounds have been given much attention in the field of word formation. In the task of recognizing noun compounds they are also more easily found, as they have an external meaning. However, endocentric compounds are not that easily found, also due to the fact that they do not always apparently have a non-compositional meaning, as stated by Plag (2003) earlier, e.g. the noun compound *gold necklace* is actually just a *golden necklace*. The compounding becomes apparent if we use phonological terms: *gold necklace* is spoken as one entity, whereas *golden necklace* is pronounced as two. Another phonological criteria is addressed by Levi (1978). The so-called *frontal stress* is a phenomenon that can distinguish compounds from other phrases, e.g. the noun compound *blackbird* from the attributive-adjective-plus-noun phrase *black bird*. Additionally, Levi states that although the presence of fronted stress denotes compounding, its absence does not, e.g. the compounds *apple pie* or *industrial revolution* have a normal stress (Levi, 1978). Thus, that noun compounding should be taken to the phonological level. Although the phenomena are not universal, the examples illustrate the notion of compound entities. A rule to recognize such exocentric compounds was formed in the guidelines: all noun-noun compounds that did not contain a modifier genitive were considered noun compounds in English. After the marking of non-lexicalized noun compounds in German caused much annotator disagreement, only lexicalized compound nouns were marked in the German subset. In this way, some noun compounds of interest, such as *Zusammengesetztes Nomen* (engl.: noun compound), were knowingly neglected. However, the annotation could become more systematized. In Russian no rule for the marking of noun compounds could be found, however, Russian did not seem to contain many noun compounds in the source texts.

6.4.2. SemRelData's Iterative Relation Definition

Before coming to the individual relation definitions, it should be stated that due to the characteristics of some relations not all of them had to be marked. The following rules and clarifying examples illustrate the characteristics used:

Rule	Example
If <i>A</i> is a synonym to <i>B</i> , then <i>B</i> is a synonym to <i>A</i> .	If <i>handbag</i> is a synonym to <i>purse</i> , then <i>purse</i> is a synonym to <i>handbag</i> .
If <i>A</i> is a hypernym of <i>B</i> , then <i>B</i> is a hyponym of <i>A</i> .	If <i>handbag</i> is a hypernym of <i>clutch</i> , then <i>clutch</i> is a hyponym of <i>handbag</i> .
If <i>A</i> is a holonym of <i>B</i> , then <i>B</i> is a meronym of <i>A</i> .	If <i>handbag</i> is a holonym of <i>handle</i> , then <i>handle</i> is a meronym of <i>handbag</i> .

Thus, only the first relation of every rule was actually annotated, the second was added during a post-processing step.

Moreover, due to the mentioned features, more annotations could be spared. As synonyms are defined as reflexive³², they share all relations. Therefore it is sufficient to annotate all relations to one

³² Contrary to Murphy's remark on synonyms not always being reflexive, in this thesis this feature is assumed (2003). Example (8) in Section 6.4.2 is a case of inference and does not strictly fulfil the requirements of synonymy.

synonym only. As hypernyms are transitive, all relations are inherited by the hyponyms, which means that it is sufficient to annotate all relations to the highest possible hypernym.

Synonymy was defined through similarity, clarifying that the relation must hold between two different words, similar to Katz, Harris, Cruse, Jackson, Chierchia and McConnel-Ginet and Hudson (as cited by Murphy, 2003).

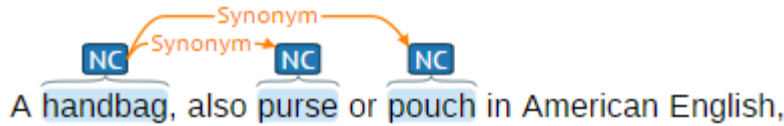


Figure 8 Example 1.1.1 of synonymy in guidelines

The other bidirectional relation that was defined was Co-hyponymy. Co-hyponyms were defined in the following way “Co-hyponyms are only annotated if there is no appropriate hypernym in the paragraph. Only co-hyponyms with a clear, common, and semantically linked hypernym are annotated.” (see 2.1).

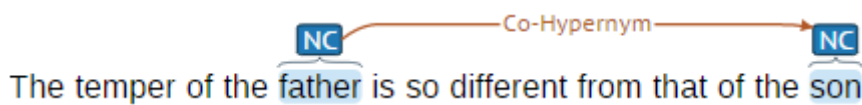


Figure 9 Example 1.2.1 of co-hyponymy with the in common hypernym family member in guidelines

The unidirectional relations were defined from the higher term of the relation to the lower. In hyperonymy this meant from the hypernym to the hyponym. Hyperonymy was defined as the *kind-of* relation. Although not explicitly outlined in the guidelines, *functional* and *taxonomic hyperonymy*, as well as *partonymy* as described by Wierzbicka (1984), were included. However, Wierzbicka’s restrictive definition of hyperonymy was not considered in the definition of hyperonymy in this thesis. As the aim of this thesis is the annotation and analysis of relations in context, restrictions to functional or taxonomic features are not substantial. Although some hypernyms may be better suited than others or may be of different grammatical category, they are still regarded as hypernyms, when the context implies it.

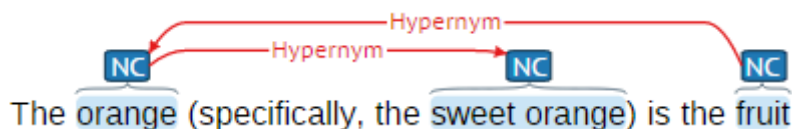


Figure 10 Example 2.1.1 of hyperonymy in guidelines

Holonymy was defined as a unidirectional relation from the holonym (whole) to the meronym (part). The subclasses described by Winston et al. were all included except the *place/area relation*³³, also using the relations described by them to show what holonymy does not include (1987). Although the subclasses described by Winston et al. were included, holonymy, like the other classical semantic relations in this thesis, was not further subclassified in the annotation. Without the differentiation of the sub-classes, Winston et al.’s transitivity feature was applied in this thesis. Thus, holonymy was defined as non-transitive, following Cruse’s (1986) definition.

³³ As the place/area relation is closely related to topological inclusion, it proved to be difficult to define in the guidelines. Moreover, all examples by Winston et al. for the place/area relation that did not consist of Named Entities could be applied to one of the other classes (Winston, Chaffin, & Herrmann, 1987).

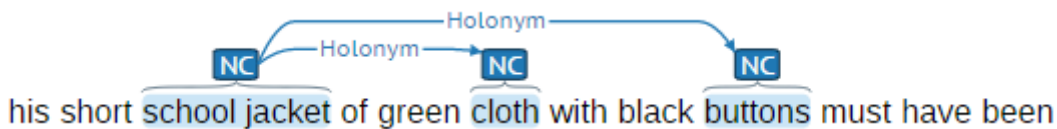


Figure 11 Example 2.3.4 of holonymy in guidelines

To efficiently mark the relations, relation features were used. Using the characteristic of transitivity in hyperonymy, all other relations had to be attached to the highest sensible hypernym, as all relations should by definition be transitively transferred to its hyponyms. As synonyms are reflexive, all other relations had to be between one of the terms in the synonym relation only, the relations being passed to the other term. Complications such as multiple occurrences of words and relations, as well as inflections of words that were in a semantic relation, also had to be considered. In the following, systematic annotator disagreements that occurred during the annotation process as well as the measures taken in order to improve the guidelines will be described.

Word ambiguity was a general problem that occurred. Annotators annotated relations, although in this context the words did not have a relation, e.g. *body* as a meronym to *person*, although in the context *body* was synonymous to *corpse*. To eliminate the annotation of such relations, the guidelines were enhanced with more examples illustrating ambiguity in order to raise the awareness of the annotators.

Holonymy appeared to be the most problematic relation. Annotators had particular problems with distinguishing between attribution and holonymy, as well as the already described similarity of the holonymic subclass of place/area and attachment (Winston et al., 1987). To solve this problem, the place/area relation was excluded from the guidelines. There was also much disagreement on holonymic relations between abstract nouns, e.g. *person* and *life* or *traffic* and *car*. The following rule was formed to prevent this: if only one of the related terms can be transformed into its plural form without changing its semantics, the terms cannot be in a semantic relationship. Furthermore, it was determined that if one of the positive rules applied to a relation, it is considered a valid relation.

Annotator disagreement also occurred in co-hyponymy due to the lack of a mutual hypernym level, e.g. in an excerpt of *The Sandman*, the characters *mother*, *father*, *sister* and *nurse* were introduced. The terms *mother*, *father* and *sister* with the mutual hypernym *family member* as well as all characters with the mutual hypernym *character* are valid annotations. At the first attempt, this problem was tentatively solved by proposing to combine all co-hyponyms with the lowest possible hypernyms. In this way more specific relations could be retrieved. However, this resulted in even more disagreement, different annotators relating different co-hyponyms with hypernyms such as female family members. Thus, in the final attempt to solve the problem, the guidelines dictated the annotation of all co-hyponyms with the highest possible hypernym. Although this decision potentially caused the loss of more specific information of the contained terms, in this way the annotation could be standardized, which is the aim of well-defined guidelines.

6.5. Inter-annotator Agreement

In this section the inter-annotator agreement as calculated using *Cohen's k* on a nominal scale will be presented. The κ is calculated by using a *contingency table*. The κ presented here was calculated using all classes presented in the contingency table with the exception of *****UNCLEAR*****, as it cannot be expected that annotators agree on a label that was created to indicate that the annotator sees a relation but cannot decide on the label³⁴.

³⁴ A calculation including the *****UNCLEAR***** label was conducted. The κ value was < 0.03 when compared with the presented values without the label, which shows that the exclusion of this class was not substantial.

The contingency table shows the agreement of the double annotation by presenting one annotator on the vertical and the other annotator on the horizontal axis, showing the different classes. Moreover, the κ between all individual annotators and the curator is shown. The matrices were built regardless of the annotated language, because with the exception of one file³⁵ all languages were annotated by two annotators, who did not overlap in another language. Strong deviations of individual κ s from the norm are reported.

In order to analyse distinct factors influencing the annotator agreement, several factors are presented separately. Influential factors are presented in the following order: annotators, paragraph size, time and genre.

6.5.1. Annotator as a Factor Influencing Annotator Agreement

All contingency tables are demonstrated in the appendix (see Table A.4, Table A.5, Table A.6, Table A.7, Table A.8, Table A.9, Table A.10 and Table A.11). The κ s of the annotators and the curator calculated through macro averaging³⁶ from these matrices are presented below.

Annotator/Curator	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Curator
Annotator 1		0.17	-	0.21	0.45
Annotator 2	0.17		0.24	0.32	0.51
Annotator 3	-	0.24		-	0.55
Annotator 4	0.21	0.32	-		0.56
Curator	0.45	0.51	0.55	0.56	

Table 6.2 κ agreement of all annotators and the curator

Table 6.2 shows that inter-annotator agreement is between 0.17 and 0.32 and has an average of 0.24. The agreement with the curator ranges between 0.45 and 0.56 with an average of 0.51. The κ s of the individual comparisons of an annotator and a curator in one language range from 0.41 – 0.59 (see Table A.4, Table A.5, Table A.6 and Table A.7). This shows that the language comparison between two coders varies only in the second decimal place. The annotator agreement presented in Table 6.2 and the detailed curator agreement (see Table A.8, Table A.9, Table A.10 and Table A.11) also represent the language dependency factor of inter-annotator agreement. According to Landis and Koch (1977)'s scale (see Table 3.2), the average agreement between the annotators is *fair* and the average agreement between the annotators and the curator is *moderate*.

However, to analyse the annotator agreement, not only the κ , but also the classes on which the annotators disagreed on are of interest. It was calculated which class was most often confused with which other class. The result was that in all pairwise comparisons of the annotators, as well as in the comparisons of annotator and curator most disagreement was caused by one annotator annotating a relation that the other annotator did not annotate at all. The second highest number in the contingency table denoted agreement, which means that if two annotators agreed on two nominals having a relation, they most frequently also agreed on its label. All classes except *Hypernym* and ****UNCLEAR**** were most often confused with hyperonymy, if the other two, previously discussed cases were ignored. Co-hyponymy was confused with holonymy in two out of total eight comparisons, whereas in the remaining six comparisons it was most often confused with hyperonymy. Hyperonymy was most often confused with holonymy in all eight comparisons, whereas ****UNCLEAR**** was confused most often with hyperonymy in two cases and in eight with holonymy.

³⁵ The sanfrancisco_de.tsv file was annotated by three annotators.

³⁶ The κ was additionally calculated by micro averaging. The variance between the two κ values calculated through micro and macro averaging was < 0.07.

6.5.2. Paragraph Size as a Factor Influencing Annotator Agreement

As the annotation task was to mark semantic relations between nominals in a paragraph, the size of the paragraph is a potential influential factor to the annotator agreement. The reasons for the potential difficulty to detect relations in longer paragraphs might be on the one hand memory, meaning that the annotator has to remember nominals for a longer text distance, on the other hand due to the tool, which in the case of very long paragraphs does not display the full text which is to be annotated.

As depicted in Table 6.2, the difference between the κ of the annotators and the curator is significant, thus an average κ for a file considering all κ s might falsify the result. Figure 12 shows the average paragraph size with the corresponding κ , the κ s of the annotators depicted in green, the κ s of the curator depicted in grey. The full table showing the precise value can be found in the appendix (see Table A.12). As can be depicted from the graph, κ is not dependent on the paragraph size.

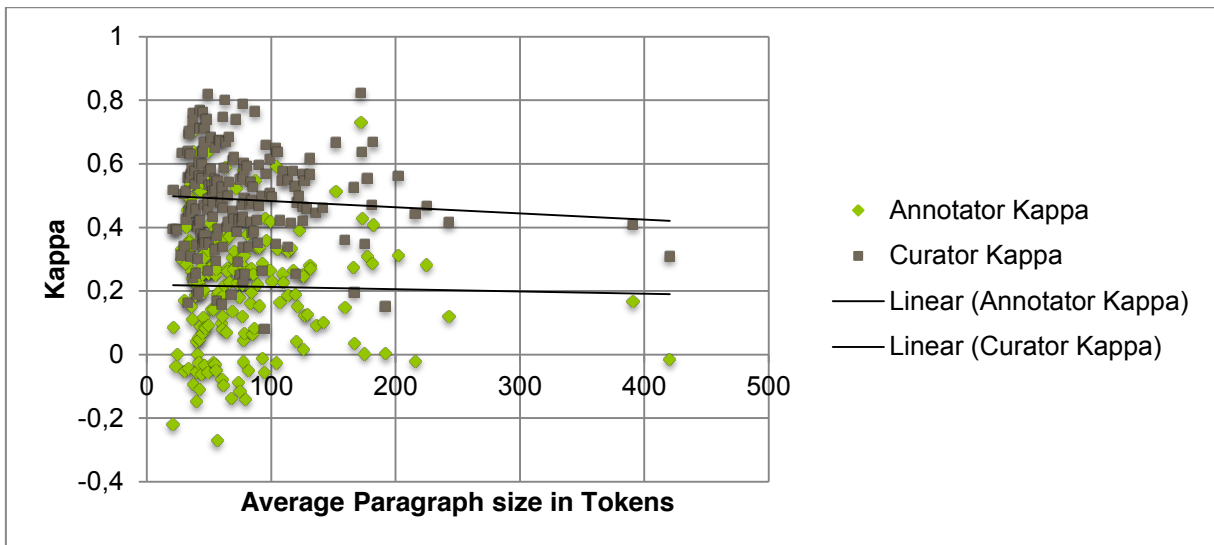


Figure 12 Paragraph size/ κ correlation

6.5.3. Time as a Factor Influencing Annotator Agreement

As already stated above, the guidelines were improved iteratively. This was done with the aim to improve the inter-annotator agreement, thus time should be analysed as an influential factor on annotator agreement. Table 6.3 shows the macro-average κ agreement of two annotators and with the curator in the corresponding time span. The spans represent different points of time to which adaptations to the guidelines were made.

Time span	Av. κ of Annotators	Av. κ with Curator
1	0.20	0.43
2	0.21	0.52
3	0.25	0.57
4	0.27	0.43

Table 6.3 Annotator agreement sorted by time spans

The table shows a steady improvement of κ between the annotators. The κ of the curator, however, drops in the last time span.

6.5.4. Genre as a Factor Influencing Annotator Agreement

Genre is one of the variables in this project, thus its influence on inter-annotator agreement is analysed. Table 6.4 shows the macro-average κ between annotators and between an annotator and a curator according to genre.

	Av. κ of Annotators	Av. κ with Curator
Encyclopaedic	0.20	0.50
Literary	0.23	0.49
News	0.23	0.53

Table 6.4 Annotator agreement sorted by genre

The table shows that the variance between the κ s in the different genres is < 0.05 for the κ between the annotators and between annotator and curator.

6.5.5. Conclusions of Influential Factors on Annotator Agreement

The analysis of the variables in this annotation task, being potential influential factors of the annotator agreement reveals that only the time factor has a measurable influence on the annotator agreement.

Table 6.3 shows a clear improvement of inter-annotator agreement, which indicates the improvement of the guidelines leading to a greater consensus of the annotators. However, the agreement with the curator drops in the last time-span. This could be due to several factors. First, the annotations in the last time-span were made under time pressure, which bears the risk of careless mistakes. Moreover, the annotations were conducted after Christmas holidays, which is also a factor which may have negatively influenced the annotator performance.

The fact that no other factor influenced the inter-annotator agreement may indicate that the concept and definition of classical semantic relations is commonly understood by the annotators and these relations are found independent of language, genre and paragraph size.

6.5.6. Conclusions on the Difference between Annotator and Curator Agreement

The comparison of all previously shown average κ s of annotators and annotators and curator shows that the average agreement with the curator is twice or more as high than the average agreement between the annotators. This leads to the assumption that the annotations of the individual annotators contain correct annotations that were found by one annotator only. This assumption is also supported by the fact that most disagreement between annotators is due to one annotator having annotated a relation that the other annotator has not annotated. As was also discussed, if annotators agreed on the related entities, they mostly also agreed on the relation class. Considering the fact that the agreement between annotators and curator is not much higher than twice the inter-annotator agreement, it may be assumed that by the double annotation and subsequent curation most of the classical semantic relations that are contained in the texts were found.

Concluding from the consistency of κ and the complementing double annotations it can be said that a uniform and automatable structure for the annotation of classical semantic relations can be found. This chapter shows that a consistently annotated dataset was created. The next chapter shows the post-processing steps that build the basis for the analysis of classical semantic relations which is conducted in Chapter 9.

7. Postprocessing and Statistics

After the annotation and curation of the dataset, the annotated relations needed to be further processed in order to fit the methods of the analysis. For the analysis, the relations needed to be represented individually and not in a tab-separated format as in the direct output of the annotation tool. Moreover, the definitions of the semantic relations provide more relations than that directly contained in the annotated dataset. The postprocessing is described in the first subchapter of this chapter.

The second subchapter deals with the statistics of the dataset and its annotations as well as the entities that were added in the postprocessing.

7.1. Postprocessing of SemRelData

In this chapter the postprocessing of the curated data is described. As described in Subchapter 2.2, the relations possess different features such as transitivity or reflexivity, which were used in this thesis.

In general, it can be said that at most only half of the existing annotations had to be actually annotated, as there is no need to mark both hypernym and hyponym relations, holonym and meronym relations and synonym and co-hyponym relations towards each other.

The rules, described in more detail in the Section 6.4 lead to the process of postprocessing, which is described in the following figure.

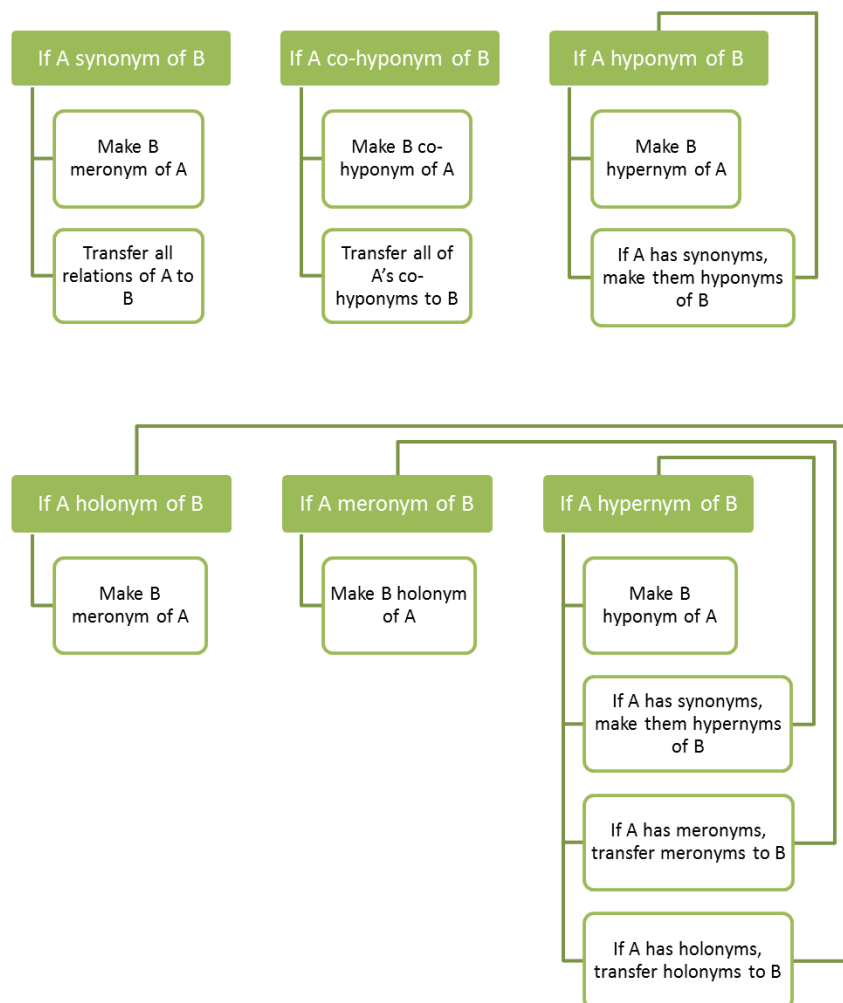


Figure 13 Postprocessing relation extraction rules

The following hypothetical example demonstrates the functionality of the postprocessing. The introductory example of *handbag* is used for this purpose. In the annotated set, *bag* would be annotated as a hypernym of *handbag*. In the postprocessing, *handbag* would be annotated as hyponym of *bag*. Furthermore, if *bag* is annotated as a holonym of *handle*, the holonymic relation would not only be passed on to *handbag*, but *handle* will also be marked as a meronym of both *bag* and *handbag*³⁷. Moreover, if *handbag* has the synonym *purse*, all relations of *handbag* are transferred to *purse*³⁸. However, relations of *handbag*, with the exception of the synonymic relation with *purse*, will not be transferred to *bag*.

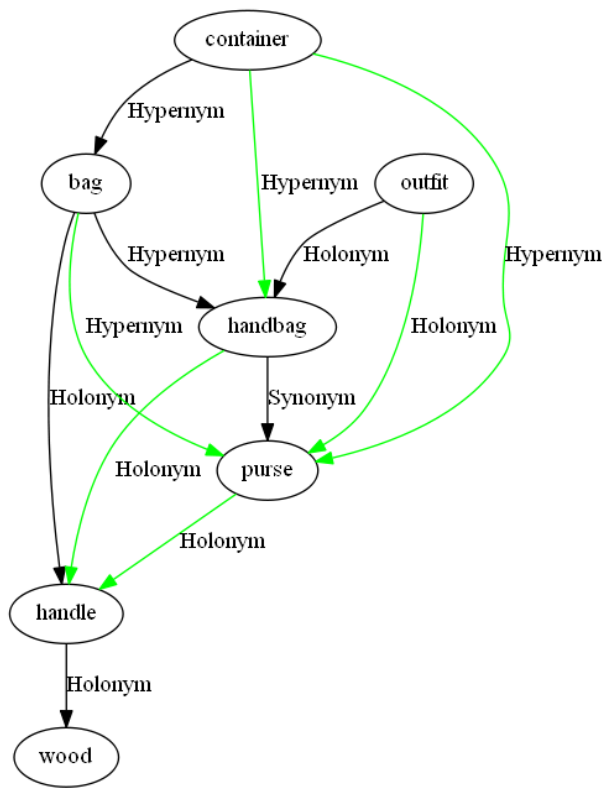


Figure 14 Graphical visualization of relations of *handbag*. Annotated relations are marked in black; annotations added in the post-processing are marked in green. Reverse annotations are not displayed.

Figure 14 shows this example. The five initially annotated relations are marked in black. The reverse annotations, such as e.g. *handbag* being the hyponym of *bag* are not displayed for reasons of clarity. The relations that would be added in the postprocessing are marked in green. All in all, 17 (including the reverse relations) additional relations would be added to this hypothetical example.

³⁷ However, if *handle* is a holonym of *wood*, this relation will not be transferred to *bag* or *handbag*, because holonymy was not defined as a transitive relation and is thus handled differently in the postprocessing.

³⁸ With the exception of the synonymic relation, which is transferred in a hyperonymic and a hyponymic relation, in the example *purse* would be a hyponym of *bag* and *bag* a hypernym of *purse*.

8. SemRelData Statistics and Characteristics

In this subchapter, the number of tokens, noun compounds, direct and transitive relations are presented. The following table shows the number of noun compounds, tokens and directly annotated relations in the subsets and the overall dataset.

Set	№ Tokens	№ NC	№ Ann. Rel.	№ Trans. Rel.
German	20,546	4,766	1,217	3,514
English	22,559	5,510	1,231	3,440
Russian	16,781	4,572	954	2,486
Encyclopaedic	7,694	2,301	982	3,170
Literary	32,727	6,519	1,587	4,328
News	19,465	6,028	833	1,942
Whole Set	59,886	14,848	3,402	9,440

Table 8.1 Statistics of SemRelData. The 1st column presents the number of noun compounds, the 2nd column presents the number of tokens, the 3rd column presents the number of annotated relations and the 4th column presents the number of transitive relations.

The resulting dataset contains approximately 60,000 tokens, 15,000 noun compounds, 3,400 annotated relations and 9,400 transitive relations.

The dataset consist of three parts and is available under CC-BY license. The first part consists of the original files in .txt format, the second part will consist of the curated files with classical semantic relation annotation in .tsv format and the third part will consist of the ontologies, including the transitive relations, of all files.

9. Results of the Analysis

In this chapter the results of the different comparisons for the analysis of SemRelData are demonstrated. In Section 9.1, the comparisons between SemRelData and WordNet, GermaNet or RusTes are shown by comparing relations between words contained in both compared datasets. In Section 9.2 a comparison with a pattern-based approach is presented. In the following two sections, the comparisons between the different languages and genres using nominal χ^2 -test are presented. Details of the calculation of χ^2 are presented in Chapter 3. For the calculation of χ^2 the numbers of half of the semantic relations, including the transitive relations, is considered³⁹. Both density of semantic relations in general and the distribution of semantic relations in different subsets are subject to this thesis, thus each factor will be analysed separately. Section 9.5 deals with the comparison of relation types in the different categories within the encyclopaedic subset. Section 9.6 deals with peculiarities, such as comparing terms with many relations to other terms with regard to the contextual role.

9.1. Comparison with Knowledge Bases

This chapter presents an exemplary entry of SemRelData in order to show general differences between the resource created in this thesis and other knowledge bases as presented in Section 2.5. In the subchapters, its subsets are compared with WordNet and its counterparts.

To compare SemRelData with knowledge bases, the relations that were extracted from the file *hose_en.txt* will be presented below. All direct relations that were annotated in the file are presented in Figure 16 in the appendix. Figure 15 shows all direct relations to the string *trousers* or *Trousers* that could be extracted from the file. The different relations are also shown in different colours – synonymous relations are marked in light blue, hyperonymic in blue and holonymic in yellow. Table 9.1 below presents all relations of *trousers*, both direct and transitive.

	Word
Synonym	pants
	long trousers
Holonym	school uniform
Meronym	legs
	fastening
Hypernym	clothing
	garments
Hyponym	shorts
	jeans
	leggings

Table 9.1 Relations of *hose_en.tsv* in SemRelData

As seen in the table, not all relations that may be transitively derived (shown in Figure 15) are actually created, e.g. *short trousers* should also be a hyponym of *trousers*, but in SemRelData it is not. This phenomenon is not restricted to this exemplary case and may have two reasons: either *short trousers* and *trousers* never occurred in the same paragraph⁴⁰, or *shorts*, being a hypernym of *trousers*,

³⁹ This was done because the numbers of all the types are symmetric to one other type in the type set, meaning the number of hypernyms is the same as the number of hyponyms, the number of holonyms is the same as the number of meronyms. As both co-hyponyms and synonyms are reflexive, their counts were halved. Considering all relations would amplify the proportions of all relations and could lead to a falsification of the results, only half of the relations was used for the analysis.

⁴⁰ As already previously described, the rules depicted in Figure 13 are restricted to paragraphs.

occurred twice in a paragraph, one occurrence being linked to *trousers*, the other occurrence being linked to *short trousers*⁴¹.

Furthermore, the table shows a double relation between shorts and trousers, which means that this relation occurred twice. Relations occurring multiple times were not treated differently in this project. In some pattern-based approaches such relations are handled as more secure.

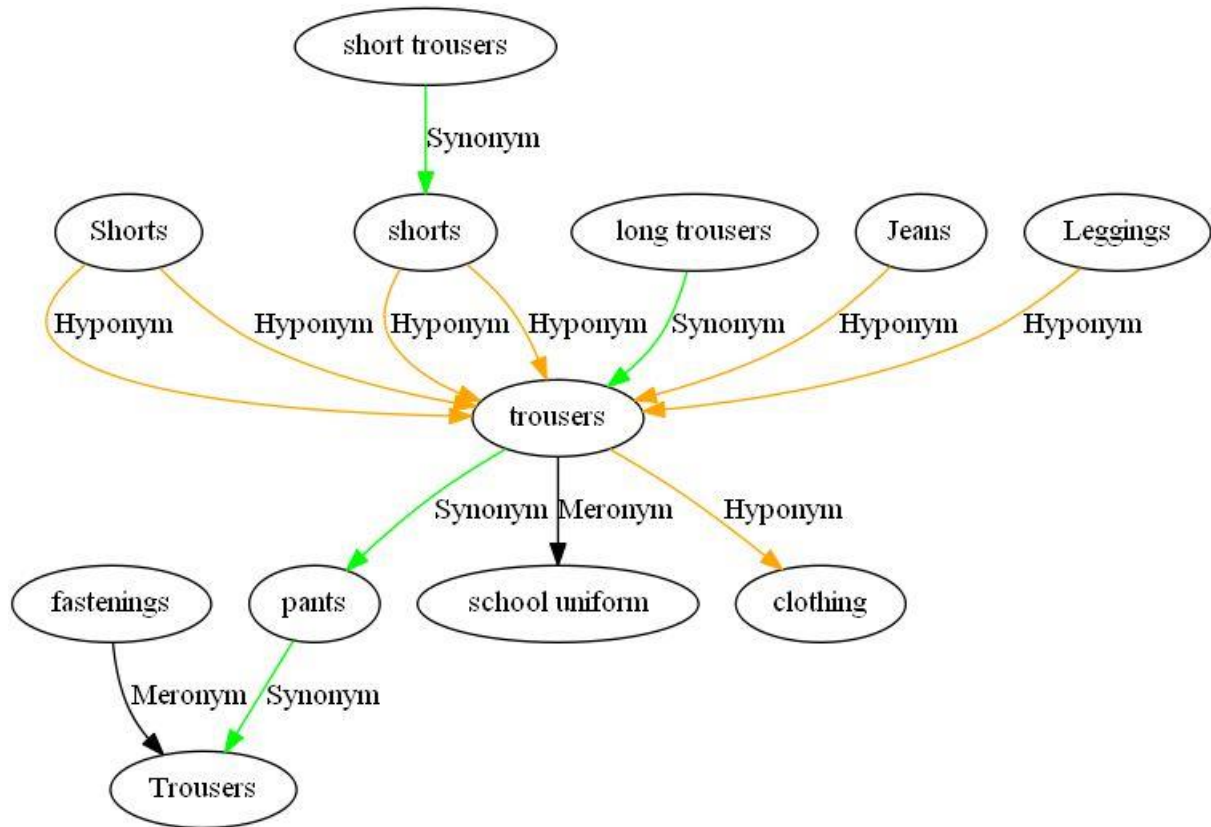


Figure 15 Graphical visualization of relations of *hose_en.tsv* in SemRelData. Synonyms are marked in green, hyponyms are marked in orange, and meronyms are marked in black.

Moreover, in this chapter the comparison with the relations contained in WordNet and its counterparts in the other two languages, GermaNet and RuTes are presented. The description of the creation and characteristics of the overall dataset, as presented in Chapter 7, applies to all subsets. All direct and transitive relations were used for the comparisons. The guidelines prescribed to ignore inflection in the annotation of relations, thus SemRelData contains inflected forms of nouns. Hence, only relations which contained words whose lemmas are both present in the other knowledge base were compared. Furthermore, all post-processing steps that were applied to SemRelData (see Chapter 7.1) were also recreated for the other knowledge bases.

As resources cannot be expected to have similar relations at the same depth, e.g. *shorts* being either considered a direct hyponym of *clothing* or a transitive hyponym through being a hyponym of *trousers* and *trousers* being a hyponym of *clothing*, depth of transitive relations was not considered in this comparison. Hence, co-hyponyms were not considered in these comparisons, as any pair of words present in any of the compared databases would be considered a co-hyponym, because in any case, they would have the top-most hypernym *Entity* in common.

⁴¹ The guidelines determined that in case of the double occurrence of one word, whereas the related word occurs only once, there should be only one relation annotation, according to specific rules (see A.2).

To analyse the relations which are not present in the other databases, 50 randomly chosen disagreements between a language subset of SemRelData and the other database were manually classified in six error types:

- 1) Relation too specific (RS): Though the relation is generally true, it is too specific e.g. *chordates* being a hyponym of *species*.
- 2) Ambiguous (A): Although the terms of the relation are present in both datasets, the meaning presented in SemRelData is missing e.g. *physiognomy* is used as a synonym of *look* in SemRelData, whereas WordNet only contains the meaning of *face*.
- 3) Contextual (C): The relation presented by SemRelData is generally not true, but exists in the given context e.g. *control* being a hypernym of *law*.
- 4) Subset too specific (SS): The subset of the terms in the relation is too specific e.g. *man* is not a hypernym of *father*, because *father* is defined as a *parent*, not as a *male human being* in WordNet.
- 5) Lemmatization error (LE): The lemmatization produced a wrong lemma, which was confused with another word e.g. *boxers*, meaning the type of *underwear*, was lemmatized as *boxer*, meaning the *athlete*.
- 6) Unclear or other (U): It is unclear why this relation is not included in the other knowledge base e.g. *icecap* is not a holonym of *ice* in WordNet) or the reason is not within the scope of the other classes (e.g. *man* is not a holonym of *hand* in WordNet *man* is a holonym of *arm* and *arm* is a holonym of *hand*, but holonymy is not transitive by the definition in SemRelData).

9.1.1. Comparison with WordNet

To compare the relations of the English subset with WordNet, the NLTK (Bird, 2006) implementation of `pywordnet`⁴² was used. For the lemmatization, NLTK using the WordNet lemmatizer was applied.

Of the 3,390 relations in the English subset, 562 were not considered, because of the above described issue with comparison of co-hyponyms. 1,902 (67.26%) could be compared with WordNet relations, as the lemmas of the terms linked by the classical semantic relations were found in WordNet. Of those 1,902 relations, 1,026 (53.94%) were present in both datasets. The following table shows the counts of the error type classification of 50 randomly chosen disagreements (the classification of all fifty relations is presented in Table A.13).

Error Type	Count
RS	4
A	2
C	9
SS	9
LE	1
U	25

Table 9.2 Disagreement analysis of WordNet and SemRelData in 50 random relations

9.1.2. Comparison with GermaNet

To compare the relations of the German subset to GermaNet, the GermaNet Java API and the GermaNet 8.0 version were used. For lemmatization, the JoBim Text API⁴³ lemmatizer using the Pretree Tool (Biemann, Quasthoff, Heyer, & Holz, 2008) was applied.

⁴² <http://osteele.com/projects/pywordnet/>

Of the 3,512 relations in the German subset, 670 were not considered, because of the above described issue of comparison of co-hyponyms. 1,284 (50.92%) could be compared with GermaNet relations, as the lemmas of the terms linked by the classical semantic relations were found in GermaNet. Of those 1,284 relations, 701 (54.59%) were present in both datasets. The following table shows the counts of the error type classification of 50 randomly chosen disagreements (the classification of all fifty relations may be found in Table A.14).

Error Type	Count
RS	8
A	7
C	6
SS	5
LE	0
U	23

Table 9.3 Disagreement analysis of GermaNet and SemRelData in 50 random relations

9.1.3. Comparison with RuTes

To compare the relations of the Russian subset with RuTes, there was no API available, so the same rules as described in Chapter 7 were applied in order to create the transitive relations⁴⁴. For the lemmatization process, pymystem3⁴⁵, which is a Python wrapper for Yandex Mystem⁴⁶, was used. Of 2,416 relations that were found in the Russian subset, 1824 were used for the comparison. 850 (46.60%) could be found in both subsets. The properties of the Russian subset limited the comparison to Hyper-, Hypo-, Holo-, and Meronyms. Thus, of those 850 relations, 596 relations could be compared due to their relation type. 288 (49.83%) relations were present in both sets. The following table shows the classification counts of the reasons for the error type classification of 50 randomly chosen disagreements (the classification of all fifty relations may be found in Table A.15).

Error Type	Count
RS	12
A	3
C	10
SS	4
LE	0
U	21

Table 9.4 Disagreement analysis of RuTes and SemRelData in 50 random relations

9.1.4. Conclusion of the Comparison with other Knowledge Bases

Summarizing the comparisons with the three knowledge bases it can be said that the distribution of the relations contained in SemRelData and a knowledge base were similar. This is also true for the results of the disagreement analysis of all three comparisons. This implies that the coverage of SemRelData is even throughout the languages.

⁴³ <http://maggie.lt.informatik.tu-darmstadt.de/jobimtext/api/>

⁴⁴ Although the transitivity is described by Loukashevich (2011), they are not explicitly instantiated due to reasons of space and data management.

⁴⁵ <https://pypi.python.org/pypi/pymystem3/0.1.1>

⁴⁶ <https://tech.yandex.ru/mystem/>

The comparisons show that about a half of the relations that were compared are present in SemRelData and an existing database. The rate of mutual relations with GermaNet and WordNet was higher than that of RuTes. One may argue that these resources are closer to each other, as GermaNet is intended to be a German version of WordNet. Moreover, due to the different structure of RuTes, synonyms could not be compared with the Russian subset. Thus, the results of the preceding sections cannot be directly compared. However, the fact that approximately 50% of the relations whose entities were both contained in SemRelData and another knowledge base shows that the approach taken in this thesis is legitimate and yielded correct results.

The further investigation of the relations which are not present in the knowledge bases, although both related entities are, revealed that 42%–50% were not contained due to unclear or miscellaneous reasons. Due to the fact that the databases were not automatically extracted from an all-encompassing corpus, it would be reasonable to expect that the databases are incomplete. Moreover the fact that the dataset created in this thesis was based on slightly different relation definitions than that of the databases implicates differences in the comparison of those.

In comparison with the other two sets, the comparison with GermaNet resulted in a higher disagreement rate due to ambiguity, meaning that the word sense of a term in SemRelData was not contained in GermaNet. This could be explained by the different generation methods and coverages of the knowledge bases, WordNet and RuTes containing nearly half as many relations as GermaNet (see Table 2.4).

Moreover, WordNet has the lowest rate of disagreement in the categories RS and A, meaning that it has the greatest coverage of specific and ambiguous terms and relations. The reason for this may be the careful creation of WordNet, which has the longest creation history and was created completely by hand.

Furthermore, the results of the comparisons revealed that 12%–18% of the relations in the disagreement analyses were contextual, proving that the approach taken in this thesis produces previously neglected relations, which are important for the analysis and processing of natural language text.

9.2. Comparison to Pattern-created Taxonomy

As already cited above, the classification and extraction of semantic relations of words is preferably done by the use of patterns. The first and most popular patterns are that of Hearst (1992), which were later enhanced by Klausner and Zhekova (2011). The implementation of JoBim Text of those patterns was applied to the English source texts that were annotated for SemRelData. The full result may be found in the appendix (see Table A.16).

As the Hearst Patterns and their extensions are composed for English hyperonymy, only the hypernym relations of the English subset were compared with the extracted relations. Those were not lemmatized as the Hearst Patterns produce both lemmatized and inflected forms of nominals.

The pattern extractor selected 112 hypernym relations using the described patterns, whereas the English subset of SemRelData contains 553. Only eight relations were contained in both sets. To analyse the difference between the two sets, 50 random relations of the 112 that were contained in the pattern-extracted hypernym set were classified according to four labels:

- 1) True (T): the relation is valid and should be present in SemRelData
- 2) Lemma (L): the relation is not present in SemRelData, because it contains lemmas or inflected forms of the related words that are different in the original text⁴⁷, e.g. the pattern-based approach extracts the relations *primate* as a hypernym of *human* and *primates* as a hypernym of *humans*, whereas only the second version is in SemRelData, as it takes exclusively the word form that was present in the text.
- 3) General (G): the relation is too general to be encountered true or only a part of a noun compound is used for the relation, which makes the relation more general⁴⁸, e.g. the relation *variety* as a hypernym of *sweet orange* can be encountered as true, but the term *variety* is too general.
- 4) False (F): the relation is wrong, e.g. *government* as a hypernym of *free trade agreement*.

Error Type	Count
T	0
L	2
G	17
F	31

Table 9.5 Disagreement analysis of relations that were only in the set extracted with the enhanced Hearst Patterns

Table 9.5 shows that 62 % of the relations in the random test were wrong and 34% too general. At this point it shall be mentioned that in the random test, there was only one occurrence of a relation classified as general, which was not contained in SemRelData due to the discussed restriction of not relating both the full noun compound and parts of the compound to the same entity. 4% of the relations were not contained in SemRelData due to deviant word forms that are formed by the Hearst Patterns. The full disagreement analysis is shown in Table A.17.

In general it can be said that Hearst Patterns do not fit the claims of this task, as the dataset is too small to work effectively. When used in natural language processing or computer linguistic tasks, only relations with a high frequency are considered. Thus the results of most relations are either wrong or too general. Although it might be assumed that Hearst Patterns do not work as well on genres other than encyclopaedic, this could not be proven in this comparison. On the one hand, there is not enough data to prove this hypothesis, on the other hand, not all of the eight relations that were in both sets were from encyclopedia text.

9.3. Comparisons between Languages

9.3.1. Comparison of Semantic Relation Density

The number of nominals varies in different languages. As this thesis examines semantic relations within nominals, the number of potential relations may also be different. To compare the density of semantic relations in the language subsets, χ^2 was calculated using the number of noun compounds. This number is related to the number of potential relations in the set and the number of all relations in the individual subsets. The contingency table of all sets is presented below:

⁴⁷ In this case, the relation had to be true for other linguistic variants of one or both related words

⁴⁸ The annotation between parts of noun compounds and other nominals was forbidden by the guidelines if the whole of the compound was already related to the word. An example of such an occurrence is the relation between *wind* and *weather factor*. The relation between *wind* and *factor* is not annotated in SemRelData.

Set	№ NC	№ Trans. Rel.	Sum
German	4,766	3,436	8,202
English	5,510	3,390	8,900
Russian	4,572	2,416	6,988
Sum	14,848	9,242	24,090

Table 9.6 Contingency table denoting the number of noun compounds and transitive relations in the language subsets

The p-value of the χ^2 -test is very small $< 10^{-18}$, meaning that the distribution of semantic relations within different languages is not even. The test was conducted for the three possible pairings of languages. The resulting p-values ranged from 10^{-8} – 10^{-19} .

9.3.2. Comparison of Semantic Relation Types

To compare the distribution of classical semantic relation types within various languages, a nominal χ^2 -test was used. For the calculation, all relations were used. The table below shows the distribution of relation types within the corresponding language.

	German	English	Russian	Sum
Synonym	77	86	63	226
Co-Hyponym	335	281	296	912
Hypernym	508	553	296	1,357
Holonym	798	775	553	2,126
Sum	1,718	1,695	1,208	4,621

Table 9.7 Distribution of Semantic Relation Types in different languages

The p-value for the distribution between all three languages is $< 10^{-6}$, which signifies that the classical relation types are not evenly distributed among languages. The pairwise comparison reveals that the distribution of relation types within German and English is not significant with a significance value of rounded 6.56%. Both pairwise comparisons with Russian are highly significant, the significance value of rounded 0.01% of the comparison with German being noticeably lower than that of the comparison with English with $p < 10^{-9}$.

The following table displays the distribution of semantic types on a percentage basis.

	German	English	Russian	Sum
Synonym	4.48	5.07	5.22	4.89
Co-Hyponym	19.50	16.58	24.50	19.74
Hypernym	29.57	32.63	24.50	29.37
Holonym	46.45	45.72	45.78	46.01

Table 9.8 Proportional distribution of semantic relation types in different languages

Table 9.8 shows that the proportions of the semantic relation type distribution is the same for all four relation types, the frequency of type being already presented in the first column of the table in an ascending order, synonyms being the least and holonymy being the most frequent relation type. This goes for all three languages, except for the Russian subset, where the number of co-hyponyms is equal to that of hypernyms. The table also shows that the variance of types is mostly between co-hyponyms and hypernyms, whereas the variance between synonyms and holonyms varies at a maximum of 0.73%. Co-hyponyms appear most frequently in the Russian subset, whereas hypernyms appear most frequently in the English subset.

Summarising the comparison of semantic relation type distribution between languages it can be said that the χ^2 -test showed that there is a highly significant difference in the distribution, especially regarding the comparison with the Russian subset. Moreover, it can be added that the proportions of the distribution are comparatively similar for all three language subsets.

9.3.3. Conclusion of Comparison between Languages

The comparison of the classical semantic relations within the different language sets showed that although the difference in the distribution of these relations is highly significant for all three languages and the three possible pairings, the distribution of semantic relation types is similar in the English and German subsets, whereas the distribution of semantic relation types in the Russian subset varies with a high significance.

This difference could be explained with the genealogic relation of the Germanic languages in contrast to the Slavic language. On the one hand, the distribution of classical semantic types themselves may depend on the language family or the culture specific linguistic encoding of information. On the other hand, it is feasible that Russian expresses the same classical semantic relations not through nominals, but through pronouns or other grammatical constructions which avoid specific mention of the referred entity, e.g. the grammar of Russian allows sentences without a subject. This is shown in the following exemplary paragraph:

“Гибрид мандарина и горького севильского апельсина выведен в 1902 году в Алжире. Произрастает, в основном, в странах Средиземноморья; маленький, оранжевого цвета и круглый с твёрдой кожурой, плотно прилегающей к сочной мякоти. На базарах с конца октября по февраль. Основные поставщики — Испания, Марокко, Италия и Алжир.”

(“Klementin”, 2015, para. 1)

(Gloss: The hybrid of a mandarine and the bitter Seville orange was first grown in 1902 in Algeria. Grows mainly in Mediterranean countries; is small, of orange colour and round with a hard peel, tightly clinging to the juicy pulp. On the market from the end of October till February. Main suppliers [are] Spain, Morocco, Italy and Algeria.)

The paragraph shows that the discussed object *clementine* is not mentioned at all. Although the second and third sentences refer to hybrid, no subject is mentioned. This may not only demonstrate the possibilities of Russian grammar, as English and particularly German allow such constructions as well, but also a language specific preference of Russian to avoid frequent use of the same term due to redundancy. However, the less frequent use of nominals in Russian cannot be proven in the context of this thesis, as the highly ambiguous affixation and free word ordering in sentences makes automatic POS-tagging less reliable than in the other two languages, e.g. in the sentence

“Листья плотные, некрупные, на коротком, чуть крылатом черешке, с зазубринками по краю и острым концом.” (“Klementin”, 2015, para. 1)

(Gloss: The leaves are thick, not big, on a short, a little winged stem, with carved edges and a sharp end.)

крылатом (engl.: winged) is tagged as a noun and черешке (engl.: stem) is tagged as an adverb by the TreeTagger. According to Vazhenina and Markov (2013), the performance of TreeTagger on unknown words in Russian is 62.44%.

Although the distribution of semantic types has a highly significant variance, the proportional distribution of the relation types is similar in the three languages, with the exception of hypernyms being less frequent in Russian than in the other two languages. This means that although the distribution of types is different, the types are used in similar proportions in the different subsets. Adding this to the previously discussed hypothesis of Russian expressing the same relations in a grammatically different way, it could be assumed that the linguistic encoding of classical semantic relations is independent of language. However, the evidences presented in this thesis are not sufficient to draw this conclusion.

9.4. Comparisons between Genres

9.4.1. Comparison of Semantic Relation Density

To compare the density of semantic relation distribution within different genres, an analogous approach to that described in 9.3.1 was conducted. Table 9.9 demonstrates the contingency table for the χ^2 -test.

Set	№ NC	№ Trans. Rel.	Sum
Encyclopaedic	2,301	3,094	5,395
Literary	6,519	4,224	10,743
News	6,028	1,924	7,952
Sum	14,848	9,242	24,090

Table 9.9 Contingency table denoting the number of noun compounds and transitive relations in the genre subsets

The χ^2 -test for all three languages resulted in the significance level of 0%. The pairwise comparison of all three languages gave result between 0 and $< 10^{-103}$. It can be concluded that the density of semantic relations is not evenly distributed among the genre subsets.

9.4.2. Comparison of Semantic Relation Types

Table 9.10 presents the contingency table of the semantic relation type distribution in the different genres that was used to calculate χ^2 .

	Encyclopaedic	Literary	News	Sum
Synonym	106	67	53	226
Co-Hyponym	122	624	166	912
Hypernym	559	451	347	1,357
Holonym	760	970	396	2,126
Sum	1,547	2,112	962	4,621

Table 9.10 Distribution of semantic relation types in different genres

The p-values of the χ^2 -test for all three genres as well as all pairwise comparisons are between 10^{13} – 10^{-20} , meaning that the hypothesis of the semantic relation types being evenly distributed between the different genres can be rejected.

	Encyclopaedic	Literary	News	Sum
Synonym	6.85	3.17	5.51	4.89
Co-Hyponym	7.89	29.55	17.26	19.74
Hypernym	36.13	21.35	36.07	29.37
Holonym	49.13	45.93	41.16	46.01

Table 9.11 Proportional distribution of semantic relation types in different genres

Table 9.11 displays the distribution of semantic types on a percentage basis. As may be derived from the above table, the least frequent semantic relation type is *Synonym*, the most frequent is *Holonym*. In the encyclopaedic subset, co-hyponyms are nearly as frequent as synonyms with 1.04% more. For the news subset, the difference between the frequency of synonyms and co-hyponyms is bigger with 11.75% more, but for both subsets co-hyponyms are the second least frequent relation type, followed by hypernyms, which are the second most frequent relation type for both categories. However, for the literary set, the order of frequency for the co-hyponyms and hypernyms is reversed. Moreover, the table shows that literary texts have the smallest number of synonyms when compared with the other two genres.

In summary, it can be said that the χ^2 -test showed that there is a highly significant difference in the distribution of semantic relations and their types between genres. Furthermore, it can be stated that the percentual distribution is different for all three subsets, especially when comparing the literary subset to the other two.

9.4.3. Conclusion of Comparisons between Genres

The comparison results of the semantic relation and relation type distribution revealed that both distributions are not even in the three different genres. However, the proportional distribution of the relation types between encyclopaedic and news texts are more similar to each other than the distribution of these in literary text. This may be an indicator towards the linguistic encoding of knowledge through classical semantic relations being dependent on the genre, as both encyclopaedic and news texts share the aim to reveal information on a restricted subject.

9.5. Comparison between Categories in Wikipedia-subset

The three subcategories – *garment*, *organ*, and *fruit* that the Wikipedia subset was constructed of were compared using the χ^2 -test. The same approach as described in 9.3 was applied by first comparing the density of semantic relations in the subset and then comparing the semantic relation type distribution.

9.5.1. Comparison of Semantic Relation Density

Table 9.12 shows the distribution of density in the analysed categories in the encyclopaedic subset of SemRelData.

Set	№ NC	№ Trans. Rel.	Sum
Garment	617	397	1,014
Organ	782	558	1,340
Fruit	902	592	1,494
Sum	2,301	1,547	3,848

Table 9.12 Contingency table denoting the number of noun compounds and transitive relations in the category subsets

The significance level of the χ^2 -test hypothesis is 40% for the overall comparison and > 22% for all pairwise comparisons, meaning that the density of semantic relations is evenly distributed between these three encyclopaedic subsets.

9.5.2. Comparison of Semantic Relation Types

The distribution of semantic relation types in the subcategories of the encyclopaedic subset is shown in Table 9.13.

	Garment	Organ	Fruit	Sum
Synonym	23	20	63	106
Co-Hyponym	49	41	32	122
Hypernym	171	135	253	559
Holonym	154	362	244	760
Sum	397	558	592	1,547

Table 9.13 Distribution of semantic relation types in different subcategories of the encyclopaedic subset

The significance level of the χ^2 -test hypothesis is $< 10^{-21}$ for the overall comparison and all pairwise comparisons, meaning that the distribution of semantic relation types is not evenly distributed between the three encyclopaedic subsets.

	Garment	Organ	Fruit	Sum
Synonym	5.79	3.58	10.64	6.85
Co-Hyponym	12.34	7.35	5.41	7.89
Hypernym	43.07	24.19	42.74	36.13
Holonym	38.79	64.87	41.22	49.13

Table 9.14 Proportional distribution of semantic relation types in different categories

As depicted in Table 9.14, the proportional distribution of semantic relation types is specific for every subcategory of the encyclopaedic subset. Synonyms are most frequent in the *fruit* category when compared with the other two categories. This is the only set in which synonyms are more frequent than co-hyponyms. Co-hyponyms are most frequent in the *garment* category when compared with the other categories. Hypernyms are most frequent for both the *garment* and the *fruit* categories, whereas holonyms are the most frequent type in the *organ* category and most frequent when compared with the other two subsets.

Summarising the above, it can be stated that the difference in the distribution of semantic relation types in different subcategories of the encyclopaedic subset is highly significant and that the proportions of the individual subcategories are particular for every of these.

9.5.3. Conclusion of Comparison between Categories

The comparison of the distribution of classical semantic types between the herein defined categories in the encyclopaedic subset showed that the semantic relations are evenly distributed, which may indicate that the density of semantic relations is genre dependent. However, as this categorisation was implemented for one genre only, this assumption lacks evidence.

The fact that *fruit* has nearly twice as many synonyms on a percentage basis than the other two categories may be explained by the choice of articles in this category. It could be assumed that the articles on exotic fruit contained more synonyms than terms that are more known in a language. However, as the proportion of these fruit and the encyclopaedic corpus is too small, this assumption cannot be proven.

9.6. Comparison of Entities with the Highest Number of Relations

In this subchapter, the entities with the highest number of classical semantic relations of every single text are analysed in order to find out whether they have a special semantic meaning in the corresponding text or whether there are parallels in their classification when comparing languages and genres. Moreover, in the case of the more or less parallel literary texts, which are all translations or translations with the corresponding original, it can be analysed whether the same entities are most frequently used in the analysed relations.

The texts were analysed in the genre subsets as this allowed the same classification of the most frequent nominals in semantic relations. The following subchapters discuss the analysis and classification of those entities in detail. For the analysis of the entities, their lemmas were used applying the lookup lists of the lemmatizers that were described in 9.1. To examine the entities with the highest number of relations, the most frequent nominals within the texts were also calculated in order to detect differences. For the calculation of the most frequent nominals, the word forms as they were used in the source texts were chosen and later on the most frequent were manually lemmatized, as automatic lemmatization would be too complex considering that this is not the main focus. Frequencies of one were not considered.

9.6.1. Entities with the Highest Number of Relations within the Encyclopaedic Subset

In the following, the classification and analysis of the three most frequent entities within the ontologies of the individual encyclopaedic texts are discussed. The entities are classified according to the placement of the word⁴⁹ which refers to the subject of the article (the full table is displayed in Table A.18).

	Subject most frequent	Subject second frequent	Subject third frequent
German	16	1	3
English	15	4	1
Russian	12	6	1
Sum	43	11	5

Table 9.15 Frequency distribution of nominals within the relations of SemRelData of the word describing the subject of the article

Table 9.15 shows that in over 71% the nominal denoting the subject of the Wikipedia article is among the most frequently used entities in the ontology of the corresponding article. In over 18% it is amongst the second most frequently used entities and in over 8% it is amongst the third most frequently used entities. Moreover it can be said that the nominal referring to the entity described in the Wikipedia article is always amongst one of the three most frequently used entity lists.

To evaluate whether the impact of most frequent nominals within semantic relations is dependent on the semantic relations and not on nominals in general, not only the most frequent nominals in relations, but the most frequent nouns in this dataset were calculated. Table A.19 shows the most frequent nouns in the corresponding files. The frequency of most frequent nominals being the defined entity in the file is the same frequency as most frequent words within relations – over 71%. The second most frequent words are even more often the defined entity of the text – about 22%, in comparison with the second most frequent entities with the highest number of relations.

Most frequent words	43
2nd most frequent words	13
3rd most frequent words	4

Table 9.16 Distribution of nominals of SemRelData of the word describing the subject of the article

The comparison of the two frequency distribution shows that the entities with the highest number of relations contain not only the most frequent nominals, but also their synonyms.

9.6.2. Entities with the Highest Number of Relations within the Literary Subset

In the following the classification and analysis of the three most frequent entities within the ontologies of the individual literary texts will be discussed. The entities are classified according to the following labels:

- 1) Person/Character (e.g. mother, carpenter, people, ...)
- 2) Description items of character, such as clothes and body parts (e.g. hair, boot, eye, ...)
- 3) Description items of locations (e.g. house, planet, ocean, ...)
- 4) Feelings/conditions (e.g. agony, peace, happiness, ...)
- 5) Other (OTH)

⁴⁹ If the word describing the subject has synonyms, all of these are considered as a valid mentioning of it.

Table 9.17 shows the distribution of the frequency positions among different classes. It shall be noted that as several entities can have the same placement due to the same number of occurrences, the individual positions can be occupied by several entities.

	Pers/Char	Description of Pers/Char	Description of place	Feeling/condition	OTH
Most frequent words	40	13	11	1	4
2nd most frequent words	29	21	12	4	8
3rd most frequent words	25	16	9	4	6

Table 9.17 Distribution of frequency placement of most frequent entities among classes

As may be seen in Table A.20, in 85% of all literary texts⁵⁰, *Person/Character* was assigned to at least one of the most frequent entities of the parallel texts. Table A.17 shows that *Person/Character* is the most frequently assigned label on all three frequency placements, followed by the description of the character. The third most frequently assigned label is *Descriptions of place*. *Feeling and condition* are the least frequently assigned labels. The table also reveals that *Other* is the second last frequently assigned label, meaning that the most frequent entities with semantic relations belong to the one of the first three labels.

In the full table, the texts are ordered in packs of three parallel texts in the three various languages. What shall be noted here is that often, but not always, the same entity is listed amongst the most frequently used in the classical semantic relations in the literary set of SemRelData.

As in the previous chapter, in order to measure the impact of frequent nominals within semantic relations in comparison with the impact of frequent nominals in the source texts, the frequent nominals in the texts were classified according to the classes that were previously presented. However, one further class – that of Named Entities (NEs), which by definition are not contained in SemRelData, was added.

	Pers/Char	Description of Pers/Char	Description of place	Feeling/condition	OTH	NE
Most frequent words	28	16	6	3	0	18
2nd most frequent words	20	11	6	9	4	16
3rd most frequent words	10	5	3	3	0	2

Table 9.18 Distribution of nominals in SemRelData of the word classified according to their function

Nominals referring to NEs are the most frequently assigned entity in the frequency distribution of the individual texts. Most of these NEs refer to person or character names. The label *Other*, being one of the least frequently assigned labels in Table 9.17 is the second most frequently assigned label in the classification of nominal frequency distribution, followed by *Person/Character*. The detailed analysis is presented in Table A.21.

9.6.3. Conclusion of Entities with the Highest Number of Relations

The analysis of most frequent entities within the relations in SemRelData revealed that these entities have an important function in the corresponding text.

However, in the case of encyclopaedic texts these terms were nearly identical to the most frequent nouns in the texts. The only additional information that the entities with the highest number of relations yielded was that the synonyms of the most frequent nominals are semantically as important as the frequently mentioned term. In most cases, the most frequent nominals as well as the entities with the

⁵⁰ Exceptions to this observation of overall 60 literary texts were the following 9 files: *bovary_de.tsv*, *bovary_ru.tsv*, *chekhov2_ru.tsv*, *france2_ru.tsv*, *goriki2_en.tsv*, *goriki2_de.tsv*, *goriki2_ru.tsv*, *sandmann2_de.tsv*, *sandmann2_ru.tsv*.

highest number of relations were equal to the defined term. This was expected, as all words in the text serve the purpose of defining this term and thus are potentially semantically related to it.

The analysis of the entities with the highest number of relations as well as the analysis of the most frequent nominals revealed that the most frequent entities are persons or NEs in the case of the most frequent nominals. However, the distribution of the other classes was not similar in the analyses presented in Sections 9.6.1 and 9.6.2. Most frequent nominals were nearly as frequently classified as either *Other* or *Person/Character*. Nevertheless, in the analysis of the entities with the highest number of relations *Other* was the least frequently assigned class. Moreover, the second most frequently assigned class in the analysis of Section 9.6.1 were nominals referring to descriptions of the main character.

The facts summarized in the previous paragraph may indicate that semantic relations serve the purpose of defining and describing characters in literary texts. In addition to the reference to persons and person names being also frequently found in the analysis of the most frequent nominals, semantic relations inform about attributes of the literary characters.

10. Conclusion

The results of this project are presented separately for the two main tasks – annotation and analysis of classical semantic relations between nominals in context. In the end an answer to the central research question of this thesis is presented. All assumptions and interpretations of results are made under the provision that the source data set is too small and restrictive to be representative of general language use. Also all the other differing factors such as author background, original or translated texts, and diachronic differences between the texts may be responsible for fluctuations in the results.

10.1. Conclusion of the Annotation Task

One of the research questions was whether it is possible to find a uniform structure for the annotation of this task. The average inter-annotator agreement of 0.24 shows that agreement on the relations can be found. Moreover, the analysis of the different influential factors such as annotator⁵¹, language, genre, paragraph size and time show that only the time factor has a measurable influence on the inter-annotator agreement. Furthermore, comparison of different languages and genres showed that the density of the analysed relations as well as the distribution of types is not even between both genres and languages. Inter-annotator agreement does not vary according to these factors, although the factors produce different distributions of relation types. Thus it could be shown that neither genre nor language influence the annotator performance. The improvement of the guidelines influences the performance of the annotators, which shows that the structure for the annotation of this task is uniform and may be even improved with time and the improvement of the guidelines. Although semantic relations are unarguably dependent on context, as will be discussed further on, the concepts of classical semantic relations seem to be universal, which supports Murphy's meta-lexical approach (2003).

The detailed analysis of the confusion matrices showed that most disagreement between annotators is caused by disagreement on the relation itself, not the type of relation, meaning that the detection but not the classification of semantic relations between nominals causes difficulty in annotation, reflected by lower inter-annotator agreement. This assumption is also supported by the nearly twice as high agreement between curator and the annotators, meaning that both annotators found correct relations that were complementary.

10.2. Conclusion of Relation Statistics

Although the most frequent relation in WordNet is hyperonymy (Fellbaum, 1998), in SemRelData the most frequent relation is holonymy. The reason for this difference may be rooted in the different approaches that formed the basis of the creation of the databases, but also the definitions of the two relations. WordNet was created manually by professionals who wanted to create a linguistic ontology. The most fundamental relation in an ontology is hyperonymy, especially regarding the parallel to scientific ontologies. Hyperonymy is also the most often confused relation between the annotators of this project, meaning that if annotators did not agree on a relation between two words, one of them had annotated it as a hyperonymic one. This supports Murphy's claim of hyperonymy being the most fundamental relation to organize knowledge (2003), probably misleading annotators to annotate relations as hyperonymic. Cruse, Lyons and Pustejovsky regard hyperonymy as one of the major structural relations (as cited by Murphy, 2003). Wierzbicka, however, believes the role of hyperonymy

⁵¹ Although annotator 1 has a divergent agreement, this can be explained with the fact that this was the only non-linguist, meaning that a similar linguistic background ensures a similar level of agreement.

to be overestimated in human thinking (1984). The fact that all relations are most often confused with hyperonymy may show that human thinking is fixed on terms being related by exactly this relations. The actual number of hypernyms when compared with holonyms may, however, indicate that hyperonymy is not as present in natural language use as classic knowledge bases suggest. This project was created with the aim to detect semantic relations in continuous text and thus probably reflects the use of classical semantic relations between nominals in natural language texts. As stated by Cruse (1986) and Winston et al. (1987), holonymy is the least concretely defined relation, consisting of many subclasses, which probably leads to the high number of holonymic relations in SemRelData. However, Winston et al. (1987) claim that holonymy is particularly important for the human understanding of lexicon structure, which may be another reason for this phenomenon.

10.3. Conclusion of Universality of Semantic Relations

A research question already touched upon in the previous paragraphs is whether the use of semantic relations and semantic relation types is universal or dependent on factors such as language, culture and genre. The answer to this question is that both the distribution of semantic relations in general as well as semantic relation types is not even neither in language nor in genre.

However, it could be shown that the distribution of semantic relation types between German and English was classified as not significant, if only marginally, whereas both Germanic languages showed a highly significant difference in the distribution of types when compared with Russian. This could be interpreted as semantic relations differing according to genealogic relatedness, more concretely it could be that genealogically related languages have similar ways of expressing semantic relations between nominals, but not semantic relations in general. It is possible that in Russian the same relations are not expressed through nominals, but through pronouns or other constructions without the explicit mention of the referred entity. Such constructions are possible in Russian due to the highly complex morphology of the language. The morphology of the language and especially its ambiguous affixing is also the reason why POS-tagging is less reliable than in the other two languages, making a comparison between the noun compound distributions inefficient. Thus the historic linguistic question of whether taxonomies are universal or culture specific that was examined by Murphy (2003) cannot be answered in this context.

The χ^2 comparison of semantic relations and semantic relation type density showed that both factors are not evenly distributed in the three different genres. The proportional distribution of the genre subsets showed that the encyclopaedic and the news subset have a more similar distribution of types than both comparisons with the literary subset. Encyclopaedic and news texts proportionally contain more similar relation types than the literary set. This could be an indicator of different semantic relation types encoding different kinds of information. Both encyclopaedic and news texts have the aim to inform the reader on one specific subject, which the literary texts do not aim at. The comparison of different categories within the encyclopaedic subset showed that the density of semantic relations is evenly distributed between all categories. This could lead to the assumption of classical semantic relation density between nominals being dependent of genre type. This assumption is also supported by the similar distribution of classes that were assigned to the most frequent entities in the ontologies of the individual files. If both the distribution of entity classes as well as the distribution of relations are genre-specific, the hypothesis of classical semantic relations being dependent on genre is probable. Nevertheless, this assumption cannot be drawn, as only the Wikipedia subset was analysed according to categories. The analysis of the proportions of semantic relation types in the categories showed that articles defining *fruit* contain nearly twice as much synonyms as the other two categories. This may be due to the fact that many of the chosen fruits were exotic and thus had many terms in the different languages. However, as articles on only seven different fruits were used, a statistical analysis of these observations would not be convincing.

10.4. Conclusion of the Contextual Approach

Another research question in this project was whether the contextual approach finds other relations than that found with non-contextual or pattern-based approaches. The comparison with the relations extracted from the same dataset using Hearst Patterns resulted in only eight similar relations with SemRelData. The classification of the relations contained in the pattern based approach showed that most extracted relations are false. However, the conclusion of that is not that Hearst Patterns are wrong in general. In automatic relation extraction, only relations with a high frequency are chosen for the final set. Yet the source set is too small to make use of this technique. Besides, Hearst Patterns may not be as efficient in genres other than encyclopaedic texts.

The comparison of SemRelData and the knowledge bases WordNet, GermaNet and RuTes showed that approximately half of the relations whose entities are contained in both sets are also present in both sets. Between 42%–50% of the random test set chosen from the relations present in SemRelData were not present in the other knowledge base due to unknown or not further specified reasons. As the test sets have been manually created, the lack of many relations is natural. Moreover, some relations were not in the dataset due to definitions in SemRelData e.g. holonymy was defined as non-transitive in SemRelData, whereas WordNet defines the relations as transitive due to a subclassification of holonymy. Only the comparison with GermaNet resulted in a relatively high error rate due to lack of ambiguous word senses. It can also be stated that in the comparison with both RuTes and GermaNet the second most frequent reason for the relation not being in the knowledge bases was the relation being too specific. The comparatively low count of terms that were classified as too specific or ambiguous in the WordNet comparison could be due to size and prevalence of WordNet, meaning that if the other knowledge bases were bigger, the distribution of the relations not contained in the knowledge base would be different. 12–18% of the relations in the random test sets were classified as contextual, meaning that this approach produces different relations than previous approaches, which could be useful for linguistic or natural language processing tasks concerned with semantics and context.

The fact that 12%–18% of the relations in the random test sets were contextual together with the fact that a pattern-based approach to the same source texts produces completely different relations shows that the approach taken in this thesis finds classical semantic relations between nominals that were previously neglected. This confirms Cruse's (1986), Lyons' (as cited by Murphy, 2003) and Murphy's (2003) statements of semantic relations being dependent on the context.

10.5. Conclusion of the Function of Semantic Relations

A further research question in this project was whether terms with many relations have a special function in the text. The answer to this question is not trivial, as it is difficult to determine which terms have a special function in a scientific way. In the case of encyclopaedic texts, the percentage of the most frequent entities within relations and the most frequent nominals in the text being the defined term is the same. This implies that semantic relations do not have a great impact on the linguistic representation of knowledge in encyclopaedic text. As the whole text suits the purpose of defining this term, it is neither a surprise that many other terms have a semantic relation to the described term, nor is it surprising that the defined term is mentioned more often than other nominals in the text. However, a closer look at the comparison shows that the most frequent entities within relation do not only contain one term that is defined in the article, but all its synonyms. With reference to the importance of classical semantic relations this could mean that they encode the information that all the synonyms of the term, although not mentioned as often and explicit in relation to the other terms, have the same relations as the frequently mentioned term. Not only information on the other semantic relations of the synonyms is encoded in this way, but all information that is provided in the text on this term. Thus, the most frequent entities within ontologies of individual encyclopaedic articles probably have a special function in encyclopaedic text, even if it is not of great importance.

The classification of the most frequent entities within the literary texts revealed that the most frequent entities are persons, which matches the results of the most frequent nominals in the texts. These revealed that *NEs*, followed by personages are the most frequent nominals. However, the distribution between the other different function classes was not similar for most frequent entities within semantic relations and nominals in general. Nominals classified as miscellaneous were nearly as frequent as terms classified as *Persons/Characters* in the distribution of most frequent nominals, whereas it was the least frequent class in the distribution of the most frequent entities within the relations. Moreover, terms referring to the description of characters were the second frequent class in the distribution of entities within relations. Thus it can be assumed that semantic relations between nominals serve the purpose of linguistic information encoding more than nominals do on their own. More specifically, semantic relations may serve the purpose of defining or specifying terms in texts, as the frequent use of attributes of literary characters in the literary texts may indicate. This would support Lyons' and Cruse's theories of terms being defined through other terms in context (Lyons as cited by Murphy, 2003; Cruse, 1986).

This leads to the most important and interesting research question of this thesis: do semantic relations have a crucial function in the linguistic encoding of knowledge? Considering the previously summarized and analysed results this question can be answered positively. Throughout languages and genres, annotators were able to equally effectively find semantic relations that they agreed upon exceedingly in the course of guideline improvement. Most importantly, they agreed on the classification of found relations, meaning that in general the concepts they were annotating were already present in their semantic understanding of language. And although the distribution of number and type within genres and languages were proven to be uneven, the importance of the most frequent terms within the relations of a text showed that terms with many relations bear an important function in the text. As the literary text made use of semantic relations in order to define persons and locations, it could be assumed that information is displayed by semantic relations in this genre. Literary texts do not have the primary aim to display information as encyclopaedic texts, thus the detection of the methods behind the information encoding cannot be as easily detected. Encyclopaedic text may repeat the important entities often, so as to reveal the information and the most important terms, whereas literary texts may reveal information in a rather concealed way, as frequent repetition of terms is a stylistic device that is only used to openly stress the importance of the repeated information.

10.6. Final Conclusion

In general it can be concluded that classical semantic relations are concepts that can be agreed upon. The context dependent approach reveals more relations than previous approaches, meaning that context is an important factor in semantic relation detection. It was shown that semantic relations are partly dependent on context, language and genre. Moreover, classical semantic relations within nominals have an important role in the linguistic representation of knowledge.

11. Further Work

In this section, possibilities to use the created dataset, open research questions as well as hypothetical subsequent work are presented.

SemRelData could be further improved by better handling the tokens which were marked as orthographic mistakes or as parts of noun compounds. Moreover, relations occurring twice in a paragraph could be synchronized. Both the synchronisation and the spell verification would not only complete the ontology, but the impact of semantic relations could be more thoroughly analysed, as all relations that are contained in the text would be represented in a more comparable and quantified way.

Furthermore, the evaluation of the dataset could also be expanded to co-hyponyms. In the comparison with the other knowledge bases, one may analyse on which level a similar hypernym can be found. For a thorough analysis of this relation, however, the guidelines should be adjusted by relating all possible co-hyponyms on several levels with an additional specification of the similar hypernym. The co-hyponymy relation is a promising relation, especially because it could provide more information on more popular classical semantic relations such as synonymy and antonymy, as these could be seen as subclasses. Some of the co-hyponyms found in SemRelData were actually antonyms, such as e.g. *heart* and *mind* or *joy* and *agony*. Thus, more information on the nature and distribution of antonyms between nominals could be researched through the study of co-hyponyms.

Although SemRelData may not be used to train machine learning algorithms in the current condition, the continuous improvement of the κ and the guidelines indicates that an automation of the annotation process is conceivable. To improve the current situation of semantic relation detection, the relations of the dataset could be further analysed automatically in order to find patterns that encode classical semantic relations beyond the scope of sentences. This could be used to automatically find more relations. As discussed earlier, classic semantic relations play a role in the linguistic encoding of knowledge. Thus, tasks that have the aim to extract knowledge would benefit from an automation of the herein discussed annotation. If a machine learning algorithm marking classical semantic relations within paragraphs of texts from diverse genres could be developed, it would improve tasks such as information retrieval, question answering, word sense disambiguation, automatic text classification, automatic text summarization, machine translation, semantic relatedness and similarity between words and documents and other context sensitive tasks, as all of these tasks already make use of semantic relations and would benefit from the contextual component of the herein presented.

After a sense disambiguation, the relations that are not contextual could be added to the existing knowledge bases. Moreover, synsets of word senses not contained in the analysed databases could be included.

A further analysis of context-dependent relations could reveal more information on the creation of context, e.g. it could be analysed, whether there are different patterns encoding context-sensitive or context-insensitive semantic relations.

To analyse whether the reason for the difference between the two Germanic languages and Russian is actually genealogical, a greater dataset with more related languages, e.g. additions of other Slavic languages and other language groups, would allow clearer and more justified statements.

Furthermore, the definitions of frequent and infrequent terms within languages could be further researched in order to analyse differences in the semantic relations. An indicator that there is a difference in the definition of terms is the high number of synonyms in the *fruit* category of the Wikipedia subset, which contained many exotic fruits each having many synonyms in comparison with the rather known fruits in the analysed language. However, to do so, more texts than used in this thesis and also more categories than *fruit* should be chosen.

Moreover, the importance of the most frequent entities within relations could be further researched, not only analysing the described term or function classes, but all entities with a focus on a possible correlation of the entities' semantic relation types and the role of the term in the text.

The differences in the distribution of semantic relations and semantic relation types in different genres could be further analysed in order to find out how different genres encode information. This knowledge may help in tasks such as information retrieval, text processing, error correction and summarization.

Reference List

- Bakeman, R., & Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*: Cambridge University Press.
- Balkanova, V., Sukhonogov, A., & Yablonskij Sergey (2004). Russian WordNet: From UML-notation to Internet/Intranet Database Implementation. *Global WordNet Conference 2004 Proceedings*, 31–38. Brno, Czech Republic, from [http://orbis-pictus.cz/id32402/jazyk/jazykove\(2da/aplikovana\(1_lingvistika/Ontologie/WordNet/Conference_2004/127.pdf](http://orbis-pictus.cz/id32402/jazyk/jazykove(2da/aplikovana(1_lingvistika/Ontologie/WordNet/Conference_2004/127.pdf).
- Biemann, C., Bordag, S., & Quasthoff, U. (2004). Lernen paradigmatischer Relationen auf iterierten Kollokationen. *LDV-Forum*, 19, 103–111, from https://www.lt.tu-darmstadt.de/fileadmin/user_upload/Group_LangTech/publications/pre-langtech/BiemannBordagQuasthoff03.pdf.
- Biemann, C., Quasthoff, U., Heyer, G., & Holz, F. (2008). ASV Toolbox – A Modular Collection of Language Exploration Tools. *Language Resources and Evaluation Conference*, 1760–1767. Marrakech, Morocco, from http://rec-conf.org/proceedings/lrec2008/pdf/447_paper.pdf.
- Bird, S. (2006). NLTK: The Natural Language Toolkit. *Proceedings of the International Conference on Computational Linguistics/Conference of the Association for Computational Linguistics on Interactive presentation sessions*, 69–72. Sydney, Australia, from <http://www.anthology.aclweb.org/P/P06/P06-4.pdf#page=79>.
- Bird, S., & Liberman, M. (2000). A Formal Framework for Linguistic Annotation. *Speech communication*, 33(1), 23–60, from <ftp://ftp.cis.upenn.edu/pub/sb/papers/cis-9901/cis-9901.pdf>.
- Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase. *Proceedings of the 2008 Association for Computing Machinery Association for Computing Machinery's Special Interest Group on Management of Data international conference on Management of data*, 1247–1250. Vancouver, Canada, from <http://ids.snu.ac.kr/w/images/9/98/sc17.pdf>.
- Bortz, J., & Weber, R. (2005). *Statistik [Statistics]: Für Human- und Sozialwissenschaftler [For humanities and social studies scientist]* (6th ed.). *Springer-Lehrbuch [Springer-Textbook]*. Berlin, Heidelberg: Springer Medizin Verlag Heidelberg [Springer Medicine Publishing House Heidelberg].
- Braslavski, P., Ustalov, D., & Mukhin, M. (2014). A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus. *Proceedings of the 13th Conference on Computational Natural Language Learning in Association for Computational Linguistics*, 101–104. Gothenburg, Sweden, from <http://www.aclweb.org/anthology/E14-2026>.
- Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249–254, from <http://www.anthology.aclweb.org/J/J96/J96-2004.pdf>.
- Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, Estevam R. Jr., & Mitchell, T. M. (2010). Toward an Architecture for Never-Ending Language Learning. *Proceedings of the 24th Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 1306–1313. Atlanta, Georgia, United States of America, from <https://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/acarlson/papers/carlson-aaai10.pdf>.
- Cohen, N. (2011). Define Gender Gap? Look Up Wikipedia's Contributor List. *New York Times*, 30(362), 1050–1056, from http://www.nytimes.com/2011/01/31/business/media/31link.html?_r=0.
- Contingency Table (2015). *Merriam Webster*. Retrieved June 20, 2015, from <http://www.merriam-webster.com/dictionary/contingency%20table>.
- Cruse, D. A. (1986). *Lexical semantics. Cambridge textbooks in linguistics*.

- Cambridge [Cambridgeshire], New York: Cambridge University Press.
- Davidov, D., & Rappoport, A. (2008). Classification of Semantic Relationships between Nominals Using Pattern Clusters. *Proceedings of Association for Computational Linguistics*, 227–235. Columbus, Ohio, United States of America, from <http://www.anthology.aclweb.org/P/P08/P08-1.pdf#page=271>.
- Entity (2015). *Merriam Webster*. Retrieved June 15, 2015, from <http://www.merriam-webster.com/dictionary/entity>.
- Fellbaum, C. (Ed.) (1998). *Language, speech, and communication. WordNet: An electronic lexical database*. Cambridge, Mass., London: MIT press.
- Fellbaum, C. (Ed.) (2013). *WordNet. The Encyclopedia of Applied Linguistics*: Wiley/Blackwell.
- Gel'venbeyn, I. G., Goncharuk, A. V., Lehel't, V., Lipatov, A. A., & Shilo, Viktor V. A. (2011). Avtomaticheskij perevod semanticheskij seti WordNet na russkij yazyk [Automatic translation of the semantic net WordNet into the Russian language]. *Trudy Mezhdunarodnogo seminara Dialog po kom'juternoj lingvistike i ejo prilozhenijam [Proceedings of the international seminar Dialog and its applications]*, Protvino, Russia, from <http://www.dialog-21.ru/Archive/2003/Goncharuk.pdf>.
- Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070), 900–901, from <http://www.nature.com/nature/journal/v438/n7070/full/438900a.html>.
- Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2007). SemEval-2007 task 04: classification of semantic relations between nominals. *SemEval-2007, Association for Computational Linguistics*, 13–18. Prague, Czech Republic.
- Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2009). Classification of semantic relations between nominals. *Language Resources & Evaluation*, 43(2), 105–121.
- Gittens, M. (2005). *Mimida; A mechanically generated Multilingual Semantic Network*, pp. 1–3, from <http://gittens.nl/gittens/topics/SemanticNetworks.pdf>.
- Gleick, J. (2013). Wikipedia's Women Problem. *The New York Review of Books*, from <http://www.nybooks.com/blogs/nyrblog/2013/apr/29/wikipedia-women-problem/>.
- Grodal, S., Gotsopoulos, A., & Suarez, F. (2014). The Co-evolution of Technologies and Categories during Industry Emergence. *Academy of Management Review*, 1–43.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2), 199–220, from <http://tomgruber.org/writing/ontolingua-kaj-1993.pdf>.
- Gvishani-Kosygina, L. A. (Ed.) (1980). *Osnovnye proizvedenija inostranoj hudozhestvennoj lieteratury: Evropa, Amerika, Avstralija: literaturno-bibliograficheskij spravocnik [Main works of foreign fiction literature: Europe, Amerika, Australia: literary-bibliographic catalogue]*. Moskow, Russia: Kniga.
- Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *International Conference on Computational Linguistics '92 Proceedings of the 14th conference on Computational linguistics*, 2, 539–545. Nantes, France, from <http://www.aclweb.org/anthology/C92-2082.pdf>.
- Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., et al. (2009). SemEval-2010 task 8: multi-way classification of semantic relations between pairs of nominals. *Conference of the North American Chapter of the Association for Computational Linguistics Workshop on Semantic Evaluations: Recent Achievements and Future Directions, Association for Computational Linguistics*, 94–99. Boulder, Colorado, United States of America.
- Henrich, V., & Hinrichs, E. (2011). Determining Immediate Constituents of Compounds in GermaNet. *Proceedings of Recent Advances in Natural Language Processing*, 420–426. Hissar, Bulgaria, from

- <http://www.anthology.aclweb.org/R/R11/R11-1.pdf#page=454>.
- Iteration (2015). *Merriam Webster*. Retrieved June 15, 2015, from <http://www.merriam-webster.com/dictionary/iteration>.
- Klaussner, C., & Zhekova, D. (2011). Lexico-Syntactic Patterns for Automatic Ontology Building. *Recent Advances in Natural Language Processing Student Research Workshop*, 109–114. Hissar, Bulgaria, from <http://www.aclweb.org/anthology/R/R11/R11-2.pdf#page=119>.
- Klementin (2015). *Wikipedia*. Retrieved June 01, 2015, from EINTRAGEN.
- Landis, R. J., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174, from http://www.dentalage.co.uk/wp-content/uploads/2014/09/landis_jr_koch_gg_1977_kappa_and_observer_agreement.pdf.
- Leech, G. (2005). Adding linguistic annotation. *Developing Linguistic Corpora: a Guide to Good Practice*, 17–29.
- Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., et al. (2014). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 167–195, from <http://semantic-web-journal.net/system/files/swj499.pdf>.
- Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.
- Lexicalization (2015). *Merriam Webster*. Retrieved June 15, 2015, from <http://www.merriam-webster.com/dictionary/lexicalization>.
- Loukashevich, N. V. (2011). *Tezaurusy v zadachah informazionnogo poiska [Thesauri in information seeking tasks]: Izdatel'stvo Moskovskogo universiteta [Moscow University Press]*.
- McEnery, T., & Wilson, A. (2004). *Corpus Linguistics. An Introduction* (2nd ed.). *Edinburgh Textbooks in Empirical Linguistics*. Edinburgh: Edinburgh University Press.
- Medushevskij, A. (2011). Stalinism kak model'. [Stalinism as a model.]: Obozrenie izdatel'skogp proekta «ROSSPEN»«Istoriya stalinisma» [Overview of the publishing project «ROSSPEN»«History of Stalinism». *Vestnik Evropy [Europe's Messenger]*, (30), from <http://magazines.russ.ru/vestnik/2011/30/m30-pr.html>.
- Miller, G. A. (1995). WordNet: A lexical database for the English language. *Communications of the ACM*, 38(11), 39–41.
- Miller, G. A., & Fellbaum, C. (1991). Semantic networks of English. *Cognition*, 41(1-3), 197–229.
- Miller, G. A., & Hristea, F. (2006). WordNet Nouns: Classes and Instances. *Computational Linguistics*, 32(1), 1–3, from <http://user.phil-fak.uni-duesseldorf.de/~bontcheva/WS0809OL/J06-1001.pdf>.
- Mitkov, R. (2004). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.
- Moore, A. (2000). *Andrew Moore's teaching resource site: Semantics - meanings, etymology and the lexicon*, from <http://www.universalteacher.org.uk/default.htm>.
- Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy, and other paradigms*. Cambridge, UK, New York: Cambridge University Press.
- Naber, D. (2004). OpenThesaurus: Building a Thesaurus with a WebCommunity, from <https://www.openthesaurus.de/download/openthesaurus.pdf>.
- Nastase, V., & Szpakowicz, S. (2003). Exploring noun-modifier semantic relations. *Fifth International Workshop on Computational Semantics*, 285–301. Tilburg, Netherlands.
- Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217–250, from <http://web.informatik.uni-mannheim.de/ponzetto/pubs/navigli12b.pdf>.
- Noy, N. F., & McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology*. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical*

- Informatics Technical Report SMI-2001-0880*, from http://liris.cnrs.fr/~amille/enseignements/Ecole_Centrale/What%20is%20an%20ontology%20and%20why%20we%20need%20it.htm.
- Plag, I. (2003). *Word-formation in English. Cambridge textbooks in linguistics*. Cambridge, New York: Cambridge University Press.
- Ratinov, L., & Roth, D. (2009). Design Challenges and Misconceptions in Named Entity Recognition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning in Association for Computational Linguistics*, 147–155. Boulder, Colorado, United States of America, from <http://www anthology.aclweb.org/WWW09/WWW09-11.pdf#page=163>.
- Reflexivity (2015). *Merriam Webster*. Retrieved June 20, 2015, from <http://www.merriam-webster.com/dictionary/reflexivity>.
- Rosario, B., & Hearst, M. (2001). Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy. *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 82–90. Pittsburgh, Pennsylvania, United States of America, from <http://www anthology.aclweb.org/WWW01/WWW01-0511.pdf>.
- Rosario, B., Hearst, M. A., & Fillmore, C. (2002). The descent of hierarchy, and selection in relational semantics. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 247–254. Philadelphia, Pennsylvania, United States of America.
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, 44–49. Manchester, United Kingdom.
- Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the Association for Computational Linguistics Special Interest Group on Linguistic data and corpus-based approaches-Workshop*, 47–50. Cambridge, Massachusetts, United States of America.
- Smith, D. G. (2000). *Woman Writers – An Exhibition of Works from the 17th Century to the present. 19th century*, from <http://www.library.unt.edu/rarebooks/exhibits/women/19th.htm>.
- Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A Core of Semantic Knowledge. *16th international World Wide Web conference*, 697–706. Banff, Alberta, Canada.
- Suhonov, A. M., & Yablonskij, S. A. (2004). Razrabotka russkogo WordNet [Implementation of a Russian WordNet]. *Trudy 6toj Vserossijskoj nauchnoj konferenzii "Elektronnye biblioteki: perspektivnye metody i tehnologii, elektronnye kolekcii", RCDL 2004 [Proceedings of the 6th All-Russian Scientific Conference "Electronic Libraries: Promising methods and technologies, RCDL 2004]*, Pushchino, Russia, from <http://rcdl.ru/doc/2004/paper28.pdf>.
- Surina, O. P. (2009). Rol' innostranyh yazykov v Rossii XVIII veka i v sovremennoj Rossii: svjaz' proshlogo s nastoyashim [The role of foreign languages in 18th century Russia and in modern Russia: link of past and present], from <http://www.yafalian.ru/konfer/080.pdf>.
- Tokar, A. (2012). *Introduction to English Morphology. Textbooks in English language and linguistics: Vol. 5*. Frankfurt: Lang, Peter, Internationaler Verlag der Wissenschaften.
- Transitive (2015). *Merriam Webster*. Retrieved June 15, 2015, from <http://www.merriam-webster.com/dictionary/transitive>.
- Turney, P. D., & Littman, M. L. (2005). Corpus-based Learning of Analogies and Semantic Relations. *Machine Learning*, 60, 251–278.
- Vazhenina, D., & Markov, K. (2013). Factored language modeling for Russian LVCSR. *International Joint Conference on Awareness Science and Technology & Ubi-Media Computing*, 205–211. Aizu-Wakamatsu City, Japan, from http://web-ext.u-aizu.ac.jp/~markov/pubs/iCAST_13.pdf.
- Wierzbicka, A. (1984). "Apples" Are Not a "Kind of Fruit": The Semantics of Human Categorization. *American Ethnologist*, 11(2), 313–328.

Winston, M. E., Chaffin, R., & Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. *Cognitive Science*, 11, 414–444.

Yimam, S. M., Eckart de Castilho, R., Gurevych, I., & Biemann, C. (2014). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations - Google Project Hosting. *Proceedings of Association for*

Computational Linguistics, 1–6. Baltimore, Maryland, United States of America. Retrieved March 21, 2015, from <https://code.google.com/p/webanno/>.

Zimmermann, A., Gravier, C., Subercaze, J., & Cruzille, Q. (2013). Nell2RFF: Read the Web, and turn it into RDF. *CEUR Workshop Proceedings, 994*, 2–8, from <http://ceur-ws.org/Vol-992/paper2.pdf>.

A. Appendix

File name	File source	Extracted on
bovary_de.txt	http://gutenberg.spiegel.de/buch/frau-bovary-2404/1	09.10.2014
bovary_en.txt	http://www.gutenberg.org/cache/epub/2413/pg2413.txt	09.10.2014
bovary_ru.txt	http://az.lib.ru/f/flober_g/text_0010.shtml	09.10.2014
krieg_de.txt	http://gutenberg.spiegel.de/buch/krieg-und-frieden-4040/1	09.10.2014
krieg_en.txt	http://www.gutenberg.org/cache/epub/2600/pg2600.txt	09.10.2014
krieg_ru.txt	http://ilibrary.ru/text/11/p.1/index.html	09.10.2014
idiot_de.txt	http://gutenberg.spiegel.de/buch/der-idiot-2098/1	09.10.2014
idiot_en.txt	http://www.gutenberg.org/cache/epub/2638/pg2638.txt	09.10.2014
idiot_ru.txt	http://az.lib.ru/d/dostoewskij_f_m/text_0070.shtml	09.10.2014
sandmann_de.txt	http://gutenberg.spiegel.de/buch/der-sandmann-3093/2	10.10.2014
sandmann_en.txt	http://www.gutenberg.org/files/32046/32046-h/32046-h.htm#sandman	10.10.2014
sandmann_ru.txt	http://rusbook.com.ua/russian_classic/beketova_ma/e_t_a_gofman_pesochnyiy_chelovek.1477	10.10.2014
france_de.txt	http://gutenberg.spiegel.de/buch/die-gotter-dursten-7856/2	27.10.2014
france_en.txt	http://www.gutenberg.org/cache/epub/24010/pg24010.txt	27.10.2014
france_ru.txt	http://www.litmir.net/br/?b=9123&p=1#section_2	27.10.2014
kipling_de.txt	http://gutenberg.spiegel.de/buch/das-dschungelbuch-2076/1	27.10.2014
kipling_en.txt	http://www.gutenberg.org/cache/epub/236/pg236.txt	27.10.2014
kipling_ru.txt	http://az.lib.ru/k/kipling_d_r/text_0070.shtml	27.10.2014
wells_de.txt	Wells, H. G. (1901). War of the Worlds. (G. A. Crüwell, Trans.). Moritz Perles, Wien. (Original work published 1898), 5-7	
wells_en.txt	http://www.gutenberg.org/cache/epub/36/pg36.txt	27.10.2014
wells_ru.txt	http://az.lib.ru/u/uells_g_d/text_1898_the_war_of_the_worlds.shtml	27.10.2014
chekhov_de.txt	http://gutenberg.spiegel.de/buch/kleine-erz-3979/26	29.10.2014
chekhov_en.txt	http://www.gutenberg.org/cache/epub/13417/pg13417.txt	29.10.2014
chekhov_ru.txt	http://lib.ru/LITRA/CHEHOW/kashtanka.txt	29.10.2014
gorki_de.txt	http://gutenberg.spiegel.de/buch/meister-erzahlungen-2859/8	29.10.2014
gorki_en.txt	http://www.gutenberg.org/cache/epub/13437/pg13437.txt	29.10.2014
gorki_ru.txt	http://www.libok.net/writer/560/kniga/38900/gorkiy_maksim/odnajdyi_osenyu/read/2	29.10.2014
gorki2_de.txt	http://gutenberg.spiegel.de/buch/meister-erzahlungen-2859/8	29.10.2014
gorki2_en.txt	http://www.gutenberg.org/cache/epub/13437/pg13437.txt	29.10.2014
gorki2_ru.txt	http://www.libok.net/writer/560/kniga/38900/gorkiy_maksim/odnajdyi_osenyu/read/2	29.10.2014
bovary2_de.txt	http://gutenberg.spiegel.de/buch/frau-bovary-2404/36	27.11.2014
bovary2_en.txt	http://www.gutenberg.org/cache/epub/2413/pg2413.txt	27.11.2014
bovary2_ru.txt	http://az.lib.ru/f/flober_g/text_0010.shtml	27.11.2014
krieg2_de.txt	http://gutenberg.spiegel.de/buch/krieg-und-frieden-4040/258	28.11.2014
sandmann2_de.txt	http://gutenberg.spiegel.de/buch/der-sandmann-3093/4	27.11.2014
sandmann2_en.txt	http://www.gutenberg.org/cache/epub/32046/pg32046.txt	27.11.2014
sandmann2_ru.txt	http://www.treffpunkt.ru/lit/read.php?id=9665&page=8&q=4	27.11.2014
kipling2_de.txt	http://gutenberg.spiegel.de/buch/das-dschungelbuch-2076/11	11.12.2014
chekhov2_de.txt	http://gutenberg.spiegel.de/buch/kleine-erz-3979/26	27.11.2014
chekhov2_en.txt	http://www.gutenberg.org/cache/epub/13417/pg13417.txt	27.11.2014
chekhov2_ru.txt	http://lib.ru/LITRA/CHEHOW/kashtanka.txt	27.11.2014
bovary3_de.txt	http://gutenberg.spiegel.de/buch/frau-bovary-2404/19	30.12.2014
bovary3_en.txt	http://www.gutenberg.org/cache/epub/2413/pg2413.txt	27.11.2014
bovary3_ru.txt	http://az.lib.ru/f/flober_g/text_0010.shtml	27.11.2014
sandmann3_de.txt	http://gutenberg.spiegel.de/buch/der-sandmann-3093/3	29.12.2014
sandmann3_ru.txt	http://rusbook.com.ua/russian_classic/beketova_ma/e_t_a_gofman_pesochnyiy_chelovek.1477/?page=6	29.12.2014

Table A.1 Sources of literary subset

File name	File source	Extracted on
durian_de.txt	http://de.wikipedia.org/wiki/Durian	06.10.2014
durian_en.txt	http://en.wikipedia.org/wiki/Durio_zibethinus	06.10.2014
durian_ru.txt	https://ru.wikipedia.org/wiki/Дуриан_цибетиновый	06.10.2014
orange_de.txt	https://de.wikipedia.org/wiki/Orange_(Frucht)	06.10.2014
orange_en.txt	https://en.wikipedia.org/wiki/Orange_(fruit)	06.10.2014
orange_ru.txt	https://ru.wikipedia.org/wiki/Апельсин	06.10.2014
apfel_de.txt	https://de.wikipedia.org/wiki/Äpfel	06.10.2014
apfel_en.txt	https://en.wikipedia.org/wiki/Malus	06.10.2014
apfel_ru.txt	https://ru.wikipedia.org/wiki/Яблона	06.10.2014
melone_de.txt	https://de.wikipedia.org/wiki/Zuckermelone	06.10.2014
melone_en.txt	https://en.wikipedia.org/wiki/Muskmelon	06.10.2014
melone_ru.txt	https://ru.wikipedia.org/wiki/Дыня	06.10.2014

File name	File source	Extracted on
clementine_de.txt	http://de.wikipedia.org/wiki/Clementine_(Frucht)	08.10.2014
clementine_en.txt	http://en.wikipedia.org/wiki/Clementine	08.10.2014
clementine_ru.txt	https://ru.wikipedia.org/wiki/Клементин	08.10.2014
kaktusfeige_de.txt	https://de.wikipedia.org/wiki/Opuntia_ficus-indica	08.10.2014
kaktusfeige_en.txt	https://en.wikipedia.org/wiki/Opuntia_ficus-indica	08.10.2014
kaktusfeige_ru.txt	https://ru.wikipedia.org/wiki/Опунция_индийская	08.10.2014
physalis_de.txt	http://de.wikipedia.org/wiki/Blasenkirschen	08.10.2014
physalis_en.txt	http://en.wikipedia.org/wiki/Physalis	08.10.2014
physalis_ru.txt	https://ru.wikipedia.org/wiki/Физалис	08.10.2014
catsuit_de.txt	http://de.wikipedia.org/wiki/Catsuit	08.10.2014
catsuit_en.txt	http://en.wikipedia.org/wiki/Catsuit	08.10.2014
catsuit_ru.txt	https://ru.wikipedia.org/wiki/Кэтсют	08.10.2014
hut_de.txt	http://de.wikipedia.org/wiki/Hut	08.10.2014
hut_en.txt	http://en.wikipedia.org/wiki/Hat	08.10.2014
hut_ru.txt	https://ru.wikipedia.org/wiki/Шляпа	08.10.2014
hose_de.txt	http://de.wikipedia.org/wiki/Hose	08.10.2014
hose_en.txt	http://en.wikipedia.org/wiki/Trousers	08.10.2014
hose_ru.txt	https://ru.wikipedia.org/wiki/Брюки	08.10.2014
boxershorts_de.txt	http://de.wikipedia.org/wiki/Boxershorts	08.10.2014
boxershorts_en.txt	http://en.wikipedia.org/wiki/Boxer_shorts	08.10.2014
boxershorts_ru.txt	https://ru.wikipedia.org/wiki/Боксёры_(одежда)	08.10.2014
weste_de.txt	http://de.wikipedia.org/wiki/Weste	28.12.2014
weste_en.txt	http://en.wikipedia.org/wiki/Waistcoat	28.12.2014
weste_ru.txt	https://ru.wikipedia.org/wiki/Жилет	28.12.2014
kilt_de.txt	http://de.wikipedia.org/wiki/Kilt	28.12.2014
kilt_en.txt	http://en.wikipedia.org/wiki/Kilt	28.12.2014
kilt_ru.txt	https://ru.wikipedia.org/wiki/Килт	28.12.2014
finger_de.txt	http://de.wikipedia.org/wiki/Finger	08.10.2014
finger_en.txt	http://en.wikipedia.org/wiki/Finger	08.10.2014
finger_ru.txt	https://ru.wikipedia.org/wiki/Палец	08.10.2014
haar_de.txt	http://de.wikipedia.org/wiki/Haar	08.10.2014
haar_en.txt	http://en.wikipedia.org/wiki/Hair	08.10.2014
haar_ru.txt	https://ru.wikipedia.org/wiki/Волосы	08.10.2014
zunge_de.txt	http://de.wikipedia.org/wiki/Zunge	08.10.2014
zunge_en.txt	http://en.wikipedia.org/wiki/Tongue	08.10.2014
zunge_ru.txt	https://ru.wikipedia.org/wiki/Язык_(анатомия)	08.10.2014
auge_de.txt	http://de.wikipedia.org/wiki/Auge	08.10.2014
auge_en.txt	http://en.wikipedia.org/wiki/Eye	08.10.2014
auge_ru.txt	https://ru.wikipedia.org/wiki/Глаз	08.10.2014
brust_de.txt	http://de.wikipedia.org/wiki/Brust	28.12.2014
brust_en.txt	http://en.wikipedia.org/wiki/Thorax	28.12.2014
brust_ru.txt	https://ru.wikipedia.org/wiki/Грудная_клетка	28.12.2014
wirbelsäule_de.txt	http://de.wikipedia.org/wiki/Wirbelsäule	28.12.2014
wirbelsäule_en.txt	http://en.wikipedia.org/wiki/Vertebral_column	28.12.2014
wirbelsäule_ru.txt	https://ru.wikipedia.org/wiki/Позвоночник	28.12.2014
ohr_de.txt	http://de.wikipedia.org/wiki/Ohr	28.12.2014
ohr_en.txt	http://en.wikipedia.org/wiki/Ear	28.12.2014
ohr_ru.txt	https://ru.wikipedia.org/wiki/Ухо	28.12.2014

Table A.2 Sources of encyclopaedic subset

Filename	Page	Publishing Date	Extracted on
space_de.txt	https://de.wikinews.org/wiki/Daisuke_Enomoto_fliegt_als_vierter_Weltraum-Tourist_zur_ISS	08.03.2006	27.11.2014
space_en.txt	https://en.wikinews.org/wiki/Daisuke_Enomoto_will_be_the_fourth_space_tourist_at_the_ISS	10.03.2006	27.11.2014
space_ru.txt	https://ru.wikinews.org/wiki/Четвёртый_космический_турист	08.03.2006	27.11.2014
sudan_de.txt	https://de.wikinews.org/wiki/Südsudan_ist_unabhängig	13.07.2011	27.11.2014
sudan_en.txt	https://en.wikinews.org/wiki/South_Sudan_gains_independence	10.07.2011	27.11.2014
sudan_ru.txt	https://ru.wikinews.org/wiki/Южный_Судан_стал_независимым_государством	09.07.2011	27.11.2014
bush_de.txt	https://de.wikinews.org/wiki/George_Bush_unterzeichnete_Gesetz_zum_Bau_eines_Zauns_an_der_Grenze_USA-Mexiko	28.10.2006	07.12.2014
bush_en.txt	https://en.wikinews.org/wiki/Bush_signs_law_to_build_fence_at_US-Mexico_border	27.10.2006	07.12.2014

Filename	Page	Publishing Date	Extracted on
bush_ru.txt	https://ru.wikinews.org/wiki/Буш_подписал_закон_о_строительстве_забора	26.10.2006	07.12.2014
nsa_de.txt	https://de.wikinews.org/wiki/Abhörmaßnahmen_der_NSA_sorgen_für_Irritationen_in_Deutschland_und_Europa	02.07.2013	18.12.2014
nsa_en.txt	https://en.wikinews.org/wiki/United_States_spies_accused_of_illegally_bugging_the_United_Nations_headquarters	26.08.2013	18.12.2014
nsa_ru.txt	https://ru.wikinews.org/wiki/Spiegel:_АНБ_США_установило_«жучки»_в_представительствах_ЕС	30.08.2013	18.12.2014
bolivia_de.txt	https://de.wikinews.org/wiki/Bolivien:_Evo_Morales_siegt_bei_der_Präsidentenwahl	19.12.2005	18.12.2014
bolivia_en.txt	https://en.wikinews.org/wiki/Evo_Morales_wins_presidential_elections_in_Bolivia	19.12.2005	18.12.2014
bolivia_ru.txt	https://ru.wikinews.org/wiki/Президентские_выборы_в_Боливии	20.12.2005	18.12.2014
wm18_de.txt	https://de.wikinews.org/wiki/Fußballweltmeisterschaft_2018_in_Russland,_2022_in_Katar	02.12.2010	18.12.2014
wm18_en.txt	https://en.wikinews.org/wiki/FIFA_announce_Russia_to_host_2018_World_Cup,_Qatar_to_host_2022_World_Cup	02.12.2010	18.12.2014
wm18_ru.txt	https://ru.wikinews.org/wiki/Чемпионат_мира_по_футболу_2018_пройдёт_в_России	02.12.2010	18.12.2014
fukushima_de.txt	https://de.wikinews.org/wiki/Atomalarm_in_Japan_-_Explosionen_im_Kernkraftwerk_Fukushima_I	14.03.2011	18.12.2014
fukushima_en.txt	https://en.wikinews.org/wiki/Earthquake-damaged_Fukushima_nuclear_power_plant_triggers_evacuation	11.03.2011	18.12.2014
fukushima_ru.txt	https://ru.wikinews.org/wiki/Японский_Чернобыль	22.08.2011	18.12.2014
räikkonen_de.txt	https://de.wikinews.org/wiki/Kimi_Räikkönen_gewann_im_März_2007_den_Großen_Preis_von_Australien	03.06.2007	19.12.2014
räikkonen_en.txt	https://en.wikinews.org/wiki/Kimi_Räikkönen_wins_2007_Australian_Grand_Prix	18.03.2007	19.12.2014
räikkonen_ru.txt	https://ru.wikinews.org/wiki/Кими_Райконнен_выиграл_Гран-при_Австралии_2007_года	19.03.2007	19.12.2014
neuzealand_de.txt	http://de.wikinews.org/wiki/100_Eisberge_auf_dem_Weg_nach_Neuseeland	04.11.2006	30.12.2014
neuzealand_en.txt	http://en.wikinews.org/wiki/100_icebergs_heading_for_New_Zealand	04.11.2006	30.12.2014
neuzealand_ru.txt	https://ru.wikinews.org/wiki/100_айсбергов_движутся_к_Новой_Зеландии	04.11.2006	30.12.2014
island_de.txt	http://de.wikinews.org/wiki/Ausbruch_des_Vulkans_Eyjafjallajökull_behindert_Luftverkehr	16.04.2010	30.12.2014
island_en.txt	http://en.wikinews.org/wiki/European_airspace_closed_by_volcanic_ash	15.04.2010	30.12.2014
island_ru.txt	https://ru.wikinews.org/wiki/Из-за_извержения_исландского_вулкана_отменяются_авиарейсы_на_севере_Европы	15.04.2010	30.12.2014
nasa_de.txt	http://de.wikinews.org/wiki/NASA:_Rasanter_Rückgang_des_„Ewigen_Eises“_in_der_Arktis	15.09.2006	30.12.2014
nasa_en.txt	http://en.wikinews.org/wiki/NASA:_Arctic_Sea's_icecap_is_melting	14.09.2006	30.12.2014
nasa_ru.txt	https://ru.wikinews.org/wiki/Льды_Арктики_тают	15.09.2006	30.12.2014
hiroshima_en.txt	http://de.wikinews.org/wiki/60._Jahrestag_des_Atombombenabwurfes_über_Hiroshima	06.08.2005	18.12.2014
hiroshima_de.txt	http://en.wikinews.org/wiki/America's_atomic_bombing_commemoration_held_in_Hiroshima	07.08.2005	18.12.2014
hiroshima_ru.txt	https://ru.wikinews.org/wiki/Всемирный_день_борьбы_за_запрещение_ядерного_оружия	06.08.2006	18.12.2014
pyeongchang_de.txt	http://de.wikinews.org/wiki/Pyeongchang_richtet_Olympische_Winterspiele_2018_aus	14.07.2011	18.12.2014
pyeongchang_en.txt	http://en.wikinews.org/wiki/South_Korean_city_wins_2018_Winter_Olympics	06.07.2011	18.12.2014
pyeongchang_ru.txt	https://ru.wikinews.org/wiki/Олимпийские_игры_2018_года_пройдут_в_южнокорейском_Пхенчхане	06.11.2011	18.12.2014
sanfrancisco_de.txt	http://de.wikinews.org/wiki/Bruchlandung_eines_südkoreanischen_Verkehrsflugzeuges_in_San_Francisco	08.07.2013	18.12.2014
sanfrancisco_en.txt	http://en.wikinews.org/wiki/Asiana_Boeing_777_crashes_upon_landing_at_San_Francisco_International_Airport	06.07.2013	18.12.2014
sanfrancisco_ru.txt	https://ru.wikinews.org/wiki/Авиакатастрофа_Boeing_777_в_Сан-Франциско	06.07.2013	18.12.2014
serbien_de.txt	http://de.wikinews.org/wiki/Serbien:_Mutmaßlicher_Kriegsverbrecher_Ratko_Mladić_verhaftet	27.05.2011	18.12.2014
serbien_en.txt	http://en.wikinews.org/wiki/Ratko_Mladić_arrested_for_war_crimes	28.05.2011	18.12.2014
serbien_ru.txt	https://ru.wikinews.org/wiki/Арестован_Ратко_Младич	26.05.2011	18.12.2014
wm10_de.txt	http://de.wikinews.org/wiki/Fußball-WM:_Tintenfisch_„Paul“_sagt_Sieg_Deutschlands_im_„kleinen_Finale“_gegen_Uruguay_voraus	09.07.2010	18.12.2014
wm10_en.txt	http://en.wikinews.org/wiki/Spain_defeat_the_Netherlands_1-0_in_	12.07.2010	18.12.2014

Filename	Page	Publishing Date	Extracted on
	extra_time_to_win_2010_FIFA_World_Cup		
wm10_ru.txt	https://ru.wikinews.org/wiki/Испания_выиграла_чемпионат_мира_по_футболу	11.07.2010	18.12.2014
unruhen_de.txt	http://de.wikinews.org/wiki/Unruhen_in_Großbritannien:_Lage_eskaliert	09.08.2011	18.12.2014
unruhen_en.txt	http://en.wikinews.org/wiki/Rioting_develops_throughout_England	09.08.2011	18.12.2014
unruhen_ru.txt	https://ru.wikinews.org/wiki/Масштабные_беспорядки_вспыхнул_и_ещё_в_нескольких_городах_Англии	09.08.2011	18.12.2014
irkutsk_de.txt	http://de.wikinews.org/wiki/Flugzeugunglück_in_Irkutsk	09.07.2006	18.12.2014
irkutsk_en.txt	http://en.wikinews.org/wiki/Passenger_airplane_crashes_in_Siberia	09.07.2006	18.12.2014
irkutsk_ru.txt	https://ru.wikinews.org/wiki/Кршшение_пассажирского_самллёта_в_Иркутшке	09.07.2006	18.12.2014
romney_de.txt	http://de.wikinews.org/wiki/Republikanische_Vorwahlen:_Florida_gieht_an_Mitt_Romney	01.02.2012	18.12.2014
romney_en.txt	http://en.wikinews.org/wiki/Mitt_Romney_wins_2012_Florida_primary	02.02.2012	18.12.2014
romney_ru.txt	https://ru.wikinews.org/wiki/Мишт_Ромни_одержал_победу_во_Флориде	01.02.2012	18.12.2014
nordkorea_de.txt	http://de.wikinews.org/wiki/Nordkorea_führt_AAatomwaffentes_durch	09.10.2006	18.12.2014
nordkorea_en.txt	http://en.wikinews.org/wiki/North_Korea_claims_it_has_conducted_a_nuclear_test	09.10.2006	18.12.2014
nordkorea_ru.txt	https://ru.wikinews.org/wiki/Ядерные_испытания_в_Северной_Корее	09.10.2006	18.12.2014

Table A.3 Sources of news texts

	Synonym	Co-Hyponym	Hypernym	Holonym	***UNCLEAR***	No Annotation	Sum
Synonym	8	2	3	2	0	48	63
Co-Hyponym	0	29	1	6	2	92	130
Hypernym	0	2	84	12	0	226	324
Holonym	0	1	1	118	0	325	445
UNCLEAR	0	0	3	20	0	70	93
No Annotation	10	65	68	231	9	1606	1989
Sum	18	99	160	389	11	2367	3044

Table A.4 Contingency table of Annotator 1 and Annotator 2

	Synonym	Co-Hyponym	Hypernym	Holonym	***UNCLEAR***	No Annotation	Sum
Synonym	11	0	5	2	0	42	60
Co-Hyponym	2	42	2	4	1	159	210
Hypernym	0	2	106	11	0	113	232
Holonym	0	1	3	117	0	138	259
UNCLEAR	0	0	1	5	0	44	50
No Annotation	20	69	144	579	19	3324	4155
Sum	33	114	261	718	20	3820	4966

Table A.5 Contingency table of Annotator 1 and Annotator 4

	Synonym	Co-Hyponym	Hypernym	Holonym	***UNCLEAR***	No Annotation	Sum
Synonym	34	1	0	1	0	27	63
Co-Hyponym	2	53	6	1	1	114	177
Hypernym	7	1	165	10	4	239	426
Holonym	3	1	7	197	10	309	527
UNCLEAR	0	0	2	2	0	18	22
No Annotation	55	119	287	530	82	4036	5109
Sum	101	175	467	741	97	4743	6324

Table A.6 Contingency table of Annotator 2 and Annotator 3

	Synonym	Co-Hyponym	Hypernym	Holonym	***UNCLEAR***	No Annotation	Sum
Synonym	13	0	1	0	0	9	23
Co-Hyponym	2	34	1	0	0	48	85
Hypernym	4	6	53	12	0	80	155
Holonym	1	0	0	69	0	53	123
UNCLEAR	1	0	1	2	0	21	25
No Annotation	27	46	100	163	26	1728	2090
Sum	48	86	156	246	26	1939	2501

Table A.7 Contingency table of Annotator 2 and Annotator 4

	Synonym	Co-Hyponym	Hypernym	Holonym	***UNCLEAR***	No Annotation	Sum
Synonym	43	0	2	1	1	50	97
Co-Hyponym	1	137	3	7	2	202	352
Hypernym	2	2	278	9	0	241	532
Holonym	0	3	2	485	1	280	771
UNCLEAR	0	0	0	0	0	0	0
No Annotation	12	76	143	624	28	5190	6073

Table A.8 Contingency table of Annotator 1 and Curator

	Synonym	Co-Hyponym	Hypernym	Holonym	***UNCLEAR***	No Annotation	Sum
Synonym	96	0	3	0	0	41	140
Co-Hyponym	7	225	5	2	3	191	433
Hypernym	13	6	549	20	7	293	888
Holonym	4	5	5	669	19	290	992
UNCLEAR	0	0	0	0	0	0	0
No Annotation	94	155	383	734	178	7564	9108
Sum	214	391	945	1425	207	8379	11561

Table A.9 Contingency table of Annotator 2 and Curator

	Synonym	Co-Hyponym	Hypernym	Holonym	***UNCLEAR***	No Annotation	Sum
Synonym	43	0	3	2	0	21	69
Co-Hyponym	1	97	1	0	2	99	200
Hypernym	3	2	292	4	3	168	472
Holonym	1	1	8	297	0	191	498
UNCLEAR	0	0	0	0	0	0	0
No Annotation	16	78	129	228	17	4320	4788
Sum	64	178	433	531	22	4799	6027

Table A.10 Contingency table of Annotator 3 and Curator

	Synonym	Co-Hyponym	Hypernym	Holonym	***UNCLEAR***	No Annotation	Sum
Synonym	54	0	2	0	1	26	83
Co-Hyponym	2	192	4	0	4	98	300
Hypernym	6	3	266	2	6	173	456
Holonym	2	4	9	250	5	340	610
UNCLEAR	0	0	0	0	0	0	0
No Annotation	19	95	101	115	57	5313	5700
Sum	83	294	382	367	73	5950	7149

Table A.11 Contingency table of Annotator 4 and Curator

File	Av. Paragraph Size in Tokens	Annotator κ	Curator κ
kaktusfeige_ru.tsv file:	21.00	-0.22	0.39
zunge_ru.tsv file:	21.50	0.09	0.52
nordkorea_ru.tsv file:	23.25	-0.04	0.40
gorki2_ru.tsv file:	24.20	0.00	0.39
bovary2_ru.tsv file:	27.60	0.30	0.31
physalis_ru.tsv file:	28.20	0.33	0.63
hut_ru.tsv file:	30.00	0.17	0.51
romney_ru.tsv file:	30.25	-0.05	0.34
wirbelsaeule_en.tsv file:	31.50	0.28	0.52
bovary2_en.tsv file:	31.60	0.40	0.43
chekhov2_ru.tsv file:	32.83	0.46	0.64
auge_ru.tsv file:	33.00	0.33	0.16
pyeongchang_ru.tsv file:	33.60	-0.04	0.56
sudan_ru.tsv file:	33.75	0.36	0.69
france2_ru.tsv file:	33.80	0.35	0.45
clementine_ru.tsv file:	34.25	0.46	0.70
space_de.tsv file:	34.75	0.50	0.56
apfel_en.tsv file:	35.00	0.49	0.63
bovary2_de.tsv file:	35.40	0.54	0.31
unruhen_ru.tsv file:	35.57	0.26	0.57
nasa_ru.tsv file:	35.67	0.15	0.45
raeikkonen_en.tsv file:	36.29	0.11	0.47
hose_de.tsv file:	36.50	0.24	0.59
weste_ru.tsv file:	37.00	0.70	0.73
zunge_de.tsv file:	37.00	-0.09	0.24
newzealand_ru.tsv file:	37.40	0.64	0.76
france2_de.tsv file:	38.40	0.23	0.54
nsa_ru.tsv file:	38.57	0.19	0.58
irkutsk_ru.tsv file:	39.50	-0.05	0.26

File	Av. Paragraph Size in Tokens	Annotator κ	Curator κ
ohr_ru.tsv file:	39.67	-0.15	0.34
serbien_ru.tsv file:	39.80	0.04	0.46
hiroshima_ru.tsv file:	40.33	0.25	0.40
gorki2_en.tsv file:	40.75	0.00	0.19
boxershorts_ru.tsv file:	41.00	0.23	0.41
space_ru.tsv file:	41.25	0.05	0.30
wm18_ru.tsv file:	41.67	-0.03	0.60
haar_ru.tsv file:	42.00	-0.11	0.20
hiroshima_en.tsv file:	42.20	0.05	0.42
weste_de.tsv file:	42.25	0.25	0.56
kipling_ru.tsv file:	42.67	0.09	0.42
wirbelsaeule_de.tsv file:	42.67	0.56	0.77
raeikkonen_ru.tsv file:	42.75	0.52	0.71
ohr_de.tsv file:	43.50	-0.06	0.38
melone_de.tsv file:	44.50	0.19	0.55
space_en.tsv file:	44.50	0.25	0.60
brust_de.tsv file:	44.67	0.28	0.64
bovary_ru.tsv file:	44.80	0.07	0.34
apfel_ru.tsv file:	45.00	0.12	0.37
island_ru.tsv file:	45.00	0.32	0.64
sandmann2_ru.tsv file:	45.47	0.54	0.76
sanfransisco_ru.tsv file:	45.62	0.20	0.36
clementine_de.tsv file:	46.00	0.30	0.67
wm10_ru.tsv file:	46.33	0.09	0.47
finger_ru.tsv file:	46.50	-0.03	0.38
pyeongchang_de.tsv file:	46.50	0.36	0.71
romney_en.tsv file:	46.86	0.08	0.35
chekhov_ru.tsv file:	48.00	0.55	0.74
catsuit_en.tsv file:	49.00	0.64	0.82
kaktusfeige_de.tsv file:	49.00	-0.06	0.26
chekhov2_en.tsv file:	49.20	0.26	0.46
nsa_en.tsv file:	49.57	0.09	0.49
orange_ru.tsv file:	49.67	0.33	0.53
bush_ru.tsv file:	50.00	0.49	0.35
chekhov2_de.tsv file:	50.00	0.26	0.57
nasa_en.tsv file:	51.22	0.41	0.68
nasa_de.tsv file:	51.43	0.14	0.58
gorki2_de.tsv file:	52.00	0.28	0.53
serbien_de.tsv file:	53.00	0.14	0.48
idiot2_en.tsv file:	53.27	0.28	0.52
wm18_de.tsv file:	53.33	-0.03	0.43
irkutsk_en.tsv file:	54.50	-0.03	0.33
melone_ru.tsv file:	55.00	0.25	0.55
romney_de.tsv file:	55.00	0.39	0.52
wm10_de.tsv file:	55.00	0.36	0.65
raeikkonen_de.tsv file:	55.50	-0.05	0.29
ohr_en.tsv file:	56.00	0.27	0.37
brust_ru.tsv file:	56.33	-0.27	0.17
wirbelsaeule_ru.tsv file:	56.67	0.19	0.55
chekhov_de.tsv file:	57.20	0.19	0.39
kilt_en.tsv file:	58.00	0.38	0.67
bovary_de.tsv file:	58.80	0.36	0.67
nordkorea_en.tsv file:	59.46	0.10	0.48
bovary_en.tsv file:	60.40	-0.08	0.16
island_en.tsv file:	60.50	0.08	0.49
bush_de.tsv file:	60.75	0.12	0.53
durian_en.tsv file:	61.00	0.18	0.46
physalis_en.tsv file:	61.33	-0.10	0.34
bovary3_ru.tsv file:	61.40	0.51	0.75
durian_de.tsv file:	62.00	0.22	0.59
krieg_ru.tsv file:	62.80	0.59	0.80
krieg_de.tsv file:	63.25	0.36	0.67
pyeongchang_en.tsv file:	64.20	0.07	0.40
newzealand_en.tsv file:	64.83	0.27	0.50
boxershorts_en.tsv file:	65.00	0.26	0.42
catsuit_de.tsv file:	66.00	0.37	0.69
irkutsk_de.tsv file:	66.00	0.23	0.54
weste_en.tsv file:	68.00	-0.14	0.19
apfel_de.tsv file:	69.00	0.14	0.45
nsa_de.tsv file:	69.00	0.30	0.61

File	Av. Paragraph Size in Tokens	Annotator κ	Curator κ
sanfransisco_en.tsv file:	69.83	0.27	0.43
haar_de.tsv file:	70.00	0.33	0.62
wm10_en.tsv file:	71.60	0.52	0.74
hiroshima_de.tsv file:	72.50	0.23	0.38
france_ru.tsv file:	73.40	-0.09	0.29
kipling_de.tsv file:	74.50	0.18	0.42
gorki_ru.tsv file:	74.60	-0.12	0.25
serbien_en.tsv file:	75.33	0.30	0.55
bush_en.tsv file:	76.56	0.12	0.47
bolivia_ru.tsv file:	76.75	0.22	0.42
krieg_en.tsv file:	77.00	0.26	0.34
chekhov_en.tsv file:	77.50	-0.02	0.23
finger_de.tsv file:	77.50	0.04	0.43
bovary3_de.tsv file:	77.60	0.58	0.79
sudan_en.tsv file:	77.64	0.28	0.56
newzealand_de.tsv file:	78.00	0.31	0.60
fukushima_ru.tsv file:	78.33	0.07	0.58
unruhen_de.tsv file:	78.50	0.26	0.49
hose_ru.tsv file:	79.00	-0.14	0.25
krieg2_de.tsv file:	80.50	0.38	0.59
france_de.tsv file:	81.80	-0.05	0.34
nordkorea_de.tsv file:	83.25	0.16	0.48
unruhen_en.tsv file:	83.75	0.26	0.54
brust_en.tsv file:	84.00	0.27	0.42
auge_de.tsv file:	84.33	0.19	0.48
kaktusfeige_en.tsv file:	85.00	0.06	0.38
bolivia_en.tsv file:	85.25	0.25	0.53
france_en.tsv file:	86.33	0.08	0.39
wells2_ru.tsv file:	87.17	0.55	0.76
orange_de.tsv file:	88.50	0.22	0.42
kipling_en.tsv file:	89.00	0.34	0.35
idiot_en.tsv file:	90.14	0.34	0.60
gorki_de.tsv file:	90.20	0.15	0.47
clementine_en.tsv file:	93.00	0.29	0.50
physalis_de.tsv file:	93.00	-0.01	0.26
orange_en.tsv file:	94.50	-0.06	0.08
kipling2_ru.tsv file:	95.67	0.43	0.66
krieg2_en.tsv file:	95.83	0.36	0.57
wells_ru.tsv file:	98.88	0.26	0.51
gorki_en.tsv file:	99.20	0.42	0.62
durian_ru.tsv file:	100.50	0.23	0.49
finger_en.tsv file:	104.00	-0.03	0.35
kilt_de.tsv file:	104.00	0.59	0.65
krieg2_ru.tsv file:	105.00	0.33	0.64
auge_en.tsv file:	107.00	0.16	0.42
wm18_en.tsv file:	109.00	0.25	0.55
idiot2_ru.tsv file:	109.40	0.23	0.58
sandmann3_ru.tsv file:	113.29	0.33	0.34
fukushima_en.tsv file:	113.33	0.19	0.55
wells2_en.tsv file:	116.00	0.33	0.42
kilt_ru.tsv file:	117.50	0.26	0.58
sudan_de.tsv file:	119.56	0.19	0.53
boxershorts_de.tsv file:	120.00	0.04	0.25
idiot_ru.tsv file:	121.00	0.15	0.48
sanfransisco_de.tsv file:	122.33	0.39	0.50
island_de.tsv file:	125.60	0.25	0.42
bolivia_de.tsv file:	126.00	0.02	0.47
wells2_de.tsv file:	126.00	0.25	0.57
idiot2_de.tsv file:	126.80	0.12	0.55
catsuit_ru.tsv file:	129.00	0.13	0.46
haar_en.tsv file:	131.00	0.28	0.62
kipling2_en.tsv file:	131.40	0.27	0.57
zunge_en.tsv file:	136.00	0.09	0.45
hut_de.tsv file:	142.00	0.10	0.46
hose_en.tsv file:	152.00	0.51	0.67
wells_en.tsv file:	159.20	0.15	0.36
fukushima_de.tsv file:	166.29	0.28	0.53
wells_de.tsv file:	166.80	0.04	0.20
bovary3_en.tsv file:	172.00	0.73	0.82
hut_en.tsv file:	173.00	0.43	0.64

File	Av. Paragraph Size in Tokens	Annotator κ	Curator κ
melone_en.tsv file:	175.00	0.00	0.35
idiot_de.tsv file:	177.50	0.31	0.55
sandmann3_de.tsv file:	181.20	0.29	0.47
kipling2_de.tsv file:	181.75	0.41	0.67
sandmann_ru.tsv file:	191.90	0.00	0.15
sandmann2_de.tsv file:	202.40	0.31	0.56
france2_en.tsv file:	216.00	-0.02	0.44
sandmann2_en.tsv file:	225.00	0.28	0.47
sandmann3_en.tsv file:	242.75	0.12	0.42
sandmann_en.tsv file:	390.75	0.17	0.41
sandmann_de.tsv file:	420.40	-0.01	0.31

Table A.12 Paragraph size/ κ correlation

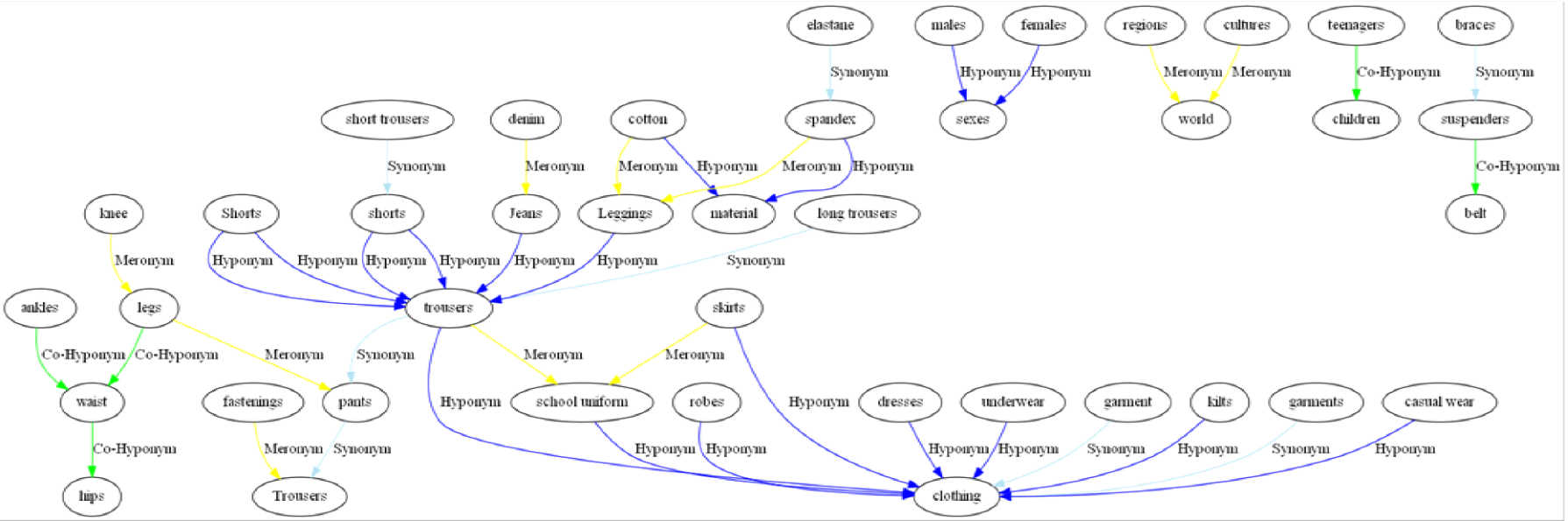


Figure 16 All relations in *hose_en*

Term 1	Relation	Term 2	Error Type
species	is a Hypernym of	chordates	S
congress	is a Holonym of	member	U
control	is a Hypernym of	law	K
casualties	is a Hyponym of	problems	K
clementine	is a Hyponym of	hybrid	Z
limb	is a Hyponym of	organ	U
meeting	is a Holonym of	women	U
Convention	is a Holonym of	members	U
crests	is a Meronym of	wavelets	U
warning	is a Hyponym of	recommendation	K
anguish	is a Hyponym of	sensation	U
appearance	is a Hypernym of	physiognomy	A
gaiters	is a Meronym of	wardrobe	U
engine	is a Meronym of	aircraft	U
men	is a Hypernym of	drivers	Z
paws	is a Holonym of	tips	U
men	is a Holonym of	hand	U
icecap	is a Holonym of	ice	U
expropriation	is a Hyponym of	interference	K
Physalis	is a Holonym of	husk	S
ballots	is a Holonym of	votes	U
state	is a Holonym of	Senator	Z
eyes	is a Meronym of	people	Z
days	is a Holonym of	noon	K
agony	is a Hypernym of	horror	U
place	is a Hypernym of	room	U
People	is a Hypernym of	celebrants	Z
martyrs	is a Meronym of	groups	K
people	is a Meronym of	groups	K
railways	is a Meronym of	infrastructure	U
children	is a Meronym of	humanity	U
area	is a Hypernym of	level	U
water	is a Meronym of	seas	U
boxers	is a Holonym of	fabric	L
man	is a Hypernym of	patriots	Z
garments	is a Hypernym of	shorts	U
dress	is a Hyponym of	garment	U
action	is a Hypernym of	negotiations	K
steeple	is a Holonym of	steps	U
men	is a Hypernym of	boys	U
riots	is a Hyponym of	protest	K
players	is a Hypernym of	captain	A
misinterpretation	is a Hyponym of	error	U
spine	is a Holonym of	vertebrae	S
company	is a Holonym of	entrepreneur	U
man	is a Hypernym of	papa	Z
travellers	is a Hyponym of	man	Z
limb	is a Hypernym of	finger	U
crops	is a Hyponym of	fruit	Z
genus	is a Hypernym of	durian	S

Table A.13 Detailed disagreement analysis of WordNet and SemRelData in 50 random relations

Term 1	Relation	Term 2	Error Type
Auge	is a Meronym of	Wirbeltieren	U
Nervenimpulse	is a Hypernym of	Reize	S
Orientierung	is a Holonym of	Sehsinns	S
Knie	is a Meronym of	Mann	U
Unterhosen	is a Hyponym of	Hosen	U
Schnitt	is a Meronym of	Hosen	U
Zauns	is a Hyponym of	Grenze	U
Staaten	is a Holonym of	Grenze	U
Barriere	is a Hypernym of	Wand	U
Augen	is a Meronym of	Tischler	Z
Clementine	is a Hyponym of	Zitruspflanzen	A
Mandarine	is a Hyponym of	Baum	A
Baum	is a Hypernym of	Mandarine	A
Daumen	is a Meronym of	Menschen	U
Kirche	is a Holonym of	Fassade	S
Stadtverwaltung	is a Holonym of	Maler	K
Tischler	is a Meronym of	Konvents	K

Term 1	Relation	Term 2	Error Type
Atommeiler	is a Holonym of	Reaktoren	S
Flusses	is a Holonym of	Wellen	U
Bund	is a Meronym of	Hosen	A
Rand	is a Meronym of	Sombrero	A
Hand	is a Meronym of	Leute	Z
Wagen	is a Holonym of	Fenstern	K
Gesichtern	is a Meronym of	Passagiere	U
Mannes	is a Hyponym of	Leute	Z
Regierungskommission	is a Holonym of	Personen	U
Familie	is a Hypernym of	Gattung	S
Menschen	is a Synonym of	Leute	Z
Pflanzenart	is a Hypernym of	Melone	Z
Eisschicht	is a Holonym of	Eis	U
Truppen	is a Meronym of	Landes	U
Baum	is a Holonym of	Orange	A
Judenkirschen	is a Meronym of	Obst-	U
Kapstachelbeere	is a Holonym of	Judenkirschen	U
Parteitag	is a Holonym of	Amtsinhaber	S
Stadt	is a Holonym of	Markt	U
Armen	is a Meronym of	Unhold	U
Worte	is a Meronym of	Zeile	U
Mann	is a Hypernym of	Sandmann	Z
Hause	is a Holonym of	Kinderstube	A
Gesicht	is a Holonym of	Katzenaugen	S
Laune	is a Hypernym of	Freude	K
Fluglinie	is a Holonym of	Sprecher	K
Referendums	is a Holonym of	Nation	S
Sicherheitstruppe	is a Holonym of	Soldaten	U
Prostitution	is a Holonym of	Prostituierten	U
Gemeinschaft	is a Hypernym of	Staatengemeinschaft	U
Krise	is a Hyponym of	Probleme	U
Familien	is a Hyponym of	Personen	U
Brand	is a Hypernym of	Brandstiftungen	K

Table A.14 Detailed disagreement analysis of GermaNet and SemRelData in 50 random relations

Term 1	Relation	Term 2	Error Type
яблоня	is a Meronym of	сад	U
политика	is a Meronym of	правительство	C
боксер	is a Hyponym of	труссы	A
позвоночник	is a Meronym of	грудная клетка	RS
мышца	is a Meronym of	грудь	RS
отверстие	is a Meronym of	грудная клетка	C
маска	is a Meronym of	одежда	U
ухо	is a Hyponym of	собака	U
палец	is a Meronym of	конечность	C
палец	is a Meronym of	ножка	A
богослужение	is a Hyponym of	собрание	RS
текст	is a Meronym of	петиция	RS
подпись	is a Meronym of	патриот	RS
человек	is a Meronym of	население	U
волна	is a Meronym of	река	C
орган	is a Meronym of	растение	C
волос	is a Meronym of	голова	U
волос	is a Meronym of	убийца	U
щека	is a Meronym of	убийца	U
спина	is a Meronym of	человек	U
двигатель	is a Meronym of	самолет	U
человек	is a Hyponym of	житель	SS
джунгли	is a Hyponym of	лес	U
господин	is a Hyponym of	человек	SS
пещера	is a Meronym of	гора	C
остров	is a Meronym of	архипелаг	RS
правительство	is a Meronym of	страна	U
министр	is a Meronym of	страна	U
орган	is a Meronym of	человек	A
человек	is a Hyponym of	млекопитающее	U
мандарин	is a Meronym of	апельсин	RS
премьер	is a Meronym of	страна	U

Term 1	Relation	Term 2	Error Type
сенатор	is a Meronym of	штат	U
башня	is a Meronym of	город	C
площадь	is a Meronym of	город	C
кровь	is a Meronym of	глаз	C
спина	is a Meronym of	отец	U
взлетно-посадочная полоса	is a Meronym of	аэропорт	RS
шасси	is a Meronym of	самолет	RS
крушение	is a Hyponym of	катастрофа	U
присяга	is a Hyponym of	церемония	U
гражданин	is a Meronym of	государство	SS
магазин	is a Hyponym of	дом	SS
вершина	is a Meronym of	холм	RS
вода	is a Meronym of	море	U
атмосфера	is a Meronym of	планета	U
дуга	is a Meronym of	позвонок	RS
язык	is a Hyponym of	вырост	RS
рукав	is a Meronym of	пиджак	U
вода	is a Meronym of	планета	C

Table A.15 Detailed disagreement analysis of RuTes and SemRelData in 50 random relations

Term 1	Term 2	Frequency
orange	variety	4
orange	popular variety	4
orange	varieties	4
orange	popular varieties	4
sweet_orange	varieties	2
sweet_orange	variety	2
satsum	varieties	2
satsum	popular varieties	2
honey	popular varieties	2
satsuma	popular varieties	2
satsuma_or_honey_sweet_orange	popular varieties	2
honey_sweet_orange	popular varieties	2
honey_sweet_orange	variety	2
honey_sweet_orange	varieties	2
honey_sweet_orange	popular variety	2
satsum	popular variety	2
satsuma_or_honey_sweet_orange	popular variety	2
satsuma	varieties	2
honey	variety	2
satsuma_or_honey_sweet_orange	varieties	2
sweet_orange	popular variety	2
sweet_orange	popular varieties	2
satsum	variety	2
honey	varieties	2
satsuma_or_honey_sweet_orange	variety	2
honey	popular variety	2
theft	act	1
theft	acts	1
vandalism	acts	1
vandalism	act	1
"	leaked_documents	1
"	document	1
"	leaked_documents	1
"	document	1
"	documents	1
"	documents	1
"	leaked_document	1
"	leaked_document	1
human	primate	1
humans	primates	1
human	primates	1
wind	factor	2
wind	weather_factors	2
wind	weather_factor	2
wind	factors	2
keratin	protein	1

Term 1	Term 2	Frequency
cortex	areas	1
visual_cortex	areas	1
cortex	area	1
visual_cortex	area	1
camera	security_measure	1
satellite	measures	1
sensor	measures	1
camera	measures	1
cameras	measures	1
cameras	security_measures	1
satellites	measures	1
sensor	measure	1
satellite	measure	1
sensors	security_measures	1
satellites	security_measures	1
satellite	security_measure	1
sensor	security_measures	1
camera	measure	1
sensors	measures	1
satellite	security_measures	1
camera	security_measures	1
sensor	security_measure	1
trade_agreements	Bolivian_governments	1
free_trade_agreement	governments	1
free_trade_agreement	Bolivian_governments	1
free_trade_agreements	governments	1
trade_agreement	governments	1
agreement	governments	1
agreements	Bolivian_governments	1
trade_agreements	governments	1
trade_agreement	past_Bolivian_governments	1
free_trade_agreements	Bolivian_governments	1
agreement	Bolivian_government	1
free_trade_agreement	Bolivian_government	1
free_trade_agreements	past_Bolivian_governments	1
free_trade_agreement	past_Bolivian_governments	1
free_trade_agreement	past_Bolivian_government	1
trade_agreement	Bolivian_government	1
agreements	governments	1
agreement	Bolivian_governments	1
trade_agreement	Bolivian_governments	1
free_trade_agreement	government	1
trade_agreement	past_Bolivian_government	1
agreements	past_Bolivian_governments	1
trade_agreement	government	1
trade_agreements	past_Bolivian_governments	1
agreement	past_Bolivian_government	1
agreement	government	1
agreement	past_Bolivian_governments	1
cloud	me	1
clouds	me	1
mother	evenings	1
mother	evening	1
owl	beak	1
owls	beaks	1
my_mother	evening	1
dark_clouds	me	1
dark_cloud	me	1
these_my_mother	evenings	1
my_mother	evenings	1
these_my_mother	evening	1
mad	me	1
owl	beaks	1
kilt	garment	1
kilt	garments	1
kilts	garments	1

Table A.16 Result of Hearst-Pattern application on the raw data of SemRelData

Term 1	Term 2	Error Type
orange	popular variety	G
orange	popular varieties	G
sweet orange	variety	G
satsum	popular varieties	G
satsuma	popular varieties	G
honey sweet orange	popular varieties	G
honey sweet orange	varieties	G
satsum	popular variety	G
satsuma	varieties	G
satsuma or honey sweet orange	varieties	G
sweet orange	popular varieties	G
honey	varieties	G
honey	popular variety	G
vandalism	act	L
"	leaked documents	F
" "	documents	F
human	primate	L
wind	factor	G
wind	factors	G
visual cortex	areas	G
visual cortex	area	G
satellite	measures	F
camera	measures	F
cameras	security measures	F
sensor	measure	F
sensors	security measures	F
satellite	security measure	F
camera	measure	F
satellite	security measures	F
sensor	security measure	F
free trade agreement	governments	F
free trade agreements	governments	F
agreement	governments	F
trade agreements	governments	F
free trade agreements	Bolivian governments	F
free trade agreement	Bolivian government	F
free trade agreement	past Bolivian governments	F
trade agreement	Bolivian government	F
agreement	Bolivian governments	F
free trade agreement	government	F
agreements	past Bolivian governments	F
trade agreements	past Bolivian governments	F
agreement	government	F
cloud	me	F
mother	evenings	F
owl	beak	F
my mother	evening	F
dark cloud	me	F
my mother	evenings	F
mad	me	F
kilt	garments	F

Table A.17 Detailed disagreement analysis of pattern-based approach and SemRelData in 50 random relations

File name	Most frequent words	Second most frequent words	Third most frequent words	Column no of topic
apfel_de.tsv.ont.lemma:	Apfel 5	Frucht 4 bäume 4 Zierstraucher 4	Kulturapfel 3 Straucher 3 Baum 3 Granatapfel 3 Familie 3 Dickicht 3 Walder 3	1
apfel_en.tsv.ont.lemma:	tree 7 shrub 7 apple 7	orchard apple 3 crab apples 3 crabapples 3 wild apples 3 crab 3	family 2 species 2	1
apfel_ru.tsv.ont.lemma:	род 8	вид 7	яблоня 5	3
auge_de.tsv.ont.lemma:	Auge 6	Tier 5	Orientierung 4	1
auge_en.tsv.ont.lemma:	eye 8	brain 7	organism 6	1
auge_ru.tsv.ont.lemma:	животный 3	глаз 3	человек 2	2

File name	Most frequent words	Second most frequent words	Third most frequent words	Column no of topic
		зрительный система 3	сенсорный орган 2 орган 2	
boxershorts_de.tsv.ont.lemma:	Boxershort 6 Hose 6	Unterwasche 5 Unterhose 5 Short 5	Eingriff 3	1
boxershorts_en.tsv.ont.lemma:	boxer 11	undergarment 9	fabric 5	1
boxershorts_ru.tsv.ont.lemma:	трусы 5	«семейник» 3 боксер 3 «семейный» 3		2
brust_de.tsv.ont.lemma:	Brust 10	Brustkorb 7	Frau 4	1
brust_en.tsv.ont.lemma:	thorax 7 chest 7	animal 5 human 5	organ 3 muscle 3	1
brust_ru.tsv.ont.lemma:	грудной клетка 7	сосуд 6 шея 6 нерв 6 пищевод 6 грудь 6 верхушка легкий 6 трахея 6	ребро 3	1
clementine_de.tsv.ont.lemma:	Clementine 19	Mandarine 16 Orange 16	Baum 12	1
clementine_en.tsv.ont.lemma:	clementine 6 fruit 6	citrus fruits 5 oil 5	orange production 4	1
clementine_ru.tsv.ont.lemma:	плод 8	клементина 7	напиток 5	2
durian_de.tsv.ont.lemma:	Stink 8 Durian 8	Frucht 7 Zibetbaum 7 Kasefrucht 7 Durianbaum 7	Malvengewachs 6	1
durian_en.tsv.ont.lemma:	durian 7	aroma 3 fragrance 3 smell 3 odour 3	flesh 2 onion 2 tree species 2 cultivar 2 fruit 2 reaction 2 genus 2	1
durian_ru.tsv.ont.lemma:	дерево 4	фрукт 2 суффикс 2 слово 2 название 2 плод 2		
finger_de.tsv.ont.lemma:	Finger 14	Daumen 9	Mensch 6	1
finger_en.tsv.ont.lemma:	finger 16	primate 13 phalanx 13 limb 13 human 13	pinky 11 organ 11 body 11 little finger 11 thumb 11	1
finger_ru.tsv.ont.lemma:	палец 5	конечность 4	птица 3 позвоночный 3	1
haar_de.tsv.ont.lemma:	Mensch 9	Haut 7	Saugetier 4 Haar 4	3
haar_en.tsv.ont.lemma:	keratin 8	hair 5	biomaterial 4 skin 5 protein 4	2
haar_ru.tsv.ont.lemma:	волос 4 растение 4	орган 3 кожный покров 3 «волосок» 3 трихом 3	защитный покров 2 животный 2	1
hose_de.tsv.ont.lemma:	Hose 7	Bein 4	Uberhose 3	1
hose_en.tsv.ont.lemma:	trouser 12	garment 11 clothing 11	short 7	1
hose_ru.tsv.ont.lemma:	брюки 7	ширинка 6 молния 6 гульфик 6 пуговица 6 кнопка 6 прорезь-клапан 6	нога 4 одежда 4	1
hut_de.tsv.ont.lemma:	Hut 10	Rand 7	Kopfbedeckung 5	1
hut_en.tsv.ont.lemma:	hat 7	fur hat 3	head covering 2 ear-flaps 2 head 2 construction workers 2	1
hut_ru.tsv.ont.lemma:	шляпа 3	чучело 2		1

File name	Most frequent words	Second most frequent words	Third most frequent words	Column no of topic
		осадки 2 солнце 2 лента 2 перо 2 ветер 2		
kaktusfeige_de.tsv.ont.lemma:	Opuntia 3	Familie 2 Pflanzenart 2 Fruchte 2	Kaktusfeige 1 Gattung 1 Feige 1 Opuntie 1	1
kaktusfeige_en.tsv.ont.lemma:	cactus 16	fig opuntia 9 prickly pear 9 barbary fig 9 spineless cactus 9 cactus pear 9 Opuntia ficus-indica 9		2
kaktusfeige_ru.tsv.ont.lemma:	колючий груша 5 индийский фи́га 5 индейский фи́га 5 цабр 5 сабр 5 индейский смоква 5	опунция индийский[1][2 3 плод 3 растение 3		1
kilt_de.tsv.ont.lemma:	Kilt 4	Wickelrock 3 Manner 3 Rock 3 Schottenrock 3	Knie 2 Trager 2 Wolle 2	1
kilt_en.tsv.ont.lemma:	garment 7	kilt 6	cloth 3	2
kilt_ru.tsv.ont.lemma:	килт 4	килт 3 горец 3	ткань 2 плечо 2 талиа 2 сумочка 2 одежда 2	1
melone_de.tsv.ont.lemma:	Zuckermelone 5	Melone 3 Beerenfruchte 3 Kurbisgewachs 3	Art 2 Pflanzenart 2 Gattung 2 Gurke 2 Familie 2	1
melone_en.tsv.ont.lemma:	Muskmelon 7 muskmelon 7	pero 5	Cucumis melo 4 specie 4 honeydew 4	1
melone_ru.tsv.ont.lemma:	дыня 5	растение 3	семейство 2 род 2 тыкваина 2 окраска 2 Плод 2 масса 2 форма 2	1
ohr_de.tsv.ont.lemma:	Horsystem 7	Nervensystem 3 Saugetier 3 Schall 3	Sinnesorgan 2 Ohr 2 Gleichgewichtsorgan 2 Verarbeitungsstation 2 Umschalt 2 Organ 2	3
ohr_en.tsv.ont.lemma:	ear 10	human 7	Vertebrates 5 mammal 5 organ 5	1
ohr_ru.tsv.ont.lemma:	позвоночный 10	человек 9 ухо 9	орган 8 млекопитающее 8	2
orange_de.tsv.ont.lemma:	Orange 19	Apfelsine 15 appelsina 15	Frucht 10 Baum 10	1
orange_en.tsv.ont.lemma:	orange 14	fruit 9	sweet orange 9	1
orange_ru.tsv.ont.lemma:	апельсин 6 апельсиновый дерево 6	плод 4	мандарин 3 помещать 3	1
physalis_de.tsv.ont.lemma:	Physalis 29	Judenkirsche 21 Blaskenkirsche 21	Lampionblume 20	1
physalis_en.tsv.ont.lemma:	Physalis 7	fruit 6	Physalis species 4	1
physalis_ru.tsv.ont.lemma:	растение 8 физалис 8	оболочка-чехлик 3 чашечка 3	ягода 2 клюква 2 чашелистик 2 вишня 2 семейство 2 стебель 2	1

File name	Most frequent words	Second most frequent words	Third most frequent words	Column no of topic
weste_de.tsv.ont.lemma:	Weste 12	Anzug 9	Westen 7	1
weste_en.tsv.ont.lemma:	upper-body garment 6	wear 4	vest 3 waistcoat 3	3
weste_ru.tsv.ont.lemma:	одежда 4 «тройка» 4 пиджак 4 костюм 4	жилет 3		2
wirbelsaeule_de.tsv.ont.lemma:	Wirbelsaule 9	Wirbel 3	Mensch 2 Skelett 2 Kreuz 2 Steißbein 2 Wirbeltiere 2 Wirbelkanal 2 Korper 2	1
wirbelsaeule_en.tsv.ont.lemma:	vertebral column 9	spine 7 backbone 7	spinal canal 5 bone 5 vertebrate 5 vertebra 5	1
wirbelsaeule_ru.tsv.ont.lemma:	позвоночник 10	позвонок 8	позвоночный столб 6	1
zunge_de.tsv.ont.lemma:	Wirbeltier 5 Mensch 5	Zunge 4	Muskelkorper 3	2
zunge_en.tsv.ont.lemma:	tongue 11	vertebrate 4 mouth 4 human 4	teeth 3	1
zunge_ru.tsv.ont.lemma:	Язык 2 язык 2			1
catsuit_de.tsv.ont.lemma:	Kleidungsstück 4 Korper 4	Gesicht 3 Kopf 3 Trager 3	Sportbekleidung 2 Catsuit 2	3
catsuit_en.tsv.ont.lemma:	material 8	leg 2 arm 2 torso 2 catsuit 2		2
catsuit_ru.tsv.ont.lemma:	тело[1 7	ткань 6 кэtságют 6	капюшон 5 маска 5 комбинезон 5	2

Table A.18 Detailed analysis of entities with the highest number of relations in the encycloaedic subset

File name	Most frequent words	Second most frequent words	Third most frequent words	Column no of topic
apfel_de.tsv:	Arten 2			-
apfel_en.tsv:	genus 2			-
apfel_ru.tsv:	яблони 2			1
auge_de.tsv:	Augen 3	Anforderungen 2 Tieren 2 Wahrnehmung 2 Leistungsfähigkeit 2 Qualität 2 Sehen 2		1
auge_en.tsv:	eyes 3 light 3 signals 3	eye 2 image 2 optical system 2 Eyes 2 brain 2 vision 2		1
auge_ru.tsv:	животных 2	Глаз 2		2
boxershorts_de.tsv:	Hosen 2			-
boxershorts_en.tsv:	boxers 4	type 2 freedom 2 shorts 2		1
boxershorts_ru.tsv:				-
brust_de.tsv:	Brust 7	Rumpfes 2 pectoralis 2	Brustkorb 2	1
brust_en.tsv:	thorax 3	chest 2		1
brust_ru.tsv:	Грудная клетка 3	отверстие 2		1
catsuit_de.tsv:	Catsuit 2			1
catsuit_en.tsv:				-
catsuit_ru.tsv:	воротника 2 пах 2 капюшоном 2 бегунками 2			-

File name	Most frequent words	Second most frequent words	Third most frequent words	Column no of topic
clementine_de.tsv:	Clementine 2 Citrus 2 Mandarine 2 Baum 2			1
clementine_en.tsv:	clementine 5	Clementines 4	fruit 3 California 3	1
clementine_ru.tsv:	плод 3	октября 2 Алжир 2 клементин 2 поставщики 2 году 2 мандарина 2 Испания 2 семян 2		1
durian_de.tsv:				-
durian_en.tsv:	species 3 fruit 3 zibethinus 3 durian 3	odour 2 cultivars 2		1
durian_ru.tsv:	Азии 3 Малайзии 3	странах 2		-
finger_de.tsv:	Finger 3 Daumen 3	Menschen 2 Fingern 2 Phalangen 2		1
finger_en.tsv:	humans 2 finger 2 thumb 2 digit 2			1
finger_ru.tsv:	Пальцы 2 конечностей 2			1
haar_de.tsv:	Haare 3	Haut 2		1
haar_en.tsv:	hair 3	Hair 2 follicles 2 skin 2		1
haar_ru.tsv:				1
kaktusfeige_de.tsv:	ficus-indica 2			1
kaktusfeige_en.tsv:				1
kaktusfeige_ru.tsv:				1
kilt_de.tsv:	Kilt 2 Männern 2			1
kilt_en.tsv:	kilt 2 century 2			1
kilt_ru.tsv:	Килт 4	килт 3	часть 2 килты 2 ткани 2 время 2	1
melone_de.tsv:	Zuckermelone 3	Formen 2 Gurke 2		1
melone_en.tsv:	Muskmelon 2 species 2 center 2 cultivars 2 varieties 2			1
melone_ru.tsv:	дыни 2 Азия 2			1
ohr_de.tsv:	Ohr 2			1
ohr_en.tsv:	ear 4	organ 2		1
ohr_ru.tsv:	человека 2 позвоночных 2 ухо 2 колебаний 2			1
orange_de.tsv:	Orange 3	Jahrhundert 2 Citrus 2 Bitterorange 2		1
orange_en.tsv:	fruit 5	sinensis 2 orange 2 sweet orange 2		2
orange_ru.tsv:				1
physalis_de.tsv:	Physalis 2 Gattung 2 Arten 2			1
physalis_en.tsv:	species 3	fruit 3	genus 2	-

File name	Most frequent words	Second most frequent words	Third most frequent words	Column no of topic
physalis_ru.tsv:	растения 2	Физалисы 2	name 2	2
weste_de.tsv:	Weste 7	Anzug 2	Kleidungsstück 2	1
weste_en.tsv:				-
weste_ru.tsv:				-
wirbelsaeule_de.tsv:	Wirbelsäule 4	Wirbeln 2		1
wirbelsaeule_en.tsv:				-
wirbelsaeule_ru.tsv:	позвонков 5	позвоночник 3		2
zunge_de.tsv:	Zunge 2			1
zunge_en.tsv:	tongue 5			1
zunge_ru.tsv:	языка 2			1
hose_de.tsv:	Hosen 4			1
hose_en.tsv:	trousers 7	legs 3 shorts 3 world 3	waist 2 Shorts 2 Trousers 2 pants 2 form 2 clothing 2 UK 2	1
hose_ru.tsv:				-
hut_de.tsv:	Hut 4	Schutz 2 Trägers 2 Rand 2 Kopfbedeckung 2		1
hut_en.tsv:	hats 5			1
hut_ru.tsv:	Шляпа 2			1

Table A.19 Detailed analysis of frequent nouns in the encyclopaedic subset

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5
bovary2_de.t sv.ont.lemm a:	Großmutter 4	Kind 3 Vater 3 Tante 3		1,2				
bovary2_en.t sv.ont.lemm a:	grandmother 2			1				
bovary2_ru.t sv.ont.lemm a:	дедушка 2 тетка 2 бабушка 2			1				
bovary3_de.t sv.ont.lemm a:	Gesicht 4	Mann 3 Furcht 3 Feuerwehrhauptm ann 3	Hauser 2 Knie 2	2	1,3	3	2	
bovary3_en.t sv.ont.lemm a:	man 4	anxiety 3 pleasure 3	fright 2	1			2,3	
bovary3_ru.t sv.ont.lemm a:	человек 6	капитан 5	страх 4	1,2			3	
bovary_de.ts v.ont.lemma:	Tuchrock 5	Stiefel 4	Hose 4		1,2,3			
bovary_en.ts v.ont.lemma:	leg 6 school jacket 6	forehead 4 hair 4 country lad 4 wrist 4 trouser 4	stocking 3 class 3 fellow 3 boot 3	2,3	1,2,3	1		
bovary_ru.ts v.ont.lemma:	пиджачок 7	башмак 5 пantalоны 5	перчатка 4 чулок 4		1,2,3			
chekhov2_d e.tsv.ont.lem ma:	Schulter 2 Hand 2 Tischler 2 Gesicht 2 Loge 2 Galerie 2 Rang 2			1	1			
chekhov2_e n.tsv.ont.lem ma:	man 3	box 2 face 2 gallery 2 shoulder 2 hand 2		1	2	2		1,2

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5
		tier 2						
chekhov2_ru.tsv.ont.lemma:								
chekhov_de.tsv.ont.lemma:	Hund 8	Dorfkoter 5 Tischler 5 Dachs 5 Glied 5	Auge 3 Pfote 3	1,2	2,3			
chekhov_en.tsv.ont.lemma:	dog 9	face 6	yard - dog 4 dachshund 4	1,3	2			
chekhov_ru.tsv.ont.lemma:	собака 9	ухо 6	дворняжка 5 такса 5	1,3	2			
france2_de.tsv.ont.lemma:	Fenster 3	Treppe 2 Haustur 2				1,2		
france2_en.tsv.ont.lemma:	hand 5	lover 4 hair 4 father 4 arm 4 face 4 head 4 citoyenne_ 4 sweetheart 4 concierge 4		2	1,2			
france2_ru.tsv.ont.lemma:	рука 2				1			
france_de.tsv.ont.lemma:	Burger 16	Tischler 7	Versammlung 6 Mitglied 6 Überwachungsausschul 6	1,2,3				3
france_en.tsv.ont.lemma:	citoyen_ 6 member 6	man 5 hand 5 woman 5 church 5 speaker 5 child 5 meeting 5	gathering 4 assembly 4	1,2	2	2		2,3
france_ru.tsv.ont.lemma:	Церковь 6	гражданин 4 комитет 4	член 3 секция 3 кафедра 3 клирик 3 собрание 3 фасад 3 подпись 3	2,3		1,3		2,3
gorki2_de.tsv.ont.lemma:	Leiden 4	Frieden 2 Friede 2					1,2	
gorki2_en.tsv.ont.lemma:								
gorki2_ru.tsv.ont.lemma:								
gorki_de.tsv.ont.lemma:	Nordwind 3 Welle 3 Wind 3	Verkaufsstande 2 Stadt 2 Stoß 2 Leute 2 Fenster 2 Natur 2 Fluß 2 Fluss 2 Bude 2		2		2		1,2
gorki_en.tsv.ont.lemma:	wavelet 3 tavern 3 man 3	rain 2 town 2 booth 2 folk 2 body 2 shop 2 wind 2		1	2	1,2		1,2
gorki_ru.tsv.	человек 2			1		1		1

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5
ont.lemma:	город 2 ларь 2 волна 2							
idiot2_de.tsv .ont.lemma:	Furst 12	Leute 9	Hand 8 Kopf 8 Herz 8	1,2	3			
idiot2_en.tsv .ont.lemma:	prince 13	hair 8	cheek 6	1	2,3			
idiot2_ru.tsv .ont.lemma:	человек 10	рука 9 князь 9	сердце 7	1	2,3			
idiot_de.tsv .ont.lemma:	Gesicht 27	Mensch 11 Leute 11	Passagier 10	2,3	1			
idiot_en.tsv .ont.lemma:	face 16	fellow 7	man 5 cloak 5 person 5	2,3	1,3			
idiot_ru.tsv.o nt.lemma:	человек 14	лицо 13	физиономия 10	1	2,3			
kipling2_de.t sv.ont.lemm a:	Mensch 10	Elefant 9	Leute 8	1,2,3				
kipling2_en.t sv.ont.lemm a:	man 7 elephant 7 men 7	bull elephant 6 tusker 6 feast 6	driver 4 blood 4	1,2,3	2,3			2
kipling2_ru.t sv.ont.lemm a:	человек 31	слон 9	мальчик 5	1,2,3				
kipling_de.ts v.ont.lemma:	Wolf 6	Tier 5	Kind 4	1,2,3				
kipling_en.ts v.ont.lemma:	teeth 4 paw 4	creature 3 tail 3 child 3 nose 3	buck 2 men 2 meat 2	1	1,2			
kipling_ru.ts v.ont.lemma:	пещера 4 шакал 4	гора 3	ребенок 2 водобоязнь 2 безумие 2 волк 2 болезнь 2	1,3		1,2	3	3
krieg2_de.ts v.ont.lemma:	Onkel 4	Mensch 3 Vater 3		1,2				
krieg2_en.ts v.ont.lemma:	father 6 ncle 6	bosom 4 hand 4	eye 2 boy 2 man 2 nose 2 Prince 2	1,3	2,3			
krieg2_ru.tsv .ont.lemma:	человек 5	дядя 3	отец 2	1,2,3				
krieg_de.tsv. ont.lemma:	Furst 9	Hofdame 3 Uniform 3 Freund 3 Sklave 3 Majestat 3 Kaiserin 3 Antichrist 3	Schnallenschuh 2 Strumpf 2	1,2	2,3			
krieg_en.tsv. ont.lemma:	man 10	Prince 8	prince 5 grandfather 5	1,2,3				
krieg_ru.tsv. ont.lemma:	князь 5	башмак 3 лакей 3 звезда 3 императрица 3 мундир 3 чулок 3 фрейлина 3	человек 2 дед 2 рука 2	1,2,3	2,3			
sandmann2_ de.tsv.ont.le mma:	Stadt 9	Hand 7	Auge 5		2,3	1		
sandmann2_ en.tsv.ont.le mma:	eye 13	hand 12	person 10	3	1,2			
sandmann2_ ru.tsv.ont.le mma:	город 7	башня 4	рука 4 галерея 4		2	1,2,3		

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5
sandmann3_de.tsv.ont.lemma:	Mensch 9 Arm 9	Kind 4 Advokat 4 Vater 4	Mutter 3 Familie 3	1,2,3	1			
sandmann3_en.tsv.ont.lemma:	man 8 father 8 arm 8	advocate 6 family--as 6	child 5	1,2,3	1			
sandmann3_ru.tsv.ont.lemma:	отец 7	ребенок 6 матушка 6	колдун 5 песочный человек 5 сосед-аптекарь 5	1,2,3				
sandmann_de.tsv.ont.lemma:	Haus 16	Zimmer 14	Mutter 13	3		1,2		
sandmann_en.tsv.ont.lemma:	mother 10	child 8	father 7 house 7	1,2,3		3		
sandmann_ru.tsv.ont.lemma:	дверь 10 комната 10	отец 9	комнатка 7 лицо 7 дом 7 парк 7	2	3	1,3		
wells2_de.tsv.ont.lemma:	Kind 5	Fleischerjunge 3 Artillerist 3 Gefuhl 3 Fahrzeug 3 Arbeiter 3	Zweifel 2 Mensch 2 Stadt 2 Leute 2 Hauser 2 Haus 2 Straß 2	1,2,3		3	2,3	2
wells2_en.tsv.ont.lemma:	child 7	humanity 5	workman 4 artilleryman 4 visitor 4	1,3				2
wells2_ru.tsv.ont.lemma:	человек 7	дом 5	улица 4	1		2,3		
wells_de.tsv.ont.lemma:	Mensch 11	Planet 7 Stern 7	Mann 6 Zone 6 Wesen 6 Region 6 Lebewesen 6	1,3		2,3		
wells_en.tsv.ont.lemma:	planet 14	world 8	inhabitant 6 morning star 6 water 6	1		1,2,3		3
wells_ru.tsv.ont.lemma:	человек 8	существо 6 планета 6	вода 5	1,2		2		3

Table A.20 Detailed analysis of entities with the highest number of relations in the literary subset (1) is person/character; 2) is description of person/character; 3) is description of location; 4) is feeling/condition; 5) is other)

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5	6
bovary2_de.tsv:									
bovary2_en.tsv:	Bovary 2			1					
bovary2_ru.tsv:	Бовари 2			1					
bovary3_de.tsv:	Tonne 3	Binet 2 Furcht 2 Liebe 2 Wildenten 2		2				2	1,2
bovary3_en.tsv:	tub 3	Emma 2		1					
bovary3_ru.tsv:	Родольфа 2	Эмма 2 бочки 2 уток 2 Бине 2		1,2					2
bovary_de.tsv:	Rektor 2 Neuling 2				1				
bovary_en.tsv:	work 2 legs 2 fellow 2 head-master 2				1	1			1

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5	6
bovary_r u.tsv:	Новичок 2 уроки 2 воспитателю 2 ногу 2				1	1			1
chekhov 2_de.tsv:	Kashtanka 5	Wand 2 Hand 2 Menschen 2 Vergleich 2		1		2	2		2
chekhov 2_en.tsv:	Kashtanka 3	hand 2 Auntie 2 wall 2		1,2		2			
chekhov 2_ru.tsv:	Каштанка 3	Тетка 2 руки 2		1,2		2			
chekhov _de.tsv:	Kashtanka 3	Trottoir 2 Kreuzung 2 Dachs 2 Alexandritsch 2 Tag 2 Dorfkötter 2		1,2	2		2		2
chekhov _en.tsv:	pavement 2 fox 2 side 2 carpenter 2 mongrel 2 face 2 Alexandritch 2 time 2 Kashtanka 2 day 2 way 2			1	1	1	1		1
chekhov _ru.tsv:	такса 2 Каштанка 2 дворяжкой 2			1	1				
france2_ de.tsv:	Tränen 2								1
france2_ en.tsv:	Good-bye 2 tears 2								1
france2_ ru.tsv:	слезы 2								1
france_d e.tsv:	Gamelin 5	Tod 2 Mütze 2 Überwachungs- schuß 2 »Ich 2 Uhr 2 Versammlungen 2 Bezirks 2 Petition 2 Tischler 2 Kanzel 2		1	1	1	1		1
france_e n.tsv:	Gamelin 5	Section 3	petition 2 Committee 2 Surveillance 2 morning 2 meeting 2 desk 2 church 2 pulpit 2 nave 2	1					
france_r u.tsv:	Гамлен 5	секции 4	петиции 2 площади 2 собранья 2	1			2,3		3
gorki2_d e.tsv:	Seele 2								1
gorki2_e n.tsv:	dawn 2	soul 2							1,2
gorki2_r u.tsv:									
gorki_de	- 3	Seele 2					2	2	1,2

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5	6
.tsv:		Tagen 2 Satten 2 Hunger 2 Stadt 2 Gebäuden 2							
gorki_en .tsv:	mind 3	hunger 2 town 2 rain 2 man 2 days 2 buildings 2 quarter 2 night 2 river 2			2		2	2	1,2
gorki_ru .tsv:	человека 2	положение 2 ветер 2 души 2				1			2
idiot2_de .tsv:	Fürst 7	Rogoshin 6	Leutnant 3 Wangen 3 Offizier 3	2	1,3	3			
idiot2_en .tsv:	prince 9	Rogojin 7	hand 3 heart 3 face 3	2	1		3		
idiot2_ru .tsv:	Рогожин 6	Князь 4	князь 3 кадет 3	1	2,3				
idiot_de.t sv:	Lächeln 3 Wagen 3 Klasse 3 Blick 3	Petersburg 2 Kapuze 2 Anschein 2 Pelz 2 Teil 2 Gesicht 2 Leute 2 Bahn 2 Tuch 2 Ausdruck 2 Haar 2 Nacht 2 Ausland 2 Mannes 2 Augen 2		2	2	1,2	1,2		2
idiot_en.t sv:	eyes 3 expression 3	look 2 appearance 2 passengers 2 morning 2 persons 2 fellow 2 carriages 2 sort 2 neighbour 2 face 2 train 2 moment 2 night 2 day 2 cloak 2			1	1,2	2		2
idiot_ru.t sv:	что-то 3	класса 2 роста 2 всё 2 лет 2 человека 2 дороги 2 лица 2 капюшоном 2 сосед 2 Италии 2		1	1	2	2		1,2
kipling2_de.tsv:	Elefanten 11	Toomai 9	Appa 7	2,3	1				
kipling2_en.tsv:	Toomai 8	Appa 6	elephants 5	1,2	2				

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5	6
kipling2_ru.tsv:	Тумаи 7	слонов 6	Анна 5	1,3	2				
kipling_de.tsv:	Wolf 5	Tabaqui 4	Knochen 3 Höhle 3	1,2		3	4		
kipling_en.tsv:	Tabaqui 5	Wolf 4	children 3	1,2	3				
kipling_ru.tsv:	Табаки 3 Волк 3	дети 2 ума 2 джунглях 2 пещеры 2 кость 2		1	2		2		2
krieg2_de.tsv:	Nikolai 8	Peter 7 Vater 7	Desalles 3 Fäden 3	1,3	2				3
krieg2_en.tsv:	Nicholas 8	father 6 Pierre 6	Dessalles 4	1,2,3	2				
krieg2_ru.tsv:	отец 8	Пьер 7	Николенька 6	2,3	1				
krieg_de.tsv:	Fürst 3	Französisch 2 Abendgesellschaft 2 Antichrist 2		2	1				2
krieg_en.tsv:	grippe 2 nothing 2 man 2 Antichrist--I 2 Scherer 2 importance 2 reception 2 Pavlovna 2 Prince 2			1	1				1
krieg_ru.tsv:	князь 3	Шереп 2 грипп 2		2	1				2
sandmann2_de.tsv:	Clara 15 – 15	Nathanael 14	Lothar 7	1,2,3					1
sandmann2_en.tsv:	Clara 16	Nathaniel 15	Lothaire 7	1,2,3					
sandmann2_ru.tsv:	Натанаэль 11	Клара 11	Лотар 7	1,2,3					
sandmann3_de.tsv:	Macht 5 Coppelius 5	Sandmann 4 Gemüt 4	Nathanael 3 Innern 3	1,3	2	3		2	1
sandmann3_en.tsv:	mind 6	Coppelius 5 power 5	world 3 father 3 sandman 3 children 3	2	3		3		1,2
sandmann3_ru.tsv:	душу 3 сила 3 Натанаэль 3 Коппелиус 3	образы 2 слов 2 дух 2 мира 2 отца 2 Копполу 2 адвоката 2 детей 2		1,3	2,3				1,2
sandmann_de.tsv:	Sandmann 21	Mutter 12	– 11		1,2				2
sandmann_en.tsv:	Sandman 18	mother 9 father 9	room 8		1,2		3		
sandmann_ru.tsv:	человек 13	отца 7	человека 6		1,2,3				
wells2_de.tsv:	Leben 3	Seele 2 Zeit 2 Menschen 2 Kinder 2 Marsleute 2			2		2		1,2

File name	Most frequent words	Second most frequent words	Third most frequent words	1	2	3	4	5	6
		Straßen 2							
wells_en.tsv:	life 4	mind 3	hand 2 space 2 Martians 2 streets 2 men 2 time 2 night 2 day 2 children 2		3	3	3		1,2,3
wells_ru.tsv:	марсиан 3 людей 3	планеты 2 жизнь 2			1		2		2
wells_de.tsv:	Erde 4 Leben 4	Mars 3 Abkühlung 3 Menschen 3 Sonne 3 Stern 3 Oberfläche 3 Theil 3	Gedanken 2 Wesen 2 Jahrhunderts 2 Planeten 2 Luft 2 Entfernung 2 Lebens 2 Menschheit 2 thun 2 Geschlecht 2 Wasser 2 Marsbewohner 2 Jahren 2 Lebewesen 2 Meilen 2	1,2	2,3		2,3		1,2,3
wells_en.tsv:	men 9	life 6	Mars 5 world 5 earth 5	3	1		3		2
wells_ru.tsv:	Мирсе 8	жизнь 7	планету 3 воды 3 Земли 3 люди 3	1,3	3		3		2,3

Table A.21 Detailed analysis of frequent nouns in the literary subset (1) is named entity; 2) is person/character; 3) is description of person/character; 4) is description of location; 5) is feeling/condition; 6) is other)

A.1. Annotator tests

A.1.1. English version

Annotator's recruitment test

Name: _____

1. Parts of speech

Please mark the nouns or if present the noun compounds in the following text.

The orange (specifically, the sweet orange) is the [fruit](#) of the [citrus](#) species *Citrus x sinensis* in the [family Rutaceae](#).^[2] The fruit of the *Citrus sinensis* is considered a sweet orange, whereas the fruit of the *Citrus aurantium* is considered a [bitter orange](#). The orange is a [hybrid](#), possibly between [pomelo](#) (*Citrus maxima*) and [mandarin](#) (*Citrus reticulata*), which has been cultivated since ancient times.^[3]

As of 1987, orange trees were found to be the most [cultivated](#) fruit tree in the world.^[4] Orange trees are widely grown in tropical and subtropical climates for their sweet fruit. The fruit of the orange tree can be eaten fresh, or processed for its juice or fragrant peel.^[5] As of 2012, sweet oranges accounted for approximately 70% of citrus production.^[6] In 2010, 68.3 million metric tons of oranges were grown worldwide, production being particularly prevalent in [Brazil](#) and the U.S. states of [California](#)^[7] and [Florida](#).^[8]

2. Semantic terms

Definitions:

Synonyms are different words with the same meaning. Words that are synonyms are said to be synonymous, and the state of being a synonym is called synonymy.

Holonymy defines the relationship between a term denoting the whole (holonym) and a term denoting a part (meronym) of, or a member (meronym) of, the whole. Holonymy is also described as "part-of" relation.

A hyponym is a word or phrase whose semantic field is included within that of another word, its hypernym. Hyponymy is also described as "kind-of"-relation.

Please give at least two examples for each of the described semantic relations (synonymy, holonymy and hyponymy). Please also mark holonyms, meronyms, hypo- and hypernyms in the examples of holonymy and hyponymy.

Synonymy: _____

Holonymy: _____

Hyponymy: _____

3. Annotation of semantic relations

Please mark the relations described in task 2 between nouns and noun compounds in the text that is presented in task 1.

A.1.2. German version

Annotator's recruitment test

Name: _____

Please follow the instructions written in English and answer in German.

1. Parts of speech

Please mark the nouns or if present the noun compounds in the following text.

Die Orange (Ausssprache: [o'ʋanʒə] oder [o'ʋɑ:ʒə]), nördlich der Speyerer Linie auch Apfelsine (von niederdeutsch appelsina, wörtlich „Apfel aus China/Sina“) genannt, ist ein immergrüner Baum, im Speziellen wird auch dessen Frucht so genannt.^[1] Der gültige botanische Name der Orange ist Citrus × sinensis L., damit gehört sie zur Gattung der Zitruspflanzen (Citrus) in der Familie der Rautengewächse (Rutaceae). Sie stammt aus China oder Südostasien, wo sie aus einer Kreuzung von Mandarine (Citrus reticulata) und Pampelmuse (Citrus maxima) entstanden ist.^[2]

2. Semantic terms

Definitions:

Synonyms are different words with the same meaning. Words that are synonyms are said to be synonymous, and the state of being a synonym is called synonymy.

Holonymy defines the relationship between a term denoting the whole (holonym) and a term denoting a part (meronym) of, or a member (meronym) of, the whole. Holonymy is also described as “part-of” relation.

A hyponym is a word or phrase whose semantic field is included within that of another word, its hypernym. Hyponymy is also described as “kind-of”-relation.

Please give at least two examples for each of the described semantic relations (synonymy, holonymy and hyponymy). Please also mark holonyms, meronyms, hypo- and hypernyms in the examples of holonymy and hyponymy.

Synonymy: _____

Holonymy: _____

Hyponymy: _____

3. Annotation of semantic relations

Please mark the relations described in task 2 between nouns and noun compounds in the text that is presented in task 1.

A.1.3. Russian version

Annotator's recruitment test

Name: _____

Please follow the instructions written in English and answer in Russian.

1. Parts of speech

Please mark the nouns or if present the noun compounds in the following text.

Апельси́н — плод апельсинового дерева (*Citrus sinensis*), которое представляет собой^[2] гибрид мандарина (*Citrus reticulata*) и помело (*Citrus maxima*) и культивировалось в Китае ещё за 2,5 тысячи лет до н. э.

В Европу дерево было привезено португальскими мореплавателями. После этого быстро распространилась мода на выращивание апельсиновых деревьев; для этого стали строить специальные стеклянные сооружения, названные оранжереями (от фр. *orange* 'апельсин'). Теперь апельсиновые деревья растут по всему побережью Средиземного моря (а также — в Центральной Америке)^[3].

Слово «апельсин» заимствовано в русский язык из голландского (нидерландского) языка; нидерл. *appelsien* (ныне чаще употребляется форма *sinaasappel*), равно как и нем. *Apfelsine*, есть калька с фр. *potme de Chine* (буквально — «яблоко из Китая»; теперь это название во французском вытеснено словом *orange*)^[4].

2. Semantic terms

Definitions:

Synonyms are different words with the same meaning. Words that are synonyms are said to be synonymous, and the state of being a synonym is called synonymy.

Holonymy defines the relationship between a term denoting the whole (holonym) and a term denoting a part (meronym) of, or a member (meronym) of, the whole. Holonymy is also described as "part-of" relation.

A hyponym is a word or phrase whose semantic field is included within that of another word, its hypernym. Hyponymy is also described as "kind-of"-relation.

Please give at least two examples for each of the described semantic relations (synonymy, holonymy and hyponymy). Please also mark holonyms, meronyms, hypo- and hypernyms in the examples of holonymy and hyponymy.

Synonymy: _____

Holonymy: _____

Hyponymy: _____

3. Annotation of semantic relations

Please mark the relations described in task 2 between nouns and noun compounds in the text that is presented in task 1.

A.2. Guidelines

Guidelines for the Annotation of Classical Semantic Relations between Nominals 1.0

Table of Content

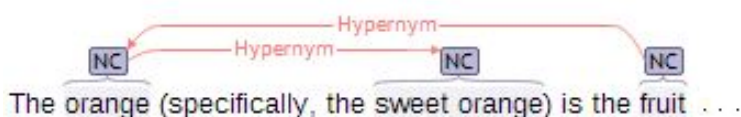
Introduction	119
Noun Compounds	119
1 Bidirectional relations	120
1.1 Synonyms	120
1.2 Co-Hyponyms	122
2 Uni-directional relations	124
2.1 Hypernyms	124
2.2 Holonyms	126
3 General Rules	129
	131
References	

Introduction

These guidelines describe the rules for the annotation of classical semantic relations between nominals in paragraphs of texts in English, German and Russian. Only relations that are both present in the context, but also applicable in natural language, are to be annotated.

Beside the creation of a knowledge base created on relations found in context, the results of this annotation task will be used for the analysis of the distribution of the further mentioned relations throughout different text genres and languages. Furthermore, terms with a prominent number of relations will be examined, especially regarding their context. Moreover, the dataset may be enlarged and used for the creation of an algorithm that automatically classifies classical semantic relations between nominals. The automatic classification of these relations may be useful in tasks such as information extraction, information retrieval, text summarization, machine translation, question answering, paraphrasing, recognizing textual entailment, thesaurus – and semantic network construction, word-sense disambiguation, and language modelling (Nastase et al., 2013).

In particular, the rules for the annotation of classical semantic relations such as synonyms, hypernyms, hyponyms, co-hyponyms, holonyms, and meronyms will be further described here. Although these guidelines are written in English only, examples will be provided in all three languages, but will not necessarily be translations of the same sentence. An introductory example of two hypernym relations is presented below:



Introductory example (“Orange”, 2014a, para. 1)

The example sentences or paragraphs may include more relations, but only those of interest for the specific task are marked for the purpose of focus.

The annotation will be performed with the annotation tool WebAnno (Yimam, 2014). For further instructions on the use of the tool, please consult the following wiki: <https://code.google.com/p/webanno/wiki/Annotation>.

The nominals will be pre-annotated using the TreeTagger (Schmid, 1995). But if annotators will find any unannotated nouns or noun compounds, they are asked to mark them, if they are in a relation which is important for this project. According to Quirk et al., a nominal usually refers to a phrase which behaves syntactically like a noun or a noun phrase (Quirk et al., 1985, p. 335). According to these guidelines, Named Entities are not annotated, as they are instances of classes rather than parts of relations. In case of doubt of whether a nominal is a Named Entity, consult the extended NoSta-D Guidelines: http://www.lrec-conf.org/proceedings/lrec2014/pdf/276_Paper.pdf (Benikova et al., 2014).

Noun Compounds

Compounds in general are an unsolved problem in linguistics. Though it is one of the most productive word formation processes for both English and German, there is no clear answer on how to systematically find this type. In these guidelines some restrictions on how to find these will be given in order to make the annotations reproducible.

The following definition is suitable for all annotated languages in these guidelines: “[...] is a word that consists of two elements, the first of which is either a root, a word or a phrase, the second of which is either a root or a word.” (Plag, 135) In the case of nominal compounds, the root⁵² of the compound has to be a noun.

In German, only lexicalized noun compounds will be annotated.

If the noun compound is exocentric, which means that the semantics of the compound are outside of the combination of the two elements separately, the compound can be identified for that reason. An example of an exocentric noun compound is *blue helmets*, not referring to *a kind of helmet*, but to *UN peacekeepers*. If the noun compound is endocentric, which means that the elements do not have an extrinsic semantic meaning, it is more complex to identify. An example of an endocentric noun compound is *sweet orange* in the introductory example, which is a kind of orange. In these guidelines, all noun phrases exclusively consisting of nouns that are not in the genitive will be considered noun compounds in the English subset. For example, *garbage can* and *cotton shirt* are noun compounds according to this rule, but *king’s will* is not.

Newly found noun compounds are annotated with the tag “NC”, without deleting already annotated nouns. Only noun compounds that are in a relation that is described in these guidelines are annotated.

1. Bidirectional relations

Some special rules have to be defined in order to prevent redundancy in bidirectional relations.

The relation is always annotated from the left word to the right word (see all examples in this chapter).


1.1. Synonyms

Synonyms are different words with the same meaning, e.g.




A handbag, also purse or pouch in American English, is a handled medium-to-large bag . . .

Example 1.1.1 (“Handbag”, 2014, para. 1)



Die Orange (Aussprache: [oˈraŋʒə] oder [oˈrãːʒə]), nördlich der Speyerer Linie auch Apfelsine (von

Example 1.1.2 (“Orange”, 2014, para. 1)




Брю́ки (нидерл.[] broek), или штаны́, — предмет верхней одежды, . . .

Example 1.1.3 (“Brjuki”, 2014, para. 1)

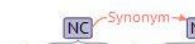
⁵² In the case of copulative compounds all the roots

Rules:


1. Different spellings are regarded as synonyms, as they are variations of the same sense unit in one language.


Color or colour (see spelling differences) is the visual perceptual property . . .

Example 1.1.4 (“Color”, 2014, para. 1)



Die Delfine oder Delphine (Delphinidae) gehören zu den Zahnwalen (Odontoceti) . . .

Example 1.1.5 (“Delfine”, 2014, para. 1)


Стрелиция, также Стрелитция (лат. __ Strelitzia) — типовой род семейства Стрелитциевые

Example 1.1.6 (“Strelizija”, 2014, para. 1)


2. Translations are not regarded as synonyms, unless they are used like regular words in the annotated language. An example of such use is built by the group of Latin or Greek terms for living things, which are used as synonyms for the language intern words in a biological setting⁵³, e.g.


Modern humans (Homo sapiens or Homo sapiens sapiens) are the only extant members of the hominin clade, . . .

Example 1.1.7 (“Human”, 2014, para. 1)


Der Mensch (Homo sapiens) ist innerhalb der biologischen Systematik ein höheres Säugetier aus der Ordnung der Primaten (Primates).

Example 1.1.8 (“Mensch”, 2014, para. 1)


Человек разумный (лат.[] Homo sapiens; в русскоязычных текстах встречается также написание Хомо Сапиенс[1] или Гомо Сапиенс[2]) —

Example 1.1.9 (“Chelovek razumnyj”, 2014, para. 1)

3. If there are other relations that are to be annotated with the synonyms, they are to be annotated with the nominal closest to the beginning of the paragraph.
See examples 1.1.1, 1.1.7, 1.1.9

⁵³ In Russian such occurrences are easier to identify, as terms written in Cyrillic are considered to be part of the language in these guidelines.

1.2. Co-Hyponyms

Co-hyponyms are only annotated if there is no appropriate hypernym in the paragraph. Only co-hyponyms with a clear, common, and semantically linked hypernym are annotated.

The common hypernym in the following examples is *family member*.

The temper of the father is so different from that of the son . . .

Example 1.2.1 (Moliere, 1668/2003)

Vater und Sohn sind in ihrer Gesinnung so gründlich verschieden, . . .

Example 1.2.2 (Moliere, 1668/1887)

Наконец, как только мне удастся -- на что я надеюсь -- найти отца и мать, . . .

Example 1.2.3 (Moliere, 1668/2009)

Co-hyponyms are combined assuming the highest possible sensible hypernym.

The nanny, the clerk, her father and mother as well as her friends came to congratulate her.

Example 1.2.4⁵⁴

Das Kindermädchen, die Sekretärin, ihr Vater und ihre Mutter, so wie ihre Freunde kamen um ihr zu gratulieren.

Example 1.2.5

Няня, секретарша, её отец и мать, так же как её друзья пришли поздравить её.

Example 1.2.6

⁵⁴ The highest possible mutual hypernym of *nanny*, *clerk*, *father*, *mother* and *friends* is *person*. Note that although *nanny* and *clerk* could be subclassified as *profession* or *mother* and *father* could be subclassified as *parent*, this is not done according to these guidelines.

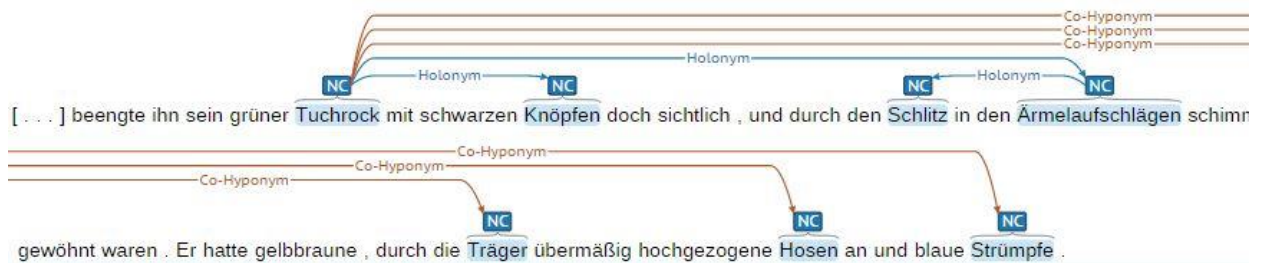
Rules:

If there are other annotations to be made from one of the co-hyponyms, the following rules apply:

1. If it is another co-hyponymic relation, it is to be annotated from the co-hyponym closest to the beginning of the paragraph.



Example 1.2.7 (Flaubert, 1856/2006)



Example 1.2.8 (Flaubert, 1856/1986)



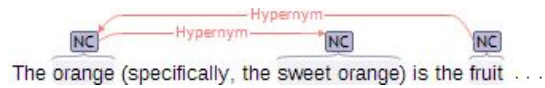
Example 1.2.9 (Flaubert, 1856/1956)

2. If it is any other relation, it is to be annotated with the related co-hyponym.
See examples 1.2.7-1.2.9.

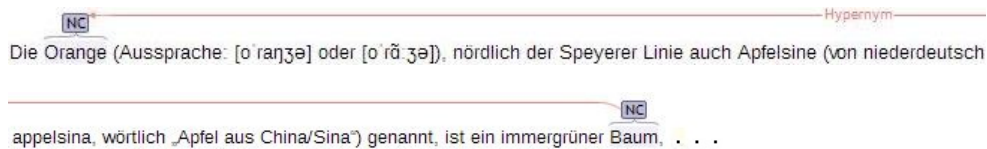
2. Uni-directional relations

2.1. Hypernyms

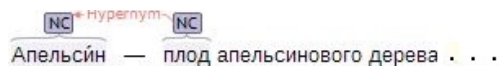
Hypernyms can be described as „kind-of“ relations. The relation is annotated from the hypernym (topic) to hyponym (minor term).



Example 2.1.1 (“Orange”, 2014a, para. 1)



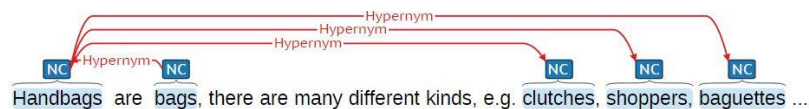
Example 2.1.2 (“Orange”, 2014b, para. 1)



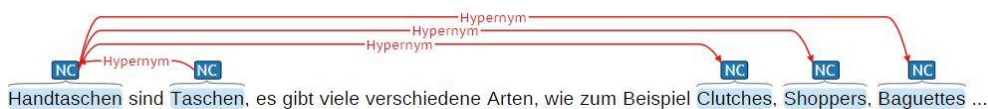
Example 2.1.3 (“Apel’sin”, 2014, para. 1)

Rules:

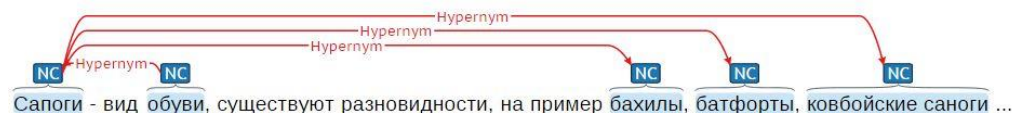
1. Too general hypernyms like *thing* are not annotated according to these guidelines. When several hypernyms are possible, the following rules apply: If they are in a hierarchy, the hierarchically lowest has to be chosen, even if other hypernyms are located closer in the text.



Example 2.1.4 ⁵⁵



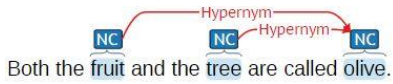
Example 2.1.5



Example 2.1.6

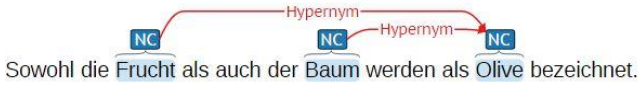
- 2.2. If they are not in a hierarchy, both nominals are marked as hypernyms.

⁵⁵ *Handbag*, not *bag* is annotated as hypernym for the mentioned kinds of purses. *Bag*, in turn is hypernym of *handbag*.



Both the **fruit** and the **tree** are called **olive**.

Example 2.1.7⁵⁶



Sowohl die **Frucht** als auch der **Baum** werden als **Olive** bezeichnet.

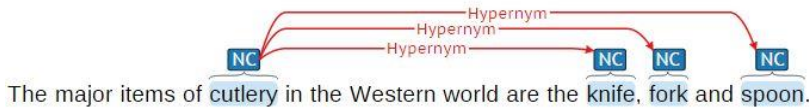
Example 2.1.8



Плод, так же как и **дерево**, называют **оливой**.

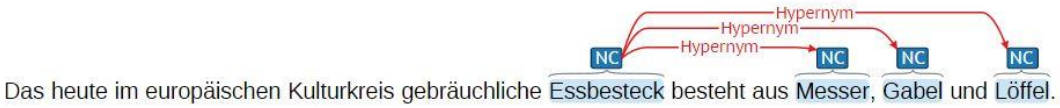
Example 2.1.9

3. Mass nouns may also be hypernyms



The major items of **cutlery** in the Western world are the **knife**, **fork** and **spoon**.

Example 2.1.10 ("Cutlery", 2014, para. 3)



Das heute im europäischen Kulturkreis gebräuchliche **Essbesteck** besteht aus **Messer**, **Gabel** und **Löffel**.

Example 2.1.11 ("Essbesteck", 2014, para. 1)



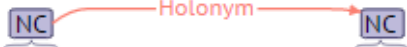
В современном европейском обществе, к самым обиходным **столовым приборам** относят **нож**, **вилку** и **ложку**.

Example 2.1.12 ("Essbesteck", 2014, para. 1 my translation)

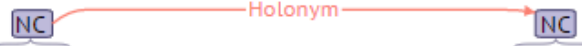
⁵⁶ As mentioned in the text, both *tree* and *fruit* are hypernyms of *olive*.

2.2. Holonyms

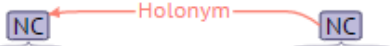
Holonymy can be described as „part-whole“-relation. The relation is annotated from the holonym (whole) to the meronym (part).

 is an evergreen tree with, in particular its fruit is called the same.

Example 2.2.1 (“Orange”, 2014b, para. 1, my translation)

 ist ein immergrüner Baum, im Speziellen wird auch dessen Frucht so genannt.

Example 2.2.2 (“Orange”, 2014b, para. 1)

 — плод апельсинового дерева

Example 2.2.3 (“Apel’sin”, 2014b, para. 1)

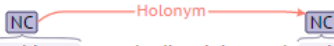
Rules:

If the following rules apply the relation shall be marked as a holonym relation:


1. A material or substance that some nominal is typically made of is considered a holonym.

 The predecessor of the suitcase is a travelling box made from vulcanized wood.

Example 2.2.4 (“Koffer”, 2014a, para. 1 my translation)

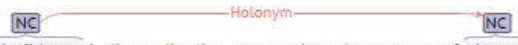
 Vorgänger des Koffers sind Reisekisten aus (vulkanisiertem) Holz.

Example 2.2.5 (“Koffer”, 2014a, para. 1)

 Предшественник чемодана - ящик для путешесвия из вулканизированной древесины.

Example 2.2.6 (“Koffer”, 2014a, para. 1 my translation)

2. Meronyms need to be parts and not random pieces of the holonym.
3. Meronyms need to be functional or physical parts of the holonym.
4. Typical components, as for example people having faces or trees having leaves are considered holonyms.
5. Nominals in a member-collection (Winston et al., 1987) relation are considered to be in a holonymical relation.

 A football team is the collective name given to a group of players selected together in the various team sports known as football .

Example 2.2.7 (“Football team”, 2014, para. 1)

Eine ^{NC} Fußballmannschaft besteht auf dem Platz aus einem Torwart und 10 Feldspielern (^{NC} 11 Spieler insgesamt) .

Example 2.2.8 (“Fußballmannschaft”, 2014, para. 1)

Каждая ^{NC} команда состоит максимум из одиннадцати ^{NC} игроков (без учета запасных) , один из которых должен быть вратарём .

Example 2.2.9 (“Futbol”, 2014, para. 1)

The following rules describe which relations are **not** considered holonymic:

1. If the sense of the meronym and holonym are of different count classes, countable and not countable, they are not considered to be in a holonymic relation

Here, “life” in general is meant, so it cannot be part of one single person:

The intellectual side of ^{NC} man already admits that ^{NC} life is an incessant struggle for existence[...] .

Example 2.2.10 (Wells, 1898)

Doch so eitel ist der ^{NC} Mensch [...], daß dort geistiges ^{NC} Leben [...] entstehen könnte.

Example 2.2.11 (Wells, 1898/1901: 6)

[...] как человек изучает в микроскопе кратковременную ^{NC} жизнь существ[...] .

Example 2.2.12 (Wells, 1898/1927)

2. Locations of nominals are not considered to be holonyms here, thus e.g. *glass* is not a holonym of *beer*, *pipe* is not a holonym for *tobacco*, *church* is not a holonym for *statue* and *car* not a holonym for *passenger*. However, if the potential meronym is a functional part of the potential holonym, but is also located in it, as e.g. *baseline* and *tennis court*, the relation is considered valid.
3. Has-property is not annotated.

Gaza sank into ^{NC} darkness, while the ^{NC} sky lighted up ...

Example 2.2.13 (Sowa, 2014 my translation)

Gaza versank in ^{NC} Dunkelheit, während der ^{NC} Himmel erleuchtete ...

Example 2.2.14 (Sowa, 2014, my translation)

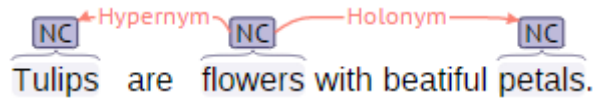
Газа погрузилась во ^{NC} тьму, а небо озарилось от ^{NC} ...

Example 2.2.15 (Sowa, 2014)

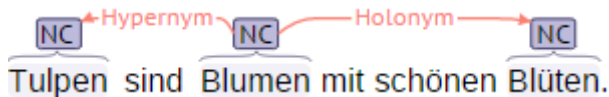
4. Genitives do not warrant a holonymic relation, please take care not to annotate holonymy only due to this grammatical indicator, e.g. *uniform* is a part of *postal clerk*, even if the phrase is *postal clerk’s uniform*.
5. Parts that are close to the whole, but not part of it, are not meronyms, e.g. clothes or symbolic artefacts of persons.
6. Functions or prerequisites are not considered meronyms.

In a holonymic relation, the following rules apply:

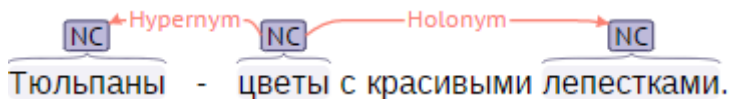
1. When assigning a holonym relation to a nominal, which is in a hypernym relation, always the highest sensible hypernym is to be chosen.



Example 2.2.16

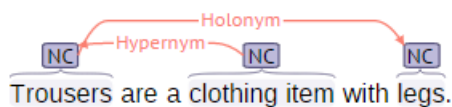


Example 2.2.17

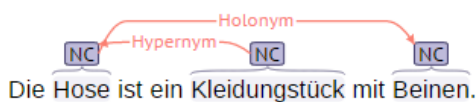


Example 2.2.18

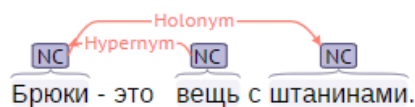
Here, on the other hand, the highest hypernym is not as sensible meronym:



Example 2.2.19⁵⁷



Example 2.2.20



Example 2.2.21

2. A meronym may have several holonyms.
If a meronym is a holonym and its meronyms are also meronyms of the hierarchically higher holonym, this shall also be marked, as holonymy is not necessarily transitive.



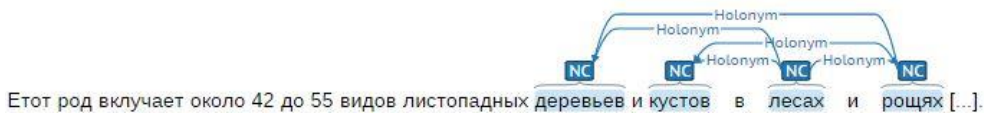
The species includes about 42 to 55 types of deciduous trees and bushes in forests and thickets [...].

Example 2.2.22 ("Äpfel", 2014, para. 1, my translation)

⁵⁷ Clothing item is not as sensible holonym to legs



Example 2.2.23 (“Äpfel”, 2014, para. 1)



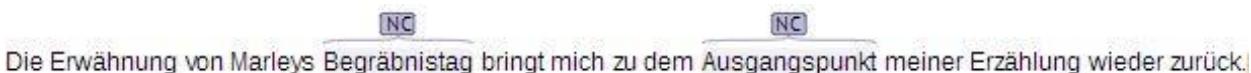
Example 2.2.24 (“Äpfel”, 2014, para. 1, my translation)

3. General Rules

After describing the specifics of the individual relations, some general rules for the annotation of classical semantic relations between nominals shall be explained.

All relations in a paragraph are to be annotated⁵⁸.

1. If a word occurs more than once in a paragraph, the following rules apply⁵⁹.
 - 1) If the identical nominals are in the same sentence, the one that is closer to the related word is to be annotated. If the distance, counted in words between the nominals, is the same between the words, the one on the right is chosen.
 - 2) If the identical nominals are in separate sentences and the related nominal is in one of these sentences, annotate the relationship between the two nominals in one sentence only.
 - 3) If the identical nominals are in one or several sentences and the related nominal is in another sentence, the nominal that is closer to the related nominal is to be annotated. In case of identical distance, apply the same rule as in 1).
 - 4) The closeness rules cease their force if they happen to coincide with synonyms. Then always the first synonym is annotated in relation to other nominals.
2. If some compounds cannot be marked fully because they are separated by other words, they are not marked.
3. Only relations that can be semantically derived from the paragraph are annotated. If relations that generally exist, but are not actually mentioned in the text, like e.g., occur, they shall not be annotated.
4. Annotated relations have to be applicable generally and not only be mentioned in one particular text.



Example 3.1⁶⁰ (Dickens, 1843/1989)

⁵⁸ This means that if a relations occurs more than once, it is annotated as many times as it occurs.

⁵⁹ Here, nominals are considered as the same word regardless of number (e.g. *steward*, *stewards*) and case (e.g. *steward*, *steward's*). Gender (e.g. *steward*, *stewardess*) is considered a differentiating factor.

⁶⁰ Although in this context, *Ausgangspunkt* (engl.: starting point) is synonymous to *Begräbnistag* (engl.: burial day), it is not a general fact, so it is not annotated.

-
5. If there are nominals of interest, which are marked wrongly due to technical reasons, annotators are asked to mark those with the tag “Text mistake”, which is to be found in the Layer “Noun Compound”. Further on the annotators shall treat the nominal as if it was marked regularly as a noun compound.
 6. If a relation between two nominals cannot be determined, but is present in the view of the annotator, it may be marked with the relation *UNCLEAR* and will be further reviewed by the curator.
 7. Units of measure such as metres, litres, minutes, etc. are not annotated.
 8. In case of word ambiguity, the sense of the word in the context of the given paragraph shall be chosen.

References

- Apel'sin (n.d.). In Wikipedia. Retrieved July, 19, 2014, from <https://ru.wikipedia.org/wiki/Апельсин>
- Äpfel (n.d.). In Wikipedia. Retrieved October, 9, 2014, from <http://de.wikipedia.org/wiki/Äpfel>
- Benikova, D., Biemann, C., & Reznicek, M. NoSta-D Named Entity Annotation for German: Guidelines and Dataset.
- Brjuki (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <https://ru.wikipedia.org/wiki/Брюки>
- Chelovek razumnyj (n.d.). In Wikipedia. Retrieved August, 13, 2014, from https://ru.wikipedia.org/wiki/Человек_разумный
- Color (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <http://en.wikipedia.org/wiki/Color>
- Cutlery (n.d.). In Wikipedia. Retrieved May, 26, 2015, from <http://en.wikipedia.org/wiki/Cutlery>
- Delfine (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <http://de.wikipedia.org/wiki/Delfine>
- Dickens, C. (1989). *Christmas Carol*. (R. Zoozmann, Trans.) (Original Work published in 1843). Retrieved October, 9, 2014, from <http://gutenberg.spiegel.de/buch/weihnachtslied-3423/1>
- Essbesteck (n.d.). In Wikipedia. Retrieved May, 26, 2015, from <http://de.wikipedia.org/wiki/Essbesteck>
- Fisalis (n.d.). In Wikipedia. Retrieved July, 19, 2014, from <https://ru.wikipedia.org/wiki/Физалис>
- Flaubert, G. (2006). *Madame Bovary*. New York: Bantam Books. (E. Marx-Aveling, Trans.) (Original Work published in 1981). Retrieved October, 9, 2014, from <http://www.gutenberg.org/cache/epub/2413/pg2413.txt>
- Flaubert, G. (1986). *Madame Bovary*. (A. Schurig, Trans.) (Original Work published in 1981). Retrieved October, 9, 2014, from <http://gutenberg.spiegel.de/buch/frau-bovary-2404/1>
- Flaubert, G. (1956). *Madame Bovary*. (A. Romm, Trans.) (Original Work published in 1981). Retrieved October, 9, 2014, from http://az.lib.ru/f/flober_g/text_0010.shtml
- Football team (n.d.). In Wikipedia. Retrieved August, 13, 2014, from http://en.wikipedia.org/wiki/Football_team
- Fußballmannschaft (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <http://de.wikipedia.org/wiki/Fußballmannschaft>
- Futbol (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <https://ru.wikipedia.org/wiki/Футбол>
- Handbag (n.d.). In Wikipedia. Retrieved July, 19, 2014, from <http://en.wikipedia.org/wiki/Handbag>
- Handtasche (n.d.). In Wikipedia. Retrieved July, 19, 2014, from <https://de.wikipedia.org/wiki/Handtasche>
- Human (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <http://en.wikipedia.org/wiki/Human>
- Koffer (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <http://de.wikipedia.org/wiki/Koffer>
- Mensch (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <http://de.wikipedia.org/wiki/Mensch>
- Molier. *Polnoe sobranie sochinenij v odnom tome*. [Full collection in one volume.] (V. Lihachaev, Trans.) Retrieved July, 19, 2014 from http://az.lib.ru/m/molxer_z/text_0250.shtml
- Moliere, J. B. (1887). *Der Geizige*. (W. H. Graf von Baudissin, Trans.). Halle a.d.S., Druck und

Verlag von Otto Hendel. (Original work published 1668) Retrieved August, 13, 2014 from <http://gutenberg.spiegel.de/buch/1921/1>

Moliere, J. B. (2003). *The Miser*. (C.H. Wall, Trans.)(Original work published 1668) Retrieved August, 13, 2014 from <http://www.gutenberg.org/cache/epub/6923/pg6923.txt>

Nastase, V., Nakov, P., Séaghdha, D. Ó., & Szpakowicz, S. (2013). Semantic relations between nominals. *Synthesis Lectures on Human Language Technologies*, 6(1), 1-119.

Orange (Frucht) (n.d.). In Wikipedia. Retrieved July, 19, 2014, from [http://de.wikipedia.org/wiki/Orange_\(Frucht\)](http://de.wikipedia.org/wiki/Orange_(Frucht))

Orange (fruit) (n.d.). In Wikipedia. Retrieved July, 19, 2014, from [http://en.wikipedia.org/wiki/Orange_\(fruit\)](http://en.wikipedia.org/wiki/Orange_(fruit))

Plag, I. (2003). *Word-formation in English*. Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1988). *A Comprehensive Grammar of the English Language*. *ELT Journal*, 42, 3.

Schmid, H. (1995). Treetagger | a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43, 28.

Sowa, E. (2014). Netjahu: Izrail gotov razshirit'operaziju v Gaze. [Netanyahu: Israil is ready to extend the operation in Gaza.]. BBC. Retrieved from http://www.bbc.co.uk/russian/international/2014/07/140718_israel_gaza_ground_operation

Strelizija (n.d.). In Wikipedia. Retrieved August, 13, 2014, from <https://ru.wikipedia.org/wiki/Стрелитция>

Wells, H. G. (1898). *War of the Worlds*. Retrieved October, 9, 2014, from <http://www.gutenberg.org/cache/epub/36/pg36.txt>.

Wells, H. G. (1927). *War of the Worlds*. (A.K. Pimenova, Trans.) (Original Work published in 1898). Retrieved October, 9, 2014, from http://az.lib.ru/u/uells_g_d/text_1898_the_war_of_the_worlds.shtml

Wells, H. G. (1901). *War of the Worlds*. (G.A. Crüwell, Trans.). Wien, Verlag von Moritz Perles. (Original work published 1898)

Winston, M. E., Chaffin, R., & Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive science*, 11(4), 417-444.

Yimam, S. M., Gurevych, I., Eckart de Castilho, R. , & Biemann, C. (2013, August). WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *ACL (Conference System Demonstrations)* (pp. 1-6).