SemRelData – Multilingual Contextual Annotation of Semantic Relations between Nominals: Dataset and Guidelines

Darina Benikova, Chris Biemann
Language Technology Group,
Computer Science Department,
TU Darmstadt
Hochschulstr. 10,
64289 Darmstadt, Germany
darina.benikova@gmail.com, biem@cs.tu-darmstadt.de

Abstract

Semantic relations play an important role in linguistic knowledge representation. Although their role is relevant in the context of written text, there is no approach or dataset that makes use of contextuality of classic semantic relations beyond the boundary of one sentence. We present the SemRelData dataset that contains annotations of semantic relations between nominals in the context of one paragraph. To be able to analyse the universality of this context notion, the annotation was performed on a multi-lingual and multi-genre corpus. To evaluate the dataset, it is compared to large, manually created knowledge resources in the respective languages. The comparison shows that knowledge bases not only have coverage gaps; they also do not account for semantic relations that are manifested in particular contexts only, yet still play an important role for text cohesion.

Keywords: semantic relations in context, multilingual, multi-genre, annotated dataset

1. Introduction

Although the role of classic semantic relations is considered an important one in context of written text, there was so far no approach or dataset that addresses and quantifies the amount of contextuality of classic semantic relations, in particular beyond the scope of one sentence.

We annotated semantic relations between nominals in the context of one paragraph. We only annotated relations that were present in the context, i.e. where the context hints at the existence of such relations.

We annotated classical semantic relations, such as *synonymy*, *hyperonymy*, *holonymy* and *co-hyponymy* between nominals. *Synonyms* are mostly defined as different words with the same meaning, *hypernyms* are superordinate terms to their subordinate *hyponyms*, *co-hyponyms* are words with the same *hypernym*, and *holonyms* are terms referring to the whole, which consists of *meronyms*.

The contribution of this paper is threefold: we present SemRelData (Semantic Relations Dataset), containing contextual semantic relations, its evaluation and an analysis of the impact of contextuality in semantic relations.

Our dataset consists of 60 language-parallel documents and ∼60.000 tokens. More specifically, it contains parallel encyclopaedic, newspaper and literary texts in English, German and Russian. The dataset was manually annotated in a double-annotator setting by students with linguistic background and subsequently curated. In the next step, the relations were extended using properties of the relations, e.g. *transitivity* of synonymy.

The setting of the multilingual and multi-genre corpus enabled us to analyse the universality of the context notion. The analysis was performed by comparing our dataset to the largest manually created or revised knowledge bases in the respective languages, i.e. WordNet (Miller and Fellbaum, 1991), GermaNet (Henrich and Hinrichs, 2011) and RuTes (Loukashevich, 2011). Annotated relations not present in these databases were further analysed and manually classified according to the reason why they were not present in the respective database.

This study highlights the significance of contextuality of semantic relations and quantitatively assesses this previously neglected phenomenon in a multi-lingual and multi-genre dataset.

This paper is structured as following: Section 2 discusses related work, Section 3 presents the creation and the parameters of the corpus created for this study, Section 4 shows the evaluation and analysis of the corpus, and Section 5 presents the conclusion and ramifications of our results for further work.

2. Related Work

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, but borrowing and trans-disciplinary approaches exist. The consensus that can be found between most involved parties is that paradigmatic semantic relations such as the classical semantic relations among words: “are somehow relevant to the structure of lexical or contextual information.” (Murphy, 2003)[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.

2.1. Classical Semantic Relations

The relations that are referred to as classical semantic relations are those that are called traditional *nym* relations by Murphy (2003) and one of their subtypes. An exact definition of such relations is necessary for a task such as presented in this study. According to Cruse (1986)[84], "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 semantic relations restricted to specific notional areas."

A relatively small number of semantic relations, such as *synonymy*, *antonymy* and *hyperonymy*, has achieved a central role in lexical semantics (Cruse, 1986). Studying the most popular semantic relations, the contextuality of such relations can be shown by comparing the results to previous databases and methods, which have been constructed with no or little contextual information. In the following subsections, the definitions of the semantic relations used in this study will be provided.

Synonymy or sometimes referred to as *poecilonymy*, is regarded as the most significant relation in the WordNet model (Miller and Fellbaum, 1991).

Murphy (2003) defines synonymy as "A synonym set includes only word-concepts that have all the same contextually relevant properties, but differ in form." (Murphy, 2003) [134]. Murphy (2003) 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).

Hyperonymy and Hyponymy According to Cruse, Lyons and Pustejovsky, *hyperonymy*[1] is one of the major structural relations (as cited in (Murphy, 2003)). Generally it is often paraphrased as the *kind-of relation* or as *set inclusion* in logical definitions. Hyperonymy is mostly defined as a unidirectional, non-reflexive and transitive (Murphy, 2003).

An example for hyperonymy would be *bag* (hypernym) and *handbag* (hyponym).

Holonymy and Meronymy *Holonymy*[2] describes the relation of the *part-whole type*. Cruse (1986) 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, e.g. *canonical holonyms*, such as *body* is to *ear*, and *facultative relations* such as *door* to *handle*.

Another crucial distinction that Cruse (1986) makes in order to define holonymy is the distinction between parts and pieces, as illustrated by e.g. "hacking a typewriter into pieces" vs. "unscrewing it into its parts". The portions in the first example are not considered meronyms of typewriter, whereas the ones in the second are considered such.

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 1: Size comparison between different databases

Cruse (1986) 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.

Winston et al. (1987) state that meronymy has often been confused or not clearly distinguished from other semantic relations such as possession, attribution and class inclusion. The consensus on the characteristics of holonymy is that it is an irreflexive and antisymmetric relation (Cruse, 1986; Winston et al., 1987).

2.2. 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 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).

2.3. Knowledge Bases containing Semantic Relations

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 1 gives a size comparison of those databases. The sizes were retrieved from the respective webpages.

2.3.1. Manually created knowledge bases

The collection of the manually created database WordNet started in 1985 (Miller, 1995; Fellbaum, 1998; Fellbaum, 2013). It consists of *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 (Navigli and Ponzetto, 2012) or Mimida[3].

GermaNet (Henrich and Hinrichs, 2011) was constructed similarly to WordNet since 1997 and is free for academic

[1] Hyperonymy is the token>type relation, whereas hyponymy is the type<token relation (Murphy, 2003). In this paper, the term hyperonymy is used preferably.

[2] Holonymy is the has-a relation, whereas its opposite meronymy is the is-part-of relation. In this study the term holonymy is preferred (Murphy, 2003).

[3] http://goo.gl/PIWbSm

use. A similar German database is OpenThesaurus[4], which is available under the GNU license. However, it only provides synonym and association relations (Naber, 2004).

RuTes (Loukashevich, 2011) is an on-going project since 1994 aimed at creating a hierarchical linguistic resource. 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. It is available under the Attribution-Non-Commercial-Share-Alike 3.0 licence[5]. There are further manually or semi-automatic created ontologies for the Russian language, such as RussNet, commercial projects by the enterprises UIS Rossija and Novosoft (Suhonov and Yablonskij, 2004), and Yet Another RussNet (YARN) (Braslavski et al., 2014), however, they are either unavailable or under development.

2.3.2. Automatically or semiautomatically created databases

DBpedia (Lehmann et al., 2014), BabelNet (Navigli and Ponzetto, 2012), Freebase (Bollacker et al., 2008), and Yet Another Great Ontology (YAGO) (Suchanek et al., 2007) use rules to extract information from Wikipedia and other sources. The Never Ending Language Learning (NELL) (Zimmermann et al., 2013) shared knowledge base tries to continuously grow by reading in new resources. The seed knowledge base was an ontology and a set of rules.

3. Corpus

3.1. Collection of dataset

In order to analyse the universality of the contextual property of semantic relations, language parallel texts from three different languages and from three different genre types were collected. 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 texts were mostly taken from Wikipedia, Wikinews and Project Gutenberg to make the dataset freely available.

3.2. Preprocessing

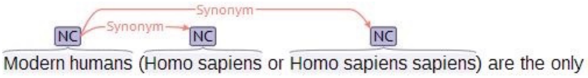
The described texts have been first divided into the paragraphs as indicated in the edition they were taken from. Afterwards the texts were part-of-speech-tagged using the TreeTagger (Schmid, 1994; Schmid, 1995) to simplify the task of annotation. In the .tsv file that was uploaded to the annotation tool WebAnno (Yimam et al., 2014), only the nouns were annotated. Not only simple nouns, but also noun compounds were of interest for the task at hand. However, this task was not conducted with German, as this language is known for its single-token-lexicalization of noun compounds. For English, all spans of continuous noun tag sequences were marked as noun compounds. For Russian no adjustment was made. Annotators were asked to correct false noun compounds if they take part in a semantic relationship.

3.3. Annotation

The annotation was performed using WebAnno[6] (Yimam et al., 2014) in a double-annotation process followed by a curation step. The annotation team consisted of four annotators, all of whom were either native or at least fluent speakers of the language they were annotating. The annotations were performed according to iterativly developed guidelines[7].

WebAnno is a highly flexible web-based annotation tool that allows free definition of span, relation, chain and slot layers and the distributed parallel annotation by several annotators. It provides an interface for curation, i.e. the merging of parallel annotations by different annotators, as well as means to compute inter-annotator agreement (IAA) between annotators.

In this project, two custom annotation layers were created. Although the first layer of type 'span' that captures noun compounds, was annotated automatically in the preprocessing step, annotators were asked to correct wrongly or only partly marked noun compounds that were in a semantic relation to other noun compounds. An exemplary automatic pre-annotation of noun compounds is presented in Figure 1. Furthermore, the noun-compound layer contained the tags *NCpart*[8] and *Textmistake*[9].



Modern humans (Homo sapiens or Homo sapiens sapiens) are the only

Figure 1: Exemplary automatic annotation of English noun compounds.

In the second layer we annotated the classical semantic relations that are of main interest in this study. The layer contained the tags *Hypernym*, *Holonym*, *Synonym* and *Co-hyponym*; furthermore, an uncertain relation could be tagged with ****UNCLEAR****. An exemplary annotation of synonyms is presented in Figure 1.

3.4. Inter-annotator agreement

Generall IAA IAA was computed pair-wisely between annotators using Cohen's κ. Table 2 shows that IAA 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 κs of the individual comparisons of an annotator and a curator in one language range from 0.41 to 0.59.

The comparison of all previously shown average κs 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. Hence, it may be assumed that by the dou-

[4] https://www.openthesaurus.de/

[5] http://creativecommons.org/licenses/by-nc-sa/3.0/deed.ru

[6] https://webanno.github.io/webanno/

[7] For full guidelines see: https://goo.gl/PXaTcu

[8] NCpart denotes a part of a noun compound, which was cut off of its second part, e.g. in the span "ball and other games" "ball" would be annotated as such.

[9] denoting spelling or tagging mistakes in the texts

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

Table 2: κ agreement of all annotators and the curator

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 3: Annotator agreement sorted by time spans

ble annotation and subsequent curation most classic semantic relations that are contained in the texts were found.

Guideline improvement Table 3 shows a clear improvement of inter-annotator agreement, which shows that the agreement increased with the iterative guideline improvement. The average agreement between the annotators is fair and the average agreement between the annotators and the curator is moderate. However, the agreement with the curator drops in the last time-span, which may be explained with a bigger workload at the end of the project leading to negligence.

3.5. Postprocessing

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 features of the individual relations, described in more detail in the Chapter 2 lead to the process of postprocessing of all curated annotations.

In this process, all features of the annotated relations are used, meaning that hyponymy is annotated to its corresponding hyperonymy, meronymy is annotated to its corresponding holonymy and synonymy is reflexively annotated. Then, transitive relations are passed on, in the case of hyperonymy and synonymy.

3.6. Statistics and Characteristics of the dataset

Table 4 shows the statistics of the resulting dataset. 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[10]. The first part consists of the original files in .txt format, the second part consists of the curated files with classical semantic relation annotation in .tsv format and the third part will consist of the ontologies of all files, including the transitive relations.

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 4: Statistics of SemRelData: 1st column: number of noun compounds, 2nd column: number of tokens, 3rd column: number of annotated relations, 4th column: the number of transitive relations

4. Evaluation and Analysis

4.1. Comparison with Knowledge Bases

In this section the comparison with the relations contained in WordNet and its counterparts in the other two languages, GermaNet and RuTes are presented. 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.

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.

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:

- Relation too specific (RS): Though the relation is generally true, it is too specific e.g. *chordates* being a hyponym of *species*.
- 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*.
- Contextual (C): The relation presented by SemRelData is generally not true, but exists in the given context e.g. *recommendation* being a hypernym of *warning*.
- 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.
- Lemmatisation error (LE): The lemmatisation produced a wrong lemma, which was confused with another word e.g. *boxers*, meaning the type of *underwear*, was lemmatised as *boxer*, meaning the *athlete*.

[10] https://www.lt.informatik.tu-darmstadt.de/de/data/semreldata/

- 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).

To compare the relations of the English subset with WordNet 3.0, the NLTK (Bird, 2006) implementation of pywordnet[11] was used. For the lemmatisation, NLTK using the WordNet lemmatiser 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.

To compare the relations of the German subset to GermaNet, the GermaNet Java API and the GermaNet 8.0 version were used. For lemmatisation, the JoBim Text API lemmatiser using the Pretree Tool (Biemann et al., 2008) was applied. 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.

To compare the relations of the Russian subset with RuTes, there was no API available, so the same rules as described in Section 3.5 were applied in order to create the transitive relations[12]. For the lemmatisation process, pymystem3[13], which is a Python wrapper for Yandex Mystem[14], 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.

Table 5 shows the counts of the error type classification of 50 randomly chosen relations for each of the analyzed languages. Summarising 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, implying that the coverage of our new SemRelData resource is even throughout the languages.

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 study is legitimate and yielded correct results.

Error Type	WordNet	GermaNet	RuTes
RS	4	8	12
A	2	7	3
C	9	6	10
SS	9	6	4
LE	1	0	0
U	25	23	21

Table 5: Disagreement analysis of knowledge bases and SemRelData in 50 random relations

Further investigation of the relations that 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 study 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 1). Moreover, WordNet has the lowest rate of disagreement in the categories RS and A, meaning that it has the largest 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. It could be assumed that WordNet is the most representative manual knowledge base, based on its pioneering role and superiority in size. Thus, we further assume that it most representatively shows the gaps in this kind of knowledge base. This is on the one hand a coverage lack of relations due to miscellaneous reasons, but on the other hand due to the negligence of contextual relations, which are relevant to information representation.

4.2. Comparison with Pattern-created Taxonomies

As described above, the automatic 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 Klaussner and Zhekova (2011). The implementation of JoBimText[15] (Biemann and Riedl, 2013) of those patterns was applied to the English source texts that were annotated for Sem-

[11] http://osteele.com/projects/pywordnet/

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

[13] https://pypi.python.org/pypi/pymystem3/0.1.1

[14] https://tech.yandex.ru/mystem/

[15] http://www.jobimtext.org

RelData. As the Hearst Patterns and their extensions are composed for English hyperonymy, only the hypernym relations of the English subset were considered. Those were not lemmatised as the Hearst Patterns produce both lemmatised 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 8 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:

- True (T): 0 instances: the relation is valid and should be present in SemRelData
- Lemma (L): 2 instances: the relation is not present in SemRelData, because it contains lemmas or inflected forms of the related words that are different in the original text, 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.
- General (G): 17 instances: 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 general , e.g. the relation variety as a hypernym of *sweet orange* can be encountered as true, but the term *variety* is too general.
- False (F): 31 instances: the relation is wrong, e.g. *government* as a hypernym of *free trade agreement*.

Distribution of instances shows that 62% of the relations in the random test were wrong and 34% too general[16]. 4% of the relations were not contained in SemRelData due to deviant word forms that are formed by the Hearst Patterns. 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, and these are aggregated over very large corpora, cf. (Panchenko et al., 2016). Thus the results of most single relation extractions are either wrong or too general.

4.3. Comparison between Languages

The number of nominals varies in different languages. 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 in Table 6.

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

Set	№NC	№trans. relations	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 6: Contingency table denoting the number of noun compounds and transitive relations in the language subsets

	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 7: Distribution of Semantic Relation Types in different languages

Table 7 shows the distribution of relation types within the corresponding language. For the calculation, all relations were used. The p-value for the distribution between all three languages is $p < 10^{-6}$, 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 $p = 0.066$. Both pairwise comparisons with Russian are highly significant, the significance value of $p = 10^{-4}$ of the comparison with German being noticeably lower than that of the comparison with English with $p < 10^{-9}$.

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. However, it is also feasible that Russian expresses the same classical semantic relations not through nominals, but through pronouns or other grammatical constructions that avoid specific mention of the referred entity, e.g. the grammar of Russian allows sentences without a subject.

4.4. Comparison between Genres

The comparison of semantic relation distribution was performed analogous to that in Section 4.3. The pairwise differences of all three genres are highly significant, with p-values close to the numeric lower bound. It can be concluded that the density of semantic relations is not evenly distributed among the genre subsets.

Table 8 presents the contingency table of the semantic relation type distribution in the different genres that was used to calculate χ^2. The p-values of the χ^2-test for all three genres as well as all pairwise comparisons are between 10^{-13} and 10^{-20}, meaning that the hypothesis of the semantic relation types being evenly distributed between the different genres

[16] 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

	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 8: Distribution of semantic relation types in different genres

can be rejected.

5. Conclusion and Future Work

In this section, we give a summary of presented results and put the consequences of these results in a wider perspective.

5.1. Summary

In this work, we have presented and evaluated the SemRelData dataset. Furthermore, we analysed the contextuality of classical semantic relations such as synonymy, hyperonymy/hyponymy and holonymy/meronymy across three languages and three genres each. We have described the annotation effort of SemRelData in detail. Further, we have analysed the data with respect to agreement with lexical-semantic knowledge bases, pattern-based extraction mechanisms and differences across languages and genres. Main takeaways of this work are the quantification of contextuality of semantic relations, the incompleteness of lexical-semantic knowledge bases and the inability of pattern-based extraction mechanisms to achieve high extraction coverage. We will now elaborate on these points in more detail before describing possible future extensions.

5.2. Ramifications for the Treatment of Semantic Relations

The results of this study have direct consequences for natural language processing tasks, as they contradict the generally accepted view that semantic relations exist out of context and can be extracted with simple patterns.

Construction and Use of Lexical Resources In the construction of lexical-semantic resources, design decisions have to be taken. While there are many possible ways of organising lexical-semantic resources, e.g. like a taxonomic tree with synsets (cf. WordNet) or a flat hierarchy of topical groupings (cf. Roget's Thesaurus), it is beyond doubt that not all semantic relations between terms that *can* possibly hold in situative contexts should actually be included: such an undertaking, even if it was feasible, would contradict the dictionary-like notion of such resources, which aims to capture the typical cases and aims to not dilute the quality of the resource by including remotely possible relations that might seem random and too situative.

We do not only confirm the inherent incompleteness of lexical resources (cf. "all resources leak" (Biemann, 2012), p.5) but also quantify the amount of leakiness. While the about 50% of missing relations (Class U in Table 5) could possibly be attained with increased efforts in in creasing the coverage of these resources, this is already worrisome in light of the long history of e.g. the WordNet project. More strikingly, however, about 20% of situatively present relations are absent because of their contextuality (Class C in Table 5), which we found to be a surprisingly high amount. This means, any approach that exclusively relies on a lexical-semantic resource for providing semantic relations for text processing suffers not only from their lack of coverage, but also from a principled upper bound on the coverage caused by these contextual relations. In consequence, lexical semantic resources should rather be used as supporting features for text understanding, than as the main driver of it.

Information Extraction and Text Understanding A further unexpected finding was the extent of coverage issues of pattern-based extraction methods, as discussed in Section 4.2. While it is well-known that such patterns lack coverage, which was also confirmed for a much larger set of patterns on web-scale corpora (Panchenko et al., 2016), the amount of erroneous extractions and the almost complete failure to extract any of our annotated relations sheds light on the severity of the problem, which is larger than we anticipated. Pattern-based extraction for taxonomy construction seems only to work when aggregating counts over very large corpora in order to eliminate noise, and patterns only extract the tip of the iceberg of what is actually expressed in these texts since the variability of expressing semantic relations – as shown in our work – is very high and patterns only capture a small part of it.

While the coverage of patters might be improved by using dependency-parse-based representation instead of the flat part-of-speech patterns, we conclude that a higher level of text understanding, including the resolution of anaphora across sentences, will be needed to substantially advance ontology learning from text, cf. (Biemann, 2005).

5.3. Further Work

Although the SemRelData dataset may probably not be used to train machine learning algorithms in the current condition because of its high variability, the continuous improvement of the inter-annotator agreement 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. The current dataset may than be used as a gold standard for the evaluation of automatic procedures, which give rise to address the coverage issues of pattern-based extractors described before.

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 annotation effort discussed herein. 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 summarisation, 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 a contextually-aware component that would add another level of text understanding. To analyse whether the reason for the difference between the two Germanic languages and Russian is actually genealogical, a larger dataset with more related languages, e.g. additions of other Slavic languages and other language families, would allow clearer and more justified statements.

Acknowledgment

The authors would like to thank the annotation team, namely Sarah Holschneider, Ilya Klyashtornyy, Angelina Romaashova and Xaver Ott, without whom the creation of the dataset and the subsequent analysis would not be possible. This work was financially supported by a prize granted by the Adolf Messer Foundation to Chris Biemann.

6. Bibliographical References

Biemann, C. and Riedl, M. (2013). Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. *Journal of Language Modelling*, 1(1):55–95.

Biemann, C., Quasthoff, U., Heyer, G., and Holz, F. (2008). ASV toolbox – a modular collection of language exploration tools. In *Proc. LREC*, pages 1760–1767, Marrakech, Morocco.

Biemann, C. (2005). Ontology learning from text: A survey of methods. *LDV Forum*, 20(2):75–93.

Biemann, C. (2012). *Structure Discovery in Natural Language*. Theory and Applications of Natural Language Processing. Springer Berlin / Heidelberg.

Bird, S. (2006). NLTK: The natural language toolkit. *Proc. COLING/ACL*, pages 69–72. Sydney, Australia.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase. *Proc. ACM SIGMOD*, pages 1247–1250. Vancouver, Canada.

Braslavski, P., Ustalov, D., and Mukhin, M. (2014). A spinning wheel for YARN: User interface for a crowdsourced thesaurus. *Proc. CoNLL*, pages 101–104. Gothenburg, Sweden.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge textbooks in linguistics. Cambridge University Press, Cambridge [Cambridgeshire] and New York.

Christiane Fellbaum, editor. (1998). *WordNet: An electronic lexical database*. Language, speech, and communication. MIT press, Cambridge, Mass. and London.

Christiane Fellbaum, editor. (2013). *WordNet. The Encyclopedia of Applied Linguistics*. Wiley/Blackwell.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proc. COLING*, 2:539–545. Nantes, France.

Henrich, V. and Hinrichs, E. (2011). Determining immediate constituents of compounds in GermaNet. *Proceedings of Recent Advances in Natural Language Processing*, pages 420–426. Hissar, Bulgaria.

Klaussner, C. and Zhekova, D. (2011). Lexico-syntactic patterns for automatic ontology building. *Proc. RANLP Student Research Workshop*, pages 109–114. Hissar, Bulgaria.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al. (2014). DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Loukashevich, N. V. (2011). *Tezaurusy v zadachah informazionnogo poiska [Thesauri in information seeking tasks]*. Izadatel'stvo Moskovskogo universiteta [Moskow University Press].

Miller, G. A. and Fellbaum, C. (1991). Semantic networks of English. *Cognition*, 41(1-3):197–229.

Miller, G. A. (1995). WordNet: A lexical database for the English language. *Communications of the ACM*, 38(11):39–41.

Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy, and other paradigms*. Cambridge University Press, Cambridge, UK and New York.

Naber, D. (2004). OpenThesaurus: Building a thesaurus with a webcommunity. Technical report.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Panchenko, A., Faralli, S., Ruppert, E., Remus, S., Naets, H., Fairon, C., Ponzetto, S., and Biemann, C. (2016). TAXI: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proc. SemEval 2016*, San Diego, CA, USA.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49. Manchester, United Kingdom.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. *Proceedings of the ACL SIG on Linguistic data and corpus-based approaches - Workshop*, pages 47–50. Cambridge, MA, USA.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: A core of semantic knowledge. *Proc. WWW*, pages 697–706. Banff, Alberta, Canada.

Suhonov, A. M. and Yablonskij, S. A. (2004). Razrabotka russkogo wordnet [implementation of a russian wordnet]. *Trudy 6toj Vserossijkoj nauchnoj konferenzii: Elektronnye biblioteki: perspektivnye metody i tehnologii, elektronnye kolekzii, RCDL 2004 [Proceedings of the 6th All-Russian Scientific Conference: Electronic Libraries: Promising methods and technologies, RCDL 2004]*.

Winston, M. E., Chaffin, R., and Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11:414–444.

Yimam, S. M., Eckart de Castilho, R., Gurevych, I., and Biemann, C. (2014). WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proc. ACL Demo Track*, pages 1–6, Baltimore, MA, USA.

Zimmermann, A., Gravier, C., Subercaze, J., and Cruzille, Q. (2013). Nell2RFF: Read the web, and turn it into RDF. *CEUR Workshop Proceedings*, 994:2–8.