Band 23 – Heft 1 – Jahrgang 2008 – ISSN 0175-1336
Zeitschrift für Computerlinguistik und Sprachtechnologie
GLDV-Journal for Computational Linguistics and Language Technology

LDV Forum

Foundations of Ontologies in Text Technology, Part II: Applications

Edited by
Uwe Mönnich and Kai-Uwe Kühnberger

Gesellschaft für linguistische Datenverarbeitung www.gldv.org

Foundations of Ontologies in Text Technology, Part II: Applications

LDV/ Impressum

LDV-Forum **ISSN 0175-1336**	Zeitschrift für Computerlinguistik und Sprachtechnologie GLDV-Journal for Computational Linguistics and Language Technology Offizielles Organ der GLDV
Herausgeber	Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV) Juniorprofessor Dr. Alexander Mehler, Universität Bielefeld, *alexander.mehler@uni-bielefeld.de* Prof. Dr. Christian Wolff, Universität Regensburg *christian.wolff@sprachlit.uni-regensburg.de*
Band 23 – 2008 – Heft 1 **Herausgeber**	Foundations of Ontologies in Text Technology – Applications Prof. Dr. Uwe Mönnich, Universität Tübingen Prof. Dr. Kai-Uwe Kühnberger, Universität Osnabrück
Anschrift der Redaktion	Prof. Dr. Christian Wolff, Universität Regensburg Institut für Medien-, Informations- und Kulturwissenschaft D-93040 Regensburg
Wissenschaftlicher Beirat	Vorstand, Beirat und Arbeitskreisleiter der GLDV *http://www.gldv.org/cms/vorstand.php, http://www.gldv.org/cms/topics.php*
Erscheinungsweise	2 Hefte im Jahr, halbjährlich zum 31. Mai und 31. Oktober. Preprints und redaktionelle Planungen sind über die Website der GLDV einsehbar (*http://www.gldv.org*).
Einreichung von Beiträgen	Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens zwei ReferentInnen begutachtet. Manuskripte sollten deshalb möglichst frühzeitig eingereicht werden. Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der AutorInnen wieder. Einreichungen sind an die Herausgeber zu übermitteln.
Bezugsbedingungen	Für Mitglieder der GLDV ist der Bezugspreis des LDV-Forums im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von 25,- € (inkl. Versand), Einzelexemplare zum Preis von 15,- € (zzgl. Versandkosten) bei der Redaktion bestellt werden.
Satz und Druck	Kai-Uwe Kühnberger und Christian Wolff, mit *LaTeX* (*pdfeTeX / MiKTeX*) und *Adobe InDesign CS3 V 5.0.2*, Druck: Druck TEAM KG, Regensburg

Uwe Mönnich, Kai-Uwe Kühnberger

Editorial

1 Introduction

The division of scientific disciplines into a theory part and a more applied or practical part is a rather common conceptualization of classifying academic fields and can be found in nearly all academic research traditions ranging from the sciences and engineering disciplines to the humanities and educational research. Although there seems to be a natural order of producing scientific results, namely an order of how to develop theory and practice – first, a theory should be developed, i.e. a conceptualization of a certain domain must be provided, and second, this theory can be tested in experiments, implementations, applications etc. – there are many examples in the history of academic disciplines where theory follows practical developments and not vice versa. Text technology is perhaps such an example: Markup standards such as RDF, OWL, or XML, coding initiatives like OLAC (Open Language Archives Community), and practical applications for retrieval purposes (often in business–related contexts) seem sometimes to get ahead of theoretical characterizations of the underlying standards. For example, a standard like OWL Full is at present theoretically not very well understood and it took some time to specify the theoretical machine models of markup languages (like XML) – actually, at a time point after the languages themselves have been accepted as de facto standards.

Nevertheless, we decided to follow the natural order of categorizing cutting–edge research in text technology into theory and practice for this present double volume "Ontologies in Text Technology" of the LDV–Forum: In the first volume, theory–related papers are collected, whereas work shedding light on applications is covered in the present second volume. Although text technology itself and, in particular, its connection to coding semantic knowledge in form of ontologies is a rather young discipline and some practical developments seem to hurry ahead of their theoretical foundations, we think that this order enables the reader to follow a more logical succession of the recent developments. This is further supported by the fact that articles contained in the first volume can be considered in many aspects as a basis for results provided in this second volume. More will be said about these interrelationships between the two types of articles in Section 3.

In any case, we as the guest editors are proud to present the second part entitled "Applications of Ontologies in Text Technology" of the double issue of the LDV–Forum to the research community. We hope that the interested reader can profit from the work collected here and in the best of all possible worlds can pick up some ideas for her own research, in order to further promote text technology and ontology design.

2 The Research Unit 437 Text Technological Information Modeling

During the last six years the development of text technology in Germany was strongly influenced by the research unit 437 "Text Technological Information Modeling" funded by the German Research Foundation (DFG). This research unit is an interdisciplinary research endeavor carried by the Universities of Bielefeld, Gießen, Dortmund, Tübingen, and Osnabrück. Starting in the year 2001, this group constitutes the largest collaborative research project devoted to text technological issues and has provided the basis for text technological research in Germany. Currently this research unit is in its final funding year. In order to get a better idea of the overall project, a concise overview of the subprojects funded during the second phase of this research unit is given:

- *Secondary Structuring of Information and Comparative Analysis of Discourse.*
  Principal Investigator: Dieter Metzing.
- *Induction of Document Grammars for the Representation of Logical Hypertextual Document Structures.*
  Principal Investigator: Alexander Mehler.
- *Text-Grammatical Foundations for the (Semi-)Automated Text-to-Hypertext Conversion.*
  Principal Investigator: Angelika Storrer.
- *Generic Document Structures in Linearly Organized Texts: Text Parsing Using Domain Ontologies and Text Structure Ontologies.*
  Principal Investigator: Henning Lobin.
- *Adaptive Ontologies on Extreme Markup Structures.*
  Principal Investigators: Uwe Mönnich, Kai-Uwe Kühnberger.

Although the research unit tries to cover all aspects of current text technological activities, it is still possible to identify certain core aspects that play a central role in all subprojects. Examples for such vertical topics of the whole research unit are ontologies, annotations, markup standards, and processing aspects of texts. All these topics play an important role in all participating projects. Some aspects of these vertical topics of the research unit are also represented in this double volume of the LDV-Forum. Whereas certain sub-projects of the collaborative research unit mentioned above are represented in Volume I focusing on the foundations of theories for developing, characterizing, coding, learning, and adapting ontological background knowledge as a crucial challenge for the semantic annotation of text documents, other sub-projects document aspects of their ongoing work in the present Volume II "Applications of Ontologies in Text Technology". We think that we can provide in this way not only a representative documentation of text technology in general, but also a representative collection illustrating the research unit 437, in particular.

3 The Structure of Volume II

This second volume collects applied work on ontology design and text technology. The rticles span a field from ontologies in discourse parsing and lexical semantics to anaphora resolution, linguistic annotations, and the automatic acquisition of formal concepts

from textual data. It is important to notice that there are many connections between the articles published in the two parts of the double volume. In particular, several foundational results presented in the first volume provided the basis for applications in the present volume. We will try to make some of these obvious connections visible while roughly summarizing important topics of the contributions collected here.

In his article "An Ontology of Linguistic Annotations", Christian Chiarcos discusses necessary design features of ontological resources for annotations mainly intended for terminological integration, and ontology–based search across linguistic resources with heterogeneous annotations. By developing a structured ontology involving selfcontained sub-ontologies, which are linked in a declarative way, he shows how a separation between the annotation documentation and its interpretation with respect to the reference terminology can be achieved. The underlying idea is a mapping process of annotations onto ontological representations, such that the full range of types of information in annotations (like syntactic, semantic, phonological etc. information) can be referenced by an ontology. A theoretical basis of the ideas spelled out in Chiarcos article can be found in the contribution of the first volume "Towards a Logical Description of Trees in Annotation Graphs" by Jens Michaelis and Uwe Mönnich.

The contribution "OWL Ontologies as a Resource for Discourse Parsing" by Maja Bärenfänger, Mirco Hilbert, Henning Lobin, and Harald Lüngen bases discourse annotations on the Rhetorical Structure Theory (Mann and Thompson, 1988) to automatically arrange discourse segments and rhetorical relations in a tree structure. The resources to extract these tree structures are based on heterogeneous types of information like discourse marker lexicons, lexico-semantic ontologies, and annotation layers of input text. The article focuses particularly on OWL ontologies and how they can be consulted by the discourse parser. An important role plays the usage of an OWL version of GermaNet and a taxonomy of rhetorical relations which was developed by the authors themselves. Certain aspects of this paper are practical applications of the theoretical results of the contribution "Domain Ontologies and Wordnets in OWL: Modelling Options" by Harald Lüngen and Angelika Storrer of the first volume.

The development of automatic extraction procedures for generating cheap but nevertheless reliable ontologies seems to be one of the most important practical challenges for text technological research (Perez and Mancho, 2003). In particular, synonymy information seems to be a good starting point for such an endeavor to identify different candidates for one and the same concept (word sense). A Kumaran, Ranbeer Makin, Vijay Pattisapu, Shaik Sharif, and Lucy Vanderwende examine in their article "Evaluating the Quality of Automatically Extracted Synonymy Information" two complementary techniques in order to automatically extract synonymy information from large corpora: First, a generic broad–coverage parser for generating bits of semantic information and second, their synthesis into sets of synonyms using word-sense disambiguation with latent semantic analysis. The authors evaluate their approaches quantitatively and qualitatively. From a general perspective this article is a further step towards the whole cycle of automatic ontology generation: extracting semantic information, expanding an ontology with additional information, and adapting this expanded ontology if necessary. In this

sense, the present paper complements the article "Automatic Ontology Extension: Resolving Inconsistencies" by Ekaterina Ovchinnikova and Kai-Uwe Kühnberger in the first volume.

A classical challenge of natural language processing concerns nominal anaphora resolution, partially because several types of knowledge have to be taken into account as, for example, morphosyntactic information and domain knowledge. The paper "Resolving Nominal Anaphora Using Hybrid Semantic Knowledge" by Daniela Göcke, Maik Stührenberg, and Tonio Wandmacher proposes a hybrid approach towards extracting automatically necessary domain knowledge: first, they propose a knowledge-free approach of distributional similarity (Paaß et al., 2004) based on latent semantic analysis, in this respect comparable to the paper "Evaluating the Quality of Automatically Extracted Synonymy Information" above, and second, they use Hearst patterns (Hearst, 1982), i.e. predicate-argument relations encoded in the syntactic structure of the text. The integration of semantic relatedness by combining information about extracted relations and cooccurrence information is used to identify the most likely antecedent in anaphora resolution tasks. The authors evaluate their approach on a corpus of German scientific and newspaper articles.

The final contribution of the present volume "Automatic Acquisition of Formal Concepts from Text" by Pablo Gamallo Otero, Gabriel Pereira Lopes, and Alexandre Agustini uses formal concept analysis (Priss, 2006) in order to implement an unsupervised learning procedure for concept acquisition from annotated corpora. The idea is to build bidimensional clusters of words and their lexico–semantic contexts. Their procedure results in a concept lattice describing a domain–specific ontology underlying the training corpus. The authors use for their evaluation a large Portuguese corpus where the tokens were extracted from a general–purpose journal and an English excerpt of the European Parliament Proceedings.

4 Acknowledgments

This volume would not have been possible without the help of many people. In the first place, the guest editors want to thank the editors-in-chief of the LDV–Forum, Alexander Mehler and Christian Wolff. Their encouragement and support in all phases of the emergence of this double volume has been irreplaceable in completing it. Furthermore we want to thank the German Research Foundation for financial support of the research unit 437 "Text Technological Information Modeling" and particularly the speaker of this research unit, Dieter Metzing.

Last but not least, the editors want to thank the program committee for their careful evaluations of the submitted papers. The quality of this volume is also a direct reflection of the work these reviewers invested. The program committee consisted of the following researchers (in alphabetical order): Irene Cramer, Thierry Declerck, Stefan Evert, Pascal Hitzler, Wolfgang Höppner, Helmar Gust, Marcus Kracht, Edda Leopold, Alessandro Moschitti, Larry Moss, Rainer Osswald, Olga Pustylnikov, Georg Rehm, Hans-Christian Schmitz, Bernhard Schröder, Uta Seewald-Heeg, Manfred Stede, Markus Stuptner, Frank Teuteberg, Yannick Versley, Johanna Völker, Armin Wegner, and Christian Wolff.

References

Hearst, M. (1982). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the Fourteenth International Conference on Computational Linguistics.

Mann, W. and Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organsation. Text, 8(3):243–281.

Paaß, G., Kindermann, J., and Leopold, E. (2004). Learning prototype ontologies by hierarchical latent semantic analysis. In Knowledge Discovery and Ontologies, Pisa, Italy.

Perez, G. A. and Mancho, M. D. (2003). A Survey of Ontology Learning Methods and Techniques. OntoWeb Delieverable 1.5.

Priss, U. (2006). Formal concept analysis in information science. Information Science and Technology, 40:521–543.

Inhaltsverzeichnis

LDV FORUM – Band 23 (1) – 2008
Foundations of Ontologies in Text Technology – Applications

Christian Chiarcos

An ontology of linguistic annotations

This paper describes development and design of an ontology of linguistic annotations, primarily word classes and morphosyntactic features, based on existing standardization approaches (e.g. EAGLES), a set of annotation schemes (e.g. for German, STTS and morphological annotations), and existing terminological resources (e.g. GOLD).
The ontology is intended to be a platform for terminological integration, integrated representation and ontology-based search across existing linguistic resources with terminologically heterogeneous annotations. Further, it can be applied to augment the semantic analysis of a given text with an ontological interpretation of its morphosyntactic analysis.

1 Background and motivation

This paper describes the development and the design of an ontology of linguistic annotations. The ontology is primarily intended as a platform for the terminological integration, integrated representation and access to existing linguistic resources with terminologically heterogeneous annotations. This means that existing annotations are mapped onto ontological representations, according to the underlying semantics a certain tag is assigned.

Beyond this, the ontology can also be applied to the ontological representation of linguistic information in a hybrid model of automated text analysis covering both semantic and morphosyntactic information. Cimiano and Reyle (2003) developed the idea that both semantic and syntactic analysis must integrated within a hybrid system using both types of information. Further, de Cea et al. (2004) proposed to model the dependencies between these two modules at the same level of conceptual representation, i.e. a system of multiple ontologies covering both the semantic concepts of an analyzed text, and the semantics of its linguistic (morphosyntactic) annotations.

Thus, the ontology-based integration of linguistic annotation terminology can be used in two different ways:

Annotation mining perspective The ontology specifies a reference inventory of terms and definitions to which different annotations refer. But also, the ontology assembles and formalizes the available annotation documentation which a user has to consult

to explore a corpus. The annotation mining perspective is basically that of a linguist searching for examples in a corpus.

NLP perspective The ontology specifies a framework for tag-set independent representation and semantic interpretation of linguistic annotations as produced, for example, by a statistical tagger. In this function, an ontology provides a semantic interpretation of linguistic annotations.

The annotation mining perspective is particularly relevant to typological and corpus linguistic research. Attempts for the standardization of morphosyntactic annotation have been made, basically presented by the lists of terms and abbreviations, e.g. the EUROTYP guidelines (König et al., 1993), but also as terminological networks and ontologies, e.g. the Generalized Ontology of Linguistic Description (GOLD) (Farrar and Langendoen, 2003).

Related research on the NLP perspective has mostly relied on the specification of a standard repertoire of linguistic terms which may be used by or must be supported by standard-conformant tag sets, the most prominent example being the EAGLES recommendations for morphosyntax (Leech and Wilson, 1996). An ontology for the linguistic annotations produced by different parsers for Spanish has been described by de Cea et al. (2004).

The classical domain of an ontology besides the annotation mining perspective and the NLP perspective is the **terminological perspective**. In this function, an ontology is employed to specify the linguistic terminology as used in an existing body of literature, a line of research currently explored by Schneider (2007), but not specifically tailored to annotation-relevant terminology.

The ontology presented here, however, is designed with a primary focus on the annotation mining perspective. It is developed in the context of the project "Sustainability of Linguistic Data" to enhance the terminological integration of the resources assembled by three German Collaborative Research Centers, CRC 441 (Tübingen, "Linguistic Data Structures") CRC 538 (Hamburg, "Multilingualism"), and CRC 632 (Potsdam/Berlin,"Information Structure"). Furthermore, the ontologies are applied for tag-set independent, ontology-based corpus querying.

This search functionality represents one of the most important fields of applications for the ontology described here (see Chiarcos (2006) for more details). Still, in the context of this volume, I concentrate on the description of the ontologies themselves, and in particular, in their function as a means for conservation and systematization of annotation documentation. Also, their potential application for the purpose of NLP applications will be shortly sketched, as the ontology also deals with annotation schemes for German, English and Russian which are technically relevant.

Here, I concentrate on part of speech (POS) and morphological annotation. Our research centers create and use morphosyntactically annotated corpora for about 42

meta tag sets and multilingual tag sets		language-specific tag sets	languages	granularity
	n/a	Tibetan tag set	Tibetan	$\geq$ 36 tags
EAGLES	generalization over existing tag sets for European languages	Susanne	English	$\approx$ 420 tags
		STTS, 3 variants	German	54 (718) tags
		Menota	Old Norse	$\approx$ 13055 tags
MULTEXT-East	adaptation of EAGLES	Russian tag set	Russian	$\geq$ 877 tags
CRC632 annotation standard	designed for typological research	n/a	$>$ 26 languages	$\approx$ 79 tags
CRC538/E2 tag set	reduced tag set for acquisition studies	n/a	German, Romance, Basque	$\geq$ 8 tags

Table 1: Tag sets and meta-tag sets for part of speech (POS) annotation in the CRCs.

languages or language stages from practically all parts of the world, cf. tab. 1. With respect to annotation schemes applied, Susanne (Sampson, 1995, English), STTS (Schiller et al., 1995, German) and the Uppsala tag set (Russian) are also technically relevant, as they are used by existing POS taggers.

The scenario of the sustainability project is that a linguist can assess the value of a given resource without being too familiar with the annotation scheme. Here, the user may encounter even greater problems hindering the direct access to the data or proper interpretation of tags: tag names are cryptic and appear in idiosyncratic variants, researchers from different communities use tags with the same names, but different definitions, tag definitions can be extremely complex, or be missing completely, or be of differing granularity.

As an example, the dialects of STTS show some degree of variation in the tag used for pronominal adverbs (`PAV`, `PROAV`, `PROP`). Such seemingly marginal variations can lead to false conclusions about the distribution of grammatical categories if they remain undetected, especially in queries with regular expressions. Further, tag sets tend to apply surface ambiguity as a criterion for the assignment of POS tags. As an example, the STTS tag `VAFIN`, intuitively interpreted as "auxiliary verb", applies to all uses of German *haben* and *sein*, in both auxiliary function ("to have, to be") and lexical use ("to own, to exist"). An ontology-based approach provides a natural base for the handling of both problems, it allows abstracting from the concrete surface form of a tag. Also, the possibility to formulate complex relationships between concepts can be used to make contra-intuitive definitions explicit.

Especially, if annotation documentation is generated from such ontological representation, sincere pitfalls of corpus research can be avoided. A widespread strategy to quickly find the tag one is looking for is to search for an appropriate example word and look up its part-of-speech tags in the corpus. For the case of `VAFIN` in STTS, this strategy is particularly treacherous, as the auxiliary use of *haben* and *sein* is not only explicated

by the abbreviation, but it also occurs more frequently than the lexical use. Using this corpus-based strategy of annotation exploration, inclusion of lexical verbs under `VAFIN` will often remain undetected. Using reference definitions to explore annotation schemes helps to avoid such problems.

2 Toward an ontology of linguistic annotations

One appealing solution to the problem of terminological heterogeneity is the standardization approach as employed by the Expert Advisory Group on Language Engineering Standards (EAGLES), an initiative of the European Commission concerned with the development of standards for large-scale language resources. In this context, Leech and Wilson (1996) formulated recommendations for morphosyntactic annotation, further referred to as the "EAGLES meta scheme". In a bottom-up approach, existing tag sets for several European languages have been considered, and commonly used terms and categories have been identified.

This surface-oriented approach, however, faced several problems. First, the outcome of the bottom-up process was merely a list of terms illustrated with examples, but not a fully developed terminological resource with concise definitions. As a consequence, incompatible interpretations of the common terminology occurred among standard-conformant tag sets, contradicting any effort of standardization (Hughes et al., 1995). Further, the standardization approach relies on a direct mapping between concrete tag sets and the meta scheme, that is, every obligatory category in the meta scheme must be implemented by a standard-conformant tag set, and every recommended feature should be implemented. This direct mapping results in a projection of complexity between tag sets and meta scheme. For example, in order to define a standard-conformant tag set for, say, Russian, the tag set needs to provide a tag for articles, which are, however, inexistent in Russian. This problem escalates as the number of languages (a standard is applied to) increases, and in fact, it has been questioned whether universal, or 'obligatory' categories exist at all (Broschart, 1997). Thus, any standardization approach is inherently restricted to a limited set of languages, and is not a general solution for a project also working with data from typological research.

As ontologies provide means for well-defined, structured terminological resources, it seems that these problems can be most easily overcome by the application of an ontology similar to the GOLD approach (Farrar and Langendoen, 2003). Instead of providing a generalization of tag sets for a fixed range of languages, it aimed to cover the full typological variety as far as possible. Finally, it took a different starting point than the EAGLES recommendation due to its orientation towards the documentation of endangered languages. As opposed to this, our joint initiative aims to achieve a unified representation and access to existing resources, which – in their quantitative majority – deal with European languages. Accordingly, we develop an ontology based

on a harmonization between EAGLES, GOLD, and the annotation schemes assembled in section 1.

The ontology is created using a three-step methodology: (i) derive an ontology from EAGLES, (ii) integrate other non-EAGLES conformant tag sets, and finally (iii) harmonize this ontology with GOLD. After an ontology for word classes, resp. part of speech tags, had been completed, this procedure was repeated for morphological features.

The result of this process is the "Reference Model", an ontology of terminology used for linguistic annotations. The basic structure of the Reference Model is derived from EAGLES, but augmented and partly redefined with reference to specific annotation schemes, formalized as "Annotation Models", and the GOLD ontology.

2.1 Building the Reference Model

As an illustration, we consider the special case of nouns. The original definition in the EAGLES recommendations (Leech and Wilson, 1996) is given as:

Nouns (N)					
1. Type:	1. Common	2. Proper			
2. Gender:	1. Masculine	2. Feminine	3. Neuter		
3. Number:	1. Singular	2. Plural			
4. Case:	1. Nominative	2. Genitive	3. Dative	4. Accusative	5. Vocative

Concentrating on the 'Type' feature as a major subclassification among two distinctive parts of speech, we can derive a rudimentary taxonomy of nouns with the concept Noun and two sub-concepts CommonNoun and ProperNoun. The initial, weak ontological representation of the EAGLES meta scheme constructed from such implicitly hierarchical structures is further refined by references to annotation schemes which introduce additional concepts that are usually not assumed for European languages. Examples for such extensions are adverbial participles in Russian, verbal nouns in Cushitic languages, and noun classifiers in Asian languages.

These categories were then aligned with the corresponding categories in the GOLD ontology (Farrar and Langendoen, 2003), which proves especially helpful for the handling of concepts whose interpretation is varying in different tag sets, such as understanding of possessive pronouns which are either regarded as determiners (because of their syntactic function), or as pronouns (because of their semantic function).

For the case of nouns, however, the linking with GOLD introduces another possible perspective on the subclassification of nouns. The concept Noun probably corresponds to Noun_G: "a broad classification of parts of speech which include substantives and nominals". The concept ProperNoun is reserved explicitly for names, and thus covers a sub-class of Substantive_G ("names of physical, concrete, relatively unchanging experiences"). As opposed to this, CommonNoun possibly represents a more general concept than Nominal_G ("whose members differ grammatically from a substantive but which

functions as one"). Especially, CommonNoun covers certain instances of Substantive$_G$ as well. As evident from this example, the GOLD definitions are based on other conceptualizations than those applied in traditional Latin-based grammars underlying most European tag sets. Hence, the Reference Model combines both sub-classifications of nouns.

2.2 Building Annotation Models

The focus of the approach is to integrate existing, heterogeneous terminologies used in existing annotations. In order to achieve sustainability of existing annotations, this also entails the premise to preserve and to systematize the information conveyed in the original annotation documentation.

Therefore, any annotation scheme is formalized within one self-contained ontology, the *Annotation Model*. The Annotation Model is created on the basis of an exhaustive collection of the available annotation documentation. However, besides the information directly formalized in the ontology, the descriptions and a selection of representative examples found in the annotation documentation are preserved and added as comments to concepts and properties in the ontology. For documentation purposes, a hypertext is created from the Annotation Model which conveys both the structure of the Annotation Model and these comments.

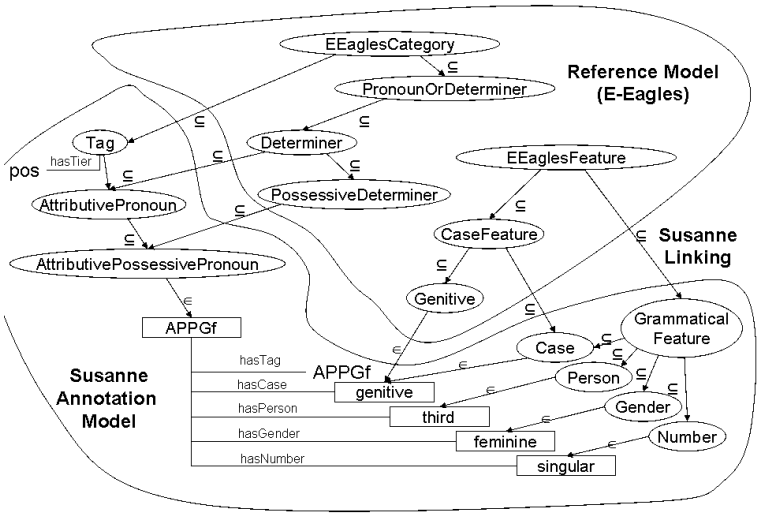
Considering the German tag set STTS as an example, a hierarchically structured Annotation Model can be derived in a similar way as described above. Unlike the EAGLES recommendations, STTS guidelines give detailed enumerations of use-cases, prototypical examples, and critical cases. Further, the aspect of hierarchical structuring is explicitly emphasized. So, the EAGLES-based Reference Model concepts Noun and the sub-concepts CommonNoun and ProperNoun can be aligned easily with the (partial) tags `N` (subsuming `NN` and `NE`), `NN` (concrete and abstract nouns, nominalizations, etc.) and `NE` (surnames, place names, etc.).

The linking between Annotation Models and the Reference Model is implemented by means of conceptual subsumption (`rdfs:subClassOf`), resulting in a complex ontological structure, see 1. An important difference as compared to the standardization approach, the linking does not only allow for underspecification and disjunction, but it also supports formulating complex linking relations with any combination of set operators.

2.3 Integrating morphological features

So far, I concentrated on the construction of a weak ontology of part of speech tags. In a second step, also grammatical features recommended by Leech and Wilson (1996) were integrated into the Reference Model. While word classes are realized as OWL classes

Figure 1: The Susanne tag `APPGf`, its representation in the Annotation Model and (partial) linking with the Reference Model.



in the ontology, grammatical features are encoded as object properties, relating word classes with concepts describing the corresponding grammatical features. Similarly, grammatical information in the corresponding Annotation Models is specified and linked to the Reference Model specifications. The linking between grammatical feature values is modeled by `rdfs:subClassOf`, the linking between object properties is modeled by `rdfs:subPropertyOf`.

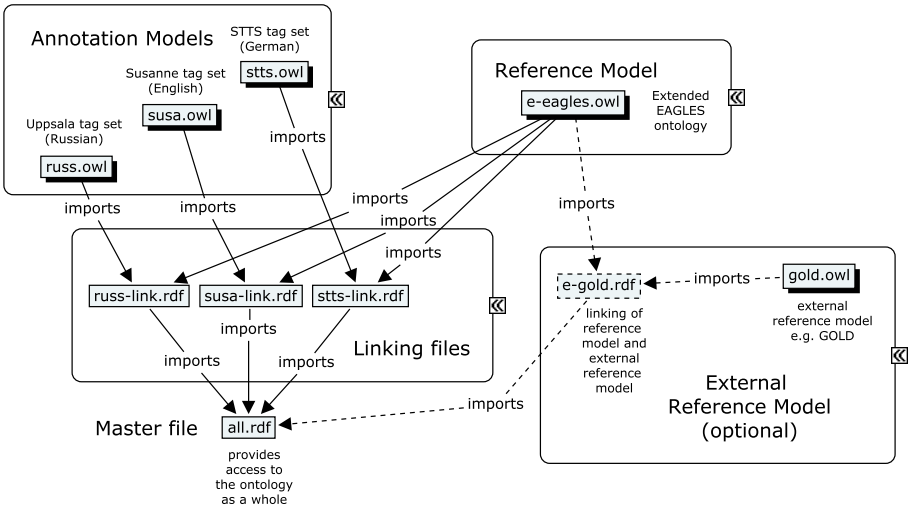
In addition to the formalization of POS tag sets enumerated in table 1, morphological information from the Susanne tag set (English), the Uppsala tag set (Russian), the TIGER annotation scheme (Brants and Hansen, 2002, German) and the CRC632 glossing guidelines are implemented in the corresponding Annotation Models.

For the ontological representation of one example tag from the Susanne tag set, `APPGf`, used for *her* as a possessive, the corresponding inheritance structure of the word class and the case property is presented in fig. 1. Using these inheritance structures, the Susanne tag `APPGf` can be rendered in terms of the Reference Model:

```
PossessiveDeterminer and hasCase(Genitive) and hasPerson(Third) and hasGender(Feminine)
and hasNumber(Singular)
```

The important difference between this description and the (similar) description in terms of the Annotation Model is that this description is tag-set neutral, and does not

Figure 2: The structured ontology.



only apply to the English *her* as a possessive, but also to the corresponding tags in other annotation schemes. The same ontological definition also applies for German *ihr* with the STTS pos tag `PPOSAT` in combination with the morphological description `*.Sg.Fem`, and in the application of the ontology for tag-set neutral corpus querying, this description may be used to retrieve the corresponding tags within different annotation schemes.

3 A structured ontology

As for the technical realization of the ontology, the ontology is broken into multiple OWL files cf. fig. 2) which respectively encode (i) the Reference Model, (ii) several Annotation Model, and (iii) the linking between a Reference Model and each particular Annotation Models.

The components as well as the ontology as a whole are defined in OWL/DL, thus enabling the processing with OWL/DL reasoners.

Reference Model The Reference Model represents the 'terminological backbone' of the structured ontology. As the skeleton of the Reference Model is originally derived from the EAGLES meta scheme as described above, it is associated with the name space `e-eagles`, i.e. extended EAGLES.

Annotation Model Annotation Models represent self-contained ontologies covering the documentation available about a particular annotation scheme. POS tags are modeled as instances, with every tag corresponding to one single instance. The surface form of this instance is defined by means of the property `hasTag`.[1]

Further, different annotation schemes employ different classifications of levels of annotation. Morphological information can be annotated on an independent annotation layer `morph`, it may be integrated with POS annotation, or together with semantic annotations on the annotation layer `gloss`. Thus, property `hasTier` specifies the name of the annotation layer where the corresponding annotation is to be found in accordance with the annotation guidelines.

Again, OWL name spaces are introduced to separate different Annotation Models (`stts, susa, russ, ...`) and Reference Model (`e-eagles`).

Linking Annotation Model and Reference Model Reference Model and Annotation Model are independent ontologies of linguistic terms. Thus, the linking between them has to be made explicit. For this purpose, we apply separate owl files which import Reference Model and Annotation Model. For every Annotation Model, say `stts.owl`, a corresponding link file `stts-link.rdf` exists. In this link file, the relationship between the STTS Annotation Model concepts and Reference Model concepts is represented in a declarative way, by means of `rdfs:Descriptions` pertaining `rdfs:subClassOf`-statements.

As both Annotation Model and Reference Model can have independent hierarchical structure, it is not necessary to assign every single tag to a concept of the Reference Model by its own. Rather, explicit references between Annotation Model concepts and Reference Model concepts are possible, thus making instances of Annotation Model concepts indirect instances of Reference Model concepts.

Linking Reference Model and external Reference Model The same mechanism as applied for the linking between Annotation Model and Reference Model may be used to relate Reference Model concepts with external ontological resources. A possible external reference model is GOLD, resp. its modified variant, with which the current Reference Model is linked to. This differentiation allows a user to differentiate between the modeling of linguistic terminologies in general (or by a specific community, that is the primary function of the external reference model) and the formalization and generalization over specific annotation guidelines. Only the latter is the primary function of the (internal)

[1] For more complex tag sets, which involve also information about morphology (such as the Uppsala tag set for Russian with 877 known tags) or semantic classes (such as the Susanne scheme for English with 420 tags), however, it is reasonable not to require a 1-to-1 mapping between instances and tags, but to rather assemble multiple tags under one instance. Thus, the property `hasTag` can be replaced by `hasTagStartingWith`, `hasTagEndingWith`, or `hasTagContaining`.

Reference Model. However, by specifying a linking between the internal Reference Model and some external reference model, the external reference model is indirectly related to the Annotation Models as well. The internal Reference Model thus serves to mediate between Annotation Models and external reference models. In this sense, the internal Reference Model provides an *interface* to the Annotation Models it is associated with.

The master file Finally, one additional file is needed which represents an interface to the ontology as a whole. Basically, this is an OWL file importing all relevant linking files (these are importing Reference Model and Annotation Models). When loading this master file, the whole ontology with all the parts becomes available to the importing program.

4 Application and evaluation

4.1 Fields of application

At the moment, the ontology focuses on the annotation mining perspective, with an application to ontology-based corpus querying and annotation documentation.

For this purpose, we have developed a problem-specific HTML visualization,[2] which enables a user to browse the ontology, in order to find out definitions of tags and concepts within an Annotation Model and their relationship to the Reference Model. As the ontology contains the comments from the original annotation documentation, it is not to be misunderstood as an ontology of linguistic terminology *in general*, as the ontologies developed by Schneider (2007) and Farrar and Langendoen (2003). Rather, the ontologies described here only concern the documentation of existing annotations, without making any claims about the use of terms beyond this. Still, it would be a great achievement to relate these or similar approaches to each other, thus directly relating terms discussed in grammatical theory with concrete linguistic annotations.

Moreover, the OntoClient was implemented, a JAVA-based pre-processor for corpus queries, which supports annotation-independent search queries by using concepts and definitions in the Reference Model. In essence, it is a specialized OWL reasoner, which translates ontological descriptions of concepts and properties into a disjunction of instances from which, then, the form of tag and the annotation can be retrieved using the `hasTag` and `hasTier` properties. Using the OntoClient, the Reference Model definitions can be applied for the formulation of tagset-neutral corpus queries, cf. Rehm et al. (2007).

[2] `http://nachhalt.sfb632.uni-potsdam.de/OntoBrowser`

4.2 Application in NLP contexts

In addition to this kind of technical application, the ontologies can be used for semantic interpretation of linguistic annotations independently of the underlying tag set. One domain where technical applications can benefit from an ontological interpretation of linguistic annotations is their natural handling of underspecification. More precisely, the accuracy and robustness of tools like taggers or parsers can be improved by the application of tool-specific ontologies.

As an example, consider the tagging of the substitutive demonstrative pronouns *der, die, das* in German. These are homonymous with the definite article and the relative pronoun, and thus, for correct identification of these pronouns, a (partial) syntactic analysis is needed. Schmid (1994)'s TreeTagger achieved a precision of 89.2% and a recall of 92.4% for the corresponding STTS tag `PDS` on the morphosyntactically analyzed Potsdam Commentary Corpus (Stede, 2004). `PDS` was misleadingly chosen for manually annotated `PDAT` (attributive demonstrative pronoun) in 5.0% of the cases and for `PRELS` (substitutive relative pronoun) in 5.8% of the cases. In terms of the ontology, this could be expressed by assigning the tag not to one ontological concept, but rather to a disjunction of the ontological concepts.[3] In this way, tool-specific underspecified ontologies for annotation schemes can be derived from an Annotation Model using a manually annotated reference corpus and the output of the corresponding tool.

On ontological, tag-set independent representation also allows to combine information from different linguistic tools such as another tagger. Considering two currently used POS tag sets for German, STTS and the Morphy tag set (Lezius et al., 1998), we find differentiations in both tag sets that are absent in the other. As such, Morphy distinguishes definite and indefinite articles, both tagged as `ART` in STTS. On the basis of an ontological representation, however, both analyses can be represented not only in parallel, but also as a conjunction, and, for this case, they can also be simplified.

$$\texttt{DefiniteArticle} \cap \texttt{Article} = \texttt{DefiniteArticle}$$

In a similar way, it is possible to enrich linguistic analyses with semantic analyses and vice versa, e.g. in the resolution of underspecification at both levels, as suggested by Cimiano and Reyle (2003). Following de Cea et al. (2004), such dependencies can benefit from the use of ontologies as a common elementary representation for both linguistic and semantic features within a text.

[3] It seems reasonable not to require the ontological interpretation of the tagger to cover any possible exception but only systematic errors. By demanding a minimal precision of 95% of the output, then, the ontological representation could be defined as the disjunction of the most frequent concepts, i.e. `PDS` and `PRELS`. A possible underspecified ontological interpretation of this disjunction was `SubstitutivePronoun`. As opposed to the original tagger output, a minimal precision of 95% percent is guaranteed for this interpretation.

The ontology presented above represents an elementary component for such a hybrid system, in particular with reference to the Annotation Models for German, English and Russian, which are employed by automatic tools.

4.3 Evaluation

The ontologies developed so far are comparably small[4] and are based only on annotation documentation, i.e. a limited selection of documents, as their source.

The hierarchical structure in the Reference Model and the Annotation Models follows from the hierarchical structure reflected in the annotation documentation resp. in the EAGLES recommendations, that is, usually one single document, for which an ontology construction procedure has been described above. For reasons of size and great homogeneity of the textual base of the ontology, an evaluation of *structural* properties of the ontologies, such as detection of cycles, seems unnecessary.

The linking was developed in co-operation with specialists for the corresponding domain and literature on the language under consideration. For non-European languages, thus, any expert knowledge that was available was dedicated to the refinement and precision of the linking, rather than its evaluation. For better-known European languages, the linking was adapted from the EAGLES recommendations, and only modified were more precise definitions of terms were provided.

As for the qualitative evaluation of the Reference Model, the implementation of several linkings with external Reference Models revealed that the conceptualizations and the definitions adopted in the Reference Model are compatible with these external Reference Models, confirming its validity. In particular, the morphosyntactic module of the OntoTag ontologies (de Cea et al., 2004) and the Data Category Registry (Ide et al., 2005) are important in this respect, as these had not been consulted during the design of the Reference Model. The morphosyntactic module of the OntoTag ontologies was developed on the basis of the EAGLES recommendations, but specifically for Spanish. The ontology differs from the Reference Model, in that the following characteristics were specified: `exhaustive`, `disjoint`, `partition` and `partOf`. These were excluded from the Reference Model in order to guarantee applicability to languages which require introduction of additional concepts. Yet, the concepts identified, the hierarchical structure and the grammatical features could be mapped onto each other, most exceptions being extensions of the OntoTag ontologies specific to Spanish.

As for the linking with the Data Category Registry categories, an OWL representation of the data categories specified by Monachini et al. (2005) was developed. The linking between this DCR ontology and the Reference Model could be established only on

[4]The Reference Model consists of 18 object properties (grammatical features), 161 classes (word classes, grammatical categories), and has a maximum inheritance depth of 5. The Uppsala Annotation Model consists of 18 object properties (grammatical features) and 79 classes (word classes and grammatical categories) with a maximum inheritance depth of 4. Further, it contains 906 instances (tags and values of grammatical features).

the basis of similarity of concept names, as Monachini et al. (2005) did not provide definitions. For word classes, however, 85.71% of top-level concepts could be linked to morphological feature types in DCR, indicating that with the exception of specifics of the typologically-oriented annotation schemes considered, the Reference Model formalizes a sub-set of DCR categories.

In this sense, the validity of the Reference Model with respect to two external knowledge sources has been shown. The high level of agreement between these is most likely due to the influence of the EAGLES recommendations that played a crucial role in the design of the Reference Model as well as in the design of the OntoTag ontologies and the DCR. More interesting, however, are the differences, which reflect different orientations of the ontologies. Those concepts that were missing in the Reference Model were either language-specific (OntoTag) or were not considered in either EAGLES or in the annotation schemes relevant to the sustainability project. The Reference Model concepts that did not find a counterpart in OntoTag or the DCR mostly originated in typologically-oriented annotation schemes, annotation schemes used by historical linguists, or the Russian Uppsala tag set, indicating that OntoTag and the DCR seem to have a stronger focus on Western European languages, or more generally, languages for which a broader range of linguistic tools exists.

5 Summary and discussion

In this paper, I have described design principles and implementation of a structured ontology of linguistic annotation. It is currently applied for purposes of annotation documentation, tag-set neutral corpus search and can also be applied in NLP contexts.

For the ontology, sustainability considerations entail the premise to preserve and to systemize existing annotations and relevant annotation documentation. In line with this conservation perspective, a structured ontology was developed which involves several self-contained ontologies, which are linked in a declarative way. Hence, a clear separation between the information drawn from the annotation documentation and its interpretation with respect to the reference terminology is established, as required by the ethics of conservation:

> The principal goal should be the stabilisation of the object or specimen. All conservation procedures should be documented and as reversible as possible, and all alterations should be clearly distinguishable from the original object or specimen. (ICOM, 2006, §2.24).

The structured ontology consists of a Reference Model specifying conventional linguistic terminology, and several Annotation Models, each representing a formalization of the annotation documentation of a given annotation scheme. Both Reference Model and the respective Annotation Models are self-contained ontologies. Between these,

however, a linking is specified which describes any Annotation Model concept in terms of the Reference Model.

As compared to related approaches, which operate on the direct mapping of annotations to an ontology of reference terms, e.g. Farrar and Langendoen (2003), de Cea et al. (2004), this structured ontology involves a high level of redundancy. The modular representation of Reference Model and Annotation Model, however, allows to view Annotation Models as a form of annotation documentation, as annotation-relevant comments are clearly separated from interpretation-relevant comments. In particular, these annotation-relevant comments are supposed to cover excerpts and examples from the original documentation which provide an informal, non-ontological definition and description of the respective concepts and properties. Also, a hypertext visualization of Annotation Models, the Reference Model and the linking has been implemented which allows a user to assess both the ontological information and these comments and thus, use the ontology as a key to annotation documentation.

Further, this modular structure is highly flexible, as it allows a user to replace any component of the system by his own specifications, that is, the linking may be altered independently from the participating Reference and Annotation Models. Similarly, an Annotation Model may be exchanged. Further, this design supports an open, extensible architecture, that is, new Annotation Models can be developed and linked to the Reference Model. Finally, a non-redundant ontological representation can be automatically retrieved from the structured ontology by unifying concepts from the Reference Model with the Annotation Model concepts that are defined as sub-concepts in the linking.

The Reference Model itself may be linked by the same mechanism to external Reference Models of linguistic terminology in general. Such external Reference Models may evolve from approaches like Farrar and Langendoen (2003) or Schneider (2007). These external Reference Models, then, must not be related to any existing Annotation Model, but instead, the linking with the Annotation Models is mediated by the (internal) Reference Model.

So far, three external Reference Models have been linked to the internal Reference Model, i.e. GOLD, the morphosyntactic component of de Cea et al. (2004)'s OntoTag ontologies, and an ontological representation of Ide et al. (2005)'s Data Category Registry. This is particularly interesting for the application of ontologies to the formulation of annotation-independent corpus queries, i.e. expressions formulated in terms of external Reference Models can be translated into queries for specific annotations on the basis of the linkings with the internal Reference Model and the Annotation Models. The internal Reference Model thus represents an *interface* to the Annotation Models, and the annotations.

Currently, Annotation Models pertaining parts of speech and morphosyntax for

German, English and Russian have been implemented. Also, Annotation Models for a typologically oriented annotation scheme has been developed, that applies not only to parts of speech and morphosyntactic annotation, but also to glossing, syntactic phrases and information structure in a broad variety of languages. Finally, several project-specific Annotation Models relevant to the CRCs (concerning historic linguistics, typological research and first language acquisition) have been created.

From these, the Annotation Models pertaining German, Russian and English are particularly relevant to text technology, as these tag sets are also used by existing tools and thus, these ontologies can be used to support the tag-set independent interpretation of automatically derived linguistic analyses. More precisely, the ontology-based approach presents a natural handling of underspecification, and by exploiting this information, the robustness of linguistic analyses in technical contexts may be improved.

References

Brants, S. and Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus. In *Proc. 3rd Conference on Language Resources and Evaluation (LREC-02)*, Las Palmas de Gran Canaria, Spain.

Broschart, J. (1997). Why Tongan does it differently: Categorial distinctions in a language without nouns and verbs. *Linguistic Typology*, 1-2:123–166.

Chiarcos, C. (2006). An ontology for heterogeneous data collections. In *Proc. Corpus Linguistics 2006*, pages 373–380, St.-Petersburg. St.-Petersburg University Press.

Cimiano, P. and Reyle, U. (2003). Ontology-based semantic construction, underspecification and disambiguation. In *Proc. Prospects and Advances in the Syntax-Semantic Interface Workshop*.

de Cea, G. A., Gómez-Pérez, A., Álvarez de Mon, I., and Pareja-Lora, A. (2004). OntoTag's linguistic ontologies. In *Proc. Int'l Conference on Information Technology, Coding and Computing (ITCC'04)*, pages 124–128, Las Vegas, Nevada.

Farrar, S. and Langendoen, D. T. (2003). A linguistic ontology for the semantic web. *GLOT International*, 7(3):97–100.

Hughes, J., Souter, C., and Atwell, E. (1995). Automatic extraction of tag set mappings from parallel annotated corpora. In *From Text to Tags: Issues in Multilingual Language Analysis, Proc. ACL-SIGDAT Workshop*, pages 10–17.

ICOM (2006). ICOM code of ethics for museums. In Hoffman, B. T., editor, *Art and Cultural Heritage. Law, Policy and Practice*. Cambridge University Press.

Ide, N., Romary, L., and de la Clergeri, E. (2005). International standard for a linguistic annotation framework. In *Proc. HLT-NAACL'03 Workshop Software Engineering and Architecture of Language Technology*.

König, E., Bakker, D., Dahl, e., Haspelmath, M., Koptjevskaja-Tamm, M., Lehmann, C., and Siewierska, A. (1993). EUROTYP Guidelines. Technical report, European Science Foundation Programme in Language Typology.

Leech, G. and Wilson, A. (1996). EAGLES recommendations for the morphosyntactic annotation of corpora. Technical report, Expert Advisory Group on Language Engineering Standards.

Lezius, W., Rapp, R., and Wettler, M. (1998). A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German. In *Proc. COLING-ACL 1998*, pages 743–747.

Monachini, M., Soria, C., and Ulivieri, M. (2005). Evaluation of existing standards for NLP lexica. draft 1.1. Technical report, LIRICS (Linguistic Infrastructure for Interoperable Resource and Systems).

Rehm, G., Eckart, R., and Chiarcos, C. (2007). An OWL- and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In *Proc. RANLP 2007: Recent Advances in Natural Language Processing*. Borovets, Bulgaria.

Sampson, G. (1995). *English for the Computer*. Clarendon Press, Oxford.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines fur das Tagging deutscher Textkorpora mit STTS. Technical report, University of Stuttgart and Universitat of Tübingen.

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

Schneider, R. (2007). A database-driven ontology for German grammar. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Data Structures for Linguistic Resources and Applications*, pages 305–314, Tübingen. Narr.

Stede, M. (2004). The Potsdam Commentary Corpus. In *Proc. ACL-04 Workshop on Discourse Annotation*, pages 96–102, Barcelona.

Maja Bärenfänger, Mirco Hilbert, Henning Lobin, Harald Lüngen

OWL ontologies as a resource for discourse parsing

1 Introduction

In the project SemDok (*Generic document structures in linearly organised texts*) funded by the German Research Foundation DFG, a discourse parser for a complex type (scientific articles by example), is being developed. Discourse parsing (henceforth DP) according to the Rhetorical Structure Theory (RST) (Mann and Taboada, 2005; Marcu, 2000) deals with automatically assigning a text a tree structure in which discourse segments and rhetorical relations between them are marked, such as Concession. For identifying the combinable segments, declarative rules are employed, which describe linguistic and structural cues and constraints about possible combinations by referring to different XML annotation layers of the input text, and external knowledge bases such as a discourse marker lexicon, a lexico-semantic ontology (later to be combined with a domain ontology), and an ontology of rhetorical relations. In our text-technological environment, the obvious choice of formalism to represent such ontologies is OWL (Smith et al., 2004). In this paper, we describe two OWL ontologies and how they are consulted from the discourse parser to solve certain tasks within DP. The first ontology is a taxononomy of rhetorical relations which was developed in the project. The second one is an OWL version of GermaNet, the model of which we designed together with our project partners.

2 Taxonomies of rhetorical relations

Already in the original conception of Rhetorical Structure Theory by Mann and Thompson (1988), (see also Mann and Taboada, 2005), rhetorical relations were grouped into classes. On a top level, there were the two groups of *multinuclear* vs. *mononuclear* relations according to the structural criterion of nuclearity. The mononuclear relations were further subdivided into *presentational* vs. *subject-matter relations* (cf. Mann and Taboada, 2005). Lower-level subgroups such as *Evidence-and-Justify* were introduced as well. The complete hierarchy is shown in Figure 1.

Hovy and Maier (1995) suggested a merger of existing hierarchies of discourse relations into one comprehensive hierarchy consisting of 65 relation categories, 43 of which were relations at the base level. Their prediction was that application-specific extensions to this merged relation set would always consist in the refinement of a relation category that was already in the hierarchy, i.e. the number of higher-level

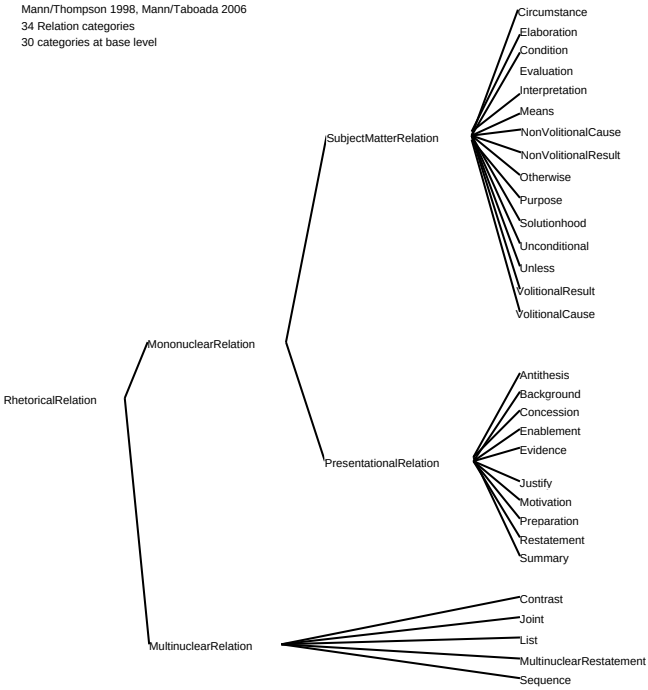


Figure 1: Hierarchy of rhetorical relations according to Mann and Thompson (1988)

relation types would always stay the same. One purpose of developing a hierarchy of discourse relations is thus to point out similarities of different relation sets by showing how they can be mapped on each other or even merged, ultimately supporting the view that a universal set of relation types exists. This hierarchy can be seen in Figure 2

In the present project, we produced corpus annotations using the original RST relation set proposed in Mann and Taboada (2005), and by an examination of these annotations and an inspection of alternative relation sets proposed in the literature (notably Carlson and Marcu (2001) and Hovy and Maier (1995)), we designed a relation hierarchy suitable for annotating the rhetorical structure of scientific journal articles in our explorative reading scenario (Lüngen et al., 2006). It consists of 70 relation types, 44 of which are basic categories in the hierarchy.

Though it seems natural to model rhetorical *relations* as OWL *properties* (<`owl:ObjectProperty`>) as we proposed in an earlier publication (Goecke et al., 2005), we finally refrained

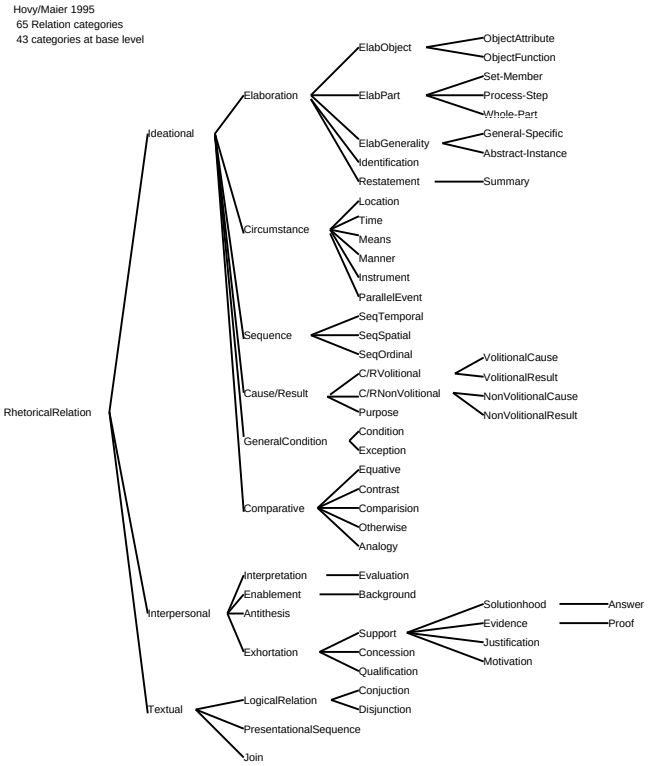


Figure 2: Hierarchy of discourse relations according to Hovy and Maier (1995)

from doing so, because we also wanted to view the properties as classes to declare disjointness between certain rhetorical relation types and to encode properties of rhetorical relations that would be inherited by their subrelations. Within OWL DL, properties can be arranged in a hierarchy but cannot be declared classes at the same time (Smith et al., 2004).[1] Thus we modelled the rhetorical relations as OWL classes, which is not so devious if one considers that it is sometimes recommended as good practice to introduce a "relation class" for the encoding of an n-ary relation in OWL (cf. Noy et al., 2006). Subrelation-hood is then marked by the <rdfs:subclassOf> construct. The use of <rdfs:subclassOf> also enabled us to include further features

[1]Since most OWL reasoners and inference tools apply to the sublanguage OWL DL, we encode our ontologies within OWL DL.

in the formalisation of our hierarchy: We introduced heavily underspecified relation classes such as MononuclearRelation, and we cross-classified all relations along the two dimensions *nuclearity* and *metafunction*, giving rise to multiple inheritance. For example, Support is both a subclass of InterpersonalRelation as well as of MononuclearRelation. (For reasons of decipherability, the links from MononuclearRelation and MultinuclearRelation are not shown in Figure 3, though.) We introduced further sub- or superrelations, when it was expedient according to our corpus analyses and with respect to our scenario (cf. Lüngen et al., 2006). The resulting hierarchy is shown in Figure 3. This "RRSet ontology" is used to combine competing hypothesis during the parsing process as described in Sect. 4.

3 Using a GermaNet-based Ontology for the automatic assignment of Elaboration

One of the most prominent RST relations in our corpus is Elaboration - it is the second most frequent relation of all. Unlike other RST relations, Elaboration is seldom signalled by syntactic or lexical discourse markers. To tackle its automatic identification and annotation, we examined instances of Elaboration in our corpus and reviewed the treatment of Elaboration in previous approaches to discourse analysis (e.g. Carlson and Marcu, 2001; Hovy and Maier, 1995; Knott et al., 2001). This led us to distinguish the different subtypes of Elaboration relations which can be seen in the taxonomy of rhetorical relations in Figure 4.

The subtaxonomy of Elaboration relations organises the subcases that can trigger different types of rhetorical links between text modules of scientific articles in our explorative reading scenario. Each subrelation has its own definition and is associated with a different set of discourse markers and linguistic or structural cues that signal it. Elaboration-definition, for example, can be determined by cues from the logical document structure (e.g. <doc:glosslist>), Elaboration-example is often signalled by the lexical discourse markers "z.B.", "Beispiel", or "beispielsweise"), whereas the subtypes of Elaboration-specification are induced by syntactic and punctuational discourse markers (e.g. a non-sentential phrase within parentheses).

However, the majority of Elaboration subtypes is not indicated by discourse markers or structural cues, but may be established by the presence of lexical-semantic relations between the central discourse entities of two discourse segments. Elaboration-derivation is signalled by conceptual relations like hyperonymy/ hyponymy, holonymy or meronymy, and lexical relations like synonymy or pertainymy indicate Elaboration-continuation, or Elaboration-restatement. Figures 5 and 6 show how holonymy (*Deutschland – Süddeutschland, Norddeutschland*) induces Elaboration-derivation, and pertainymy (*Automatisierung – automatisiert*) Elaboration-drift.

For the automatic identification of these subtypes there are two options: 1. Lexical-semantic relations may be identified in the discourse parser by performing a lookup

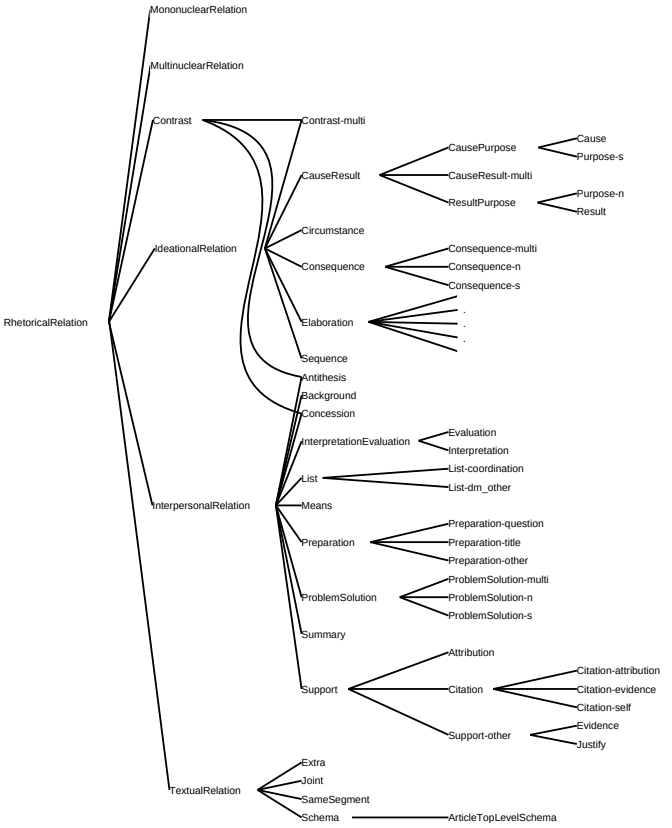


Figure 3: SemDok RRSET ontology (save the subclasses of ELABORATION)

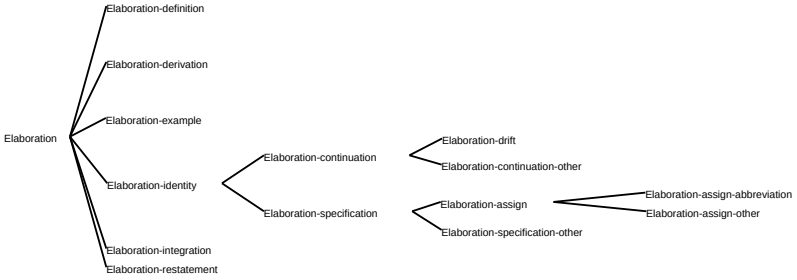


Figure 4: SemDok hierarchy of ELABORATION relations)

in an OWL version of the lexico-semantic net GermaNet (Kunze et al., 2007). In this approach, GermaNet is directly consulted from the parser. 2. Lexical-semantic relations may be calculated in auxiliary components and be made available to the parser in the form of additional annotation layers of the input text. As auxiliary components, we envisage a lexical chainer and/or an anaphora resolution component as developed in out partner projects HyTex (Holler et al., 2004), and Sekimo (Goecke et al., this volume). As the coverage of our corpus by GermaNet 5.0 seems not high enough for a direct approach – 30.74% of all noun tokens and 59.17% of all noun types in our corpus are not contained in GermaNet – we will first focus on the second option.

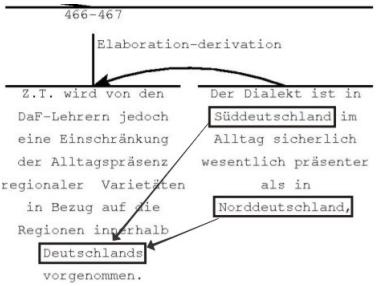


Figure 5: Holonymy as a cue for ELABORATION-DERIVATION

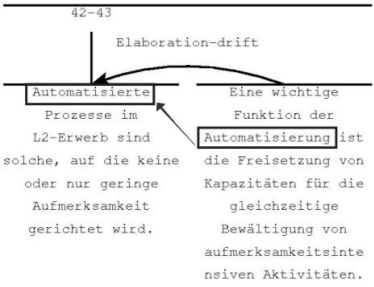


Figure 6: Pertainymy as a cue for ELABORATION-DRIFT

4 Generalised utilisation of OWL ontologies in the GAP

We consider the process of DP as an iterative application of a more general parser architecture which accepts different annotation layers as input data and produces a new annotation layer as its output, see Figure 7. In each of the consecutive instantiations of the so-called *Generalised Annotation Parser* (GAP), a different set of resources is employed to control it.

The core of the GAP is a bottom-up passive chart parser, implemented in Prolog. It takes the primary textual data and their n XML annotation layers as its input, which are first converted to a Prolog fact base. The behaviour of the parser is controlled by a set of application-dependent reduce rules formulated in XML, which, for the most part, are derived from a discourse marker lexicon. The conditions of their application are expressed as declarative constraints between the $n+1$ annotation layers. The conditions for several subcases of ELABORATION relations expressed in Sect. 3, for example, are formulated as reduce rules. The reduce rules set is converted to Prolog, so that they can directly be used by the chart parser. The constraints that are part of the reduce rules make

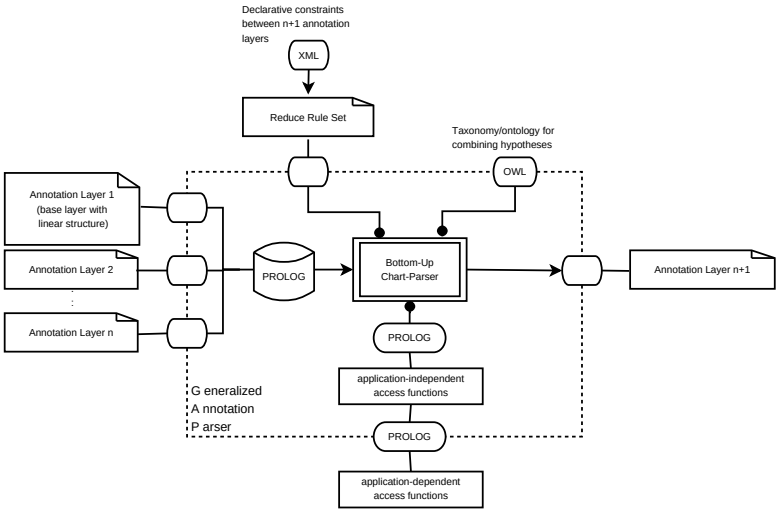


Figure 7: Generalised Annotation Parser GAP

use of access predicates which express connections between different annotation layers. The set of access predicates can be divided into application-independent ones, such as `identity`($layer_i$`:`$element_x$`,` $layer_j$`:`$element_y$`)` or `text-inclusion`($textvalue$`,` $layer_i$`:`$element_x$`)`, and application-dependent ones which can refer to the schema information of annotation layers.

As in most parsing applications, it can happen that more than one reduce rule is applicable in a reduce step. Such situations depend on the one hand on the reduce rule set and on the other hand on the structure of the input annotation layers, specifically, when there are ambiguous discourse markers, such as the German conjunction *aber*, which, similar to English "but" can signal CONCESSION or CONSTRAST-MULTI (cf. Figures 8 and 9[2]). If such an ambiguity cannot be resolved e.g. because no further, supporting discourse markers are present, it leads to competing hypotheses about the combination of segments and therefore to a *set* of possible output annotation hierarchies (two in case of the example). This has two types of negative consequences:

1. In a bottom-up parsing approach, which is mostly taken in DP, the number of alternatives that have to be pursued increases, thus reducing parsing efficiency in terms of time and memory.

[2]Text segments in Figures 5, 8, and 9 are from Baßler and Spiekermann (2001); text segments in Figure 6 are from Bärenfänger and Beyer (2001).

2. When parsing results are evaluated against reference annotations of discourse structures, the non-matching hypothesis will count as an ordinary recall error, although CONSTRAST-MULTI and CONCESSION are semantically quite similar.

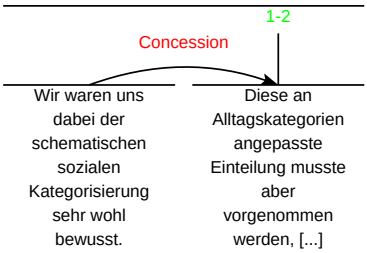


Figure 8: *aber* signalling CONCESSION

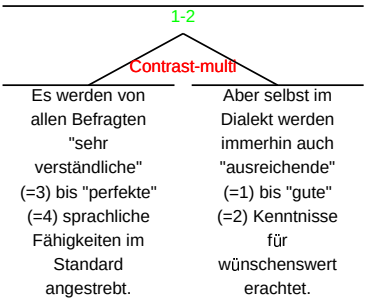


Figure 9: *aber* signalling CONTRAST-MULTI

Besides introducing local ambiguity packing (Tomita, 1987), the first situation can be remedied by replacing the two hypotheses by one hypothesis with the label of the lowest common superordinate relation according to the RRSET hierarchy, which is CONTRAST in the the example. Such a combination rule can be derived from the OWL *subclassOf* property that holds between classes of an application-dependent OWL DL ontology. Whenever two or more competing hypotheses about relation instances have been emitted in the parsing process, the parser consults the RRSET ontology (Sect. 2) and check whether the n relation names of the competing hypotheses have one or more lowest common superclasses within a certain range, e.g. within a so-called *reduced relation set*. For each lowest common superclass found, the hypotheses are merged into one, and the superclass is taken as the relation label of the new hypothesis, representing an underspecified relation instance.

In the second situation, in order to differentiate between hard-core recall errors and those caused by semantically similar relations that have been recognised at the same time, an additional evaluation can be conducted where relation labels in the parsing results as well as in the reference annotations are first replaced by labels from a reduced relation set, as e.g. done in Soricut and Marcu (2003). Such a replacement can also be effected by a look-up in the RRSET ontology.

Like the OWL ontology of GermaNet, the RRSet ontology is converted to Prolog and consulted by the parser using the Thea OWL Library for Prolog (Vassiliadis, 2006), which in turn uses the SWI-Prolog's Semantic Web library[3]

[3] `http://www.swi-prolog.org/`

5 Conclusion

In this article, we sketched the SemDok RRSet relation taxonomy for rhetorical relations in scientific journal articles which was designed based on corpus investigations and previously proposed hierarchies of discourse relations. We described how it was coded in the Web Ontology Language OWL, and how the OWL-based ontology will be consulted as a knowledge base by a discourse parser. As a second example of the utilisation of ontologies in discourse parsing, methods to identify subtypes of the ELABORATION relation using an OWL version of the lexico-semantic net GermaNet were described.

In the GAP, local ambiguity packing is currently employed rather than looking up the RRSET ontology during parsing. However, the RRSET will be used in the evaluation of parsing results as described, and is also used to generate a relations file for manual annotations of discourse structures using O'Donnells RSTTool (O'Donnell, 2000). As for the identification of ELABORATION relations, a comprehensive approach analysing annotations of anaphora and lexical chains is pursued.

References

Bärenfänger, O. and Beyer, S. (2001). Zur Funktion der mündlichen L2-Produktion und zu den damit verbundene kognitiven Prozessesn für den Erwerb der fremdsprachlichen Sprechfertigkeit. *Linguistik Online*, 8. `http://www.linguistik-online.de`.

Baßler, H. and Spiekermann, H. (2001). Dialekt und Standardsprache im DaF-Unterricht. Wie Schüler urteilen - wie Lehrer urteilen. *Linguistik Online*, 9. `http://www. linguistik-online.de`.

Carlson, L. and Marcu, D. (2001). Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA. ISI-TR-545.

Goecke, D., Lüngen, H., Sasaki, F., Witt, A., and Farrar, S. (2005). GOLD and discourse: Domain- and community-specific extensions. In *Proceedings of the 2005 E-MELD-Workshop*, Boston, MA.

Holler, A., Maas, J.-F., and Storrer, A. (2004). Exploiting coreference annotations for text-to-hypertext conversion. In *Proceeding of LREC*, volume II, pages 651–654. Lisboa.

Hovy, E. and Maier, E. (1995). Parsimonious or profligate: How many and which discourse structure relations? Unpublished paper, `http://www.isi.edu/natural-language/people/hovy/publications.html`.

Knott, A., Oberlander, J., O'Donnell, M., and Mellish, C. (2001). Beyond elaboration: The interaction of relations and focus in coherent text. In Sanders, T., Schilperoord, J., and Spooren, W., editors, *Text representation: Linguistic and psycholinguistic aspects*, volume 8 of *Human Cognitive Processing*, pages 181–196. Benjamins, Amsterdam.

Kunze, C., Lemnitzer, L., Lüngen, H., and Storre, A. (2007). Repräsentation und verknüpfung allgemeinsprachlicher und terminologischer wortnetze in owl. *Zeitschrift für Sprachwissenschaft*. To appear.

Lüngen, H., Lobin, H., Bärenfänger, M., Hilbert, M., and Puskàs, C. (2006). Text parsing of a complex genre. In *Proceedings of the Conference on Electronic Publishing (ELPUB)*, pages 247–256, Bansko, Bulgaria.

Mann, W. C. and Taboada, M. (2005). RST – Rhetorical Structure Theory. W3C page. `http://www.sfu.ca/rst`.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organisation. *Text*, 8(3):243–281.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, MA.

Noy, N., Rector, A., and (eds.) (2006). Defining n-ary relations on the semantic web. Technical report, W3C Working Group Note. `http://www.w3.org/TR/swbp-n-aryRelations`.

O'Donnell, M. (2000). RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In *Proceedings of the International Natural Language Generation Conference (INLG'2000)*, pages 253 – 256, Mitzpe Ramon, Israel.

Smith, M. K., Welty, C., McGuiness, D. L., and (eds.) (2004). OWL Web Ontology Language guide. Technical report, W3C recommendation. `http://www.w3.org/TR/2004/REC-owl-guide-20040210`.

Soricut, R. and Marcu, D. (2003). Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL)*, Edmonton, Canada.

Tomita, M. (1987). An efficient augmented-context-free parsing algorithm. *Computational Linguistics*, 13(1-2):31–46.

Vassiliadis, V. (2006). Thea. A web ontology language - OWL library for [SWI] Prolog. Web-published manual, `http://www.semanticweb.gr/TheaOWLLib/index.htm`, visited 15.7.2006.

A Kumaran, Ranbeer Makin, Vijay Pattisapu, Shaik Sharif, Lucy Vanderwende

Evaluating the Quality of Automatically Extracted Synonymy Information

Automatic extraction of semantic information, if successful, offers to languages with little or poor resources, the prospects of creating ontological resources inexpensively, thus providing support for common-sense reasoning applications in those languages. In this paper we explore the automatic extraction of synonymy information from large corpora using two complementary techniques: a generic broad-coverage parser for generation of bits of semantic information, and their synthesis into sets of synonyms using automatic sense-disambiguation. To validate the quality of the synonymy information thus extracted, we experiment with English, where appropriate semantic resources are already available. We cull synonymy information from a large corpus and compare it against synonymy information available in several standard sources. We present the results of our methodology, both quantitatively and qualitatively, that indicate good quality synonymy information may be extracted automatically from large corpora using the proposed methodology.

1 Introduction

Automatic extraction of semantic information, if successful, will prove to be invaluable for languages with little or poor resources in the way of dictionaries, thesauri, etc. It opens up an unprecedented level of access to obscure and under-represented languages by enabling such projects as automated compilation of lexica, content organization, and multilingual information retrieval. In this project, we explore the automatic construction of synonymy information from large corpora, using two complementary techniques: a generic broad-coverage parsing for generation of bits of semantic information and their synthesis into sets of synonyms based on automatic sense-disambiguation methodologies. To validate the quality of the synonymy information extracted by our methodology, we experiment first with English, where appropriate semantic resources are already available as reference. We cull synonymy information from a large corpus, and compare it against the synonymy information available in multiple sources, specifically, the Oxford English Dictionary (14) and WordNet (4).

We first present a *naive* approach, where we assemble sets of synonyms under the assumption of transitive synonymy. While the quantitative and qualitative analysis

of synonym sets thus constructed present an endemic problem of semantic drift, we present a solution methodology based on sense disambiguation to synthesize better quality synonym sets. Finally, we present some quantitative and qualitative evaluations of the results of our refined approach, including, inter alia, comparisons with our naïve approach and WordNet data, as well as discussion of possibilities for this technique.

2 Automatic Synonym Extraction from Large Corpora

In this section, we provide a brief description of the two large resources that we used in our experimentation, i.e., WordNet and MindNet.

2.1 WordNet

The Princeton WordNet (4) is a manually constructed lexical database organized by word meanings (as opposed to word forms, as in a dictionary). A part of WordNet, namely, the noun synonyms, resembles a thesaurus. Its hierarchical semantic structure describes hypernymy/hyponymy, holonymy/meronymy, and synonymy/antonymy between words. Different word senses are addressed by writing multiple, enumerated lexical entries (they are effectively treated as if they were different words). WordNet is being used as a lexical knowledge base in a wide variety of information retrieval (IR) applications. Since WordNet is hand-crafted, it is thorough, expensive and unique. It is thorough because it has been created by professional lexicographers; specifically, (4) states that "in terms of coverage, WordNet's goals differ little from those of a good standard college-level dictionary". It is expensive, having taken decades to compile for English alone. WordNets for other languages have been and are being compiled (6), but are available primarily in Western European (3) languages, and even then in most languages, not as complete as the English WordNet. Given the time and resources needed to develop WordNet in a language, it may be a daunting task for most languages of the world, which are constrained by economic resources, market potential, or linguistic expertise.

2.2 MindNet

MindNet is an automatic ontology extraction system for unrestricted text in English (23) (16) that has also been successfully adapted to Japanese (21) as well. MindNet builds a logical form graph for each sentence using a broad-coverage parser, and extracts semantic relationships among words in that sentence. Such extracted knowledge is accumulated in MindNet, from which all semantic relationships between two words may be explored explicitly through an explorer interface[1]. The corpora we use for extracting

[1] An online explorer of dictionary-based MindNet is available at http://research.microsoft.com/mnex/.

semantic information are two machine-readable dictionaries (MRDs), The American Heritage Dictionary, 3rd ed. (AHD) and Longman's Dictionary of Contemporary English (LDOCE). Although MindNet can be used with any corpus, we use MRDs in order to produce optimal output for constructing inferences.

For the sake of discussion in the following sections, it is important to emphasize the following two caveats: First, the relationships extrated by MindNet hold between words, where as WordNet is organized by the word senses. Second, for extracting synonymy information, it has been shown that simpler pattern matching techniques may perform well, in (2) and (7). However, we use a broad-coverage parser, due to its ready availability and due to our goal of ultimately extracting all types of relationships (12).

3 Naive Approach

First, we compiled all the synonymy relationships MindNet extracted from the MRDs. This compilation consists of expressions of the form "*A syn B*", essentially encoding the fact that A and B are synonymous in some context. From this we synthesized a set of synsets, wherein if "*A syn B*" and "*B syn C*" were found in the extracted expressions, then A, B, and C are put into the same synset (i.e., it is inferred that "*A syn C*"). The naïve approach is thus characterized by transitive synonymy any set of nodes connected transitively to each other are grouped into the same synset. In addition to syn relationships, we incorporated nodes from the hypernymy/hyponymy relations output of MindNet, in order to cover those WordNet leaf synsets that are primarily singletons. In the following two sections we analyze the quality of such synthesis of synsets.

3.1 Quantitative Evaluation of the Naïve Approach

The naïve approach, working on the *syn* and *hyp* relationships extracted from AHD and LDOCE, produced 49,693 synsets. Figure 1 shows a quantitative comparison of synsets formed by MindNet, with those of WordNet.

CHARACTERISTICS	WORDNET	MINDNET
Total Words	117,097	63,230
Total Synsets	81,426	49,693
Avg Words/Synset	1.78	1.27
Avg Synsets/Word	1.24	1.00

Figure 1: Comparison of WordNet and MindNet Synsets

We observe that we have only a little more than half as many words and synsets as WordNet, possibly due to the limited extent of corpora that was analyzed by MindNet, resulting in far less number of words for which syn information is extracted. We believe that extracting from larger and more diverse corpora might alleviate this relative shortcoming. We see that the synonymy relationships are markedly richer in WordNet, as indicated by higher averages of words in synset and vice-versa. This is an unsolvable shortcoming of our naïve approach, as a given word could be a part of only one synset, whereas in WordNet, a polysemous word is common to several synsets. Hence, our synthesis must be enhanced to account for polysemous words (which is addressed in the next section). Our subsequent analysis in this section focuses on the quality of the synsets thus extracted by our naïve method, and not on quantity.

3.2 Qualitative Evaluation of the Naïve Approach

We first analyzed the distribution of sizes of the extracted synsets; as shown in Figure 2, we find that the majority (94%) of the synsets were singletons, produced primarily by the hyp relationships, for which there were no corresponding syn relationships available. Comparatively, 57% of the WordNet synsets are singletons.

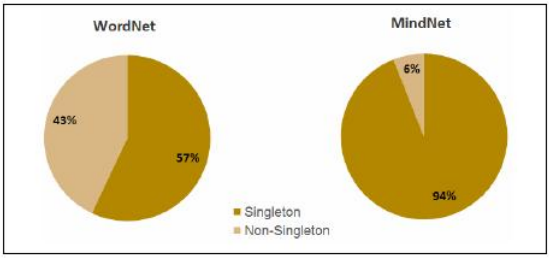


Figure 2: Comparison of WordNet and MindNet Synset Sizes

The disparity in proportion of singleton synsets between the two can be due to a variety of reasons. Part of the explanation is the obvious fact that automatic extraction of lexical information underperforms manual construction of it. Another is that WordNet covers many more words than MindNet's source MRD's. Since the singletons synsets are good, by definition, we examined for quality the remaining 6% (2,882) non-singleton sets, in the subsequent analysis.

We manually inspected the output synsets of the naïve approach against the Oxford English Dictionary (OED) and the two source machine readable dictionaries. Checking against the OED, an independent source when compared with AHD and LDOCE, has several advantages; first, it prevents artificially high results from using the same corpus

as both an input corpus and output test (19). In addition, it adds insight into the variability of dictionaries, as large differences were observed, when the synsets gleaned from AHD + LDOCE were cross-verified with OED, while only minimal variations were expected. Such differences may reveal importance of corpus choice in automatic ontology extraction; though we aim to be able to handle unrestricted text, we are also interested in exploring the implication of corpus choices on the output quality. The motivation for manually checking the synset output against the AHD and LDOCE is primarily to evaluate the global and local performance of the logical forms produced by broad-coverage parser; that is, while the synonymy information captured in the logical forms produced from a single definition in AHD or LDOCE is expected to be correct, we also wish to verify the global consistency of the synonymy information gleaned from logical forms produced by parsing multiple definitions for a set of related words.

For this manual qualitative analysis, we distinguish between *well-formed* or *ill-formed* synsets, which refer only to the quality of the synsets, and not to the quality of the MindNet data. Our criteria for whether a synset is well- or ill-formed is an approximation of lexicographers' consensus via manually checking the output against a variety of lexical resources: OED, AHD, LDOCE, and WordNet. Essentially, a synset is classified as well-formed, if each pair of the words from that synset are synonymous, when checked against OED, AHD, LDOCE and WordNet. By this method, we found that about 87% (2,517) of the extracted synsets were well-formed, and the rest were ill-formed. Next, we analyzed manually all the ill-formed synsets and classified them into different categories, depending on the reasons for their ill-formedness; Figure 3 gives a classification of these ill-formed synsets.

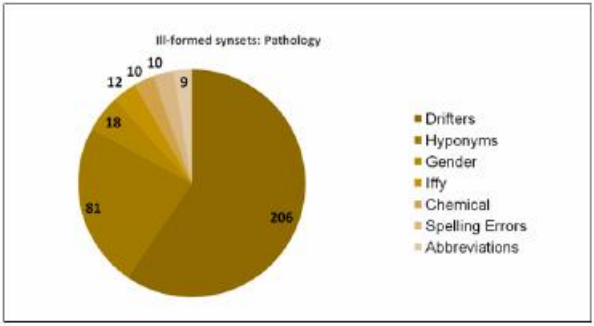


Figure 3: Pathology of MindNet Synthesized Synsets

The drifters form the majority of the ill-formed synsets (206 synsets, constituting of nearly 60% of the ill-formed synsets). Drifters are synsets like { *board, committee, plank*

} that spread across more than one consistent semantic space. In the above example the ill-formed synset, { *board, committee, plank* } contains two different semantic spaces, namely, { *board, committee* } and { *board, plank* }. If a synset contains at least one pair of words that are not synonymous, but included due to transitive synonymy, then that synset is classed as a drifter. In our naïve extraction, one pathological case of the drifter synset had nearly 9,500 entries.

In addition to drifters, we find several other classes of problems. Nouns, for example, that appear in coordinate phrases and that can be parsed as either a noun or an adjective, parse preferentially as nouns, resulting often in incorrectly synthesized synsets; for example, the definition of a dictionary entry for "*calculus*" as, "*differential or integral calculus*", incorrectly yields "*differential syn integral calculus*". Such wrong parses resulted in hyponyms and hypernyms grouped in a synset, and accounted for nearly 23% of the errors (tagged as Hyponyms, in Figure 2). Some idiosyncratic, but simpler to correct, parse errors involved chemical names, in which the presence of paran-theses lead to wrong parses; for example, the entry for lead arsenate, whose empirical formula is $Pb_3(AsO_4)_2$, results in a synset { Pb_3, *lead_arsenate*, AsO_4, *azurite*, *erythrite*}, along with that of azurite and erythrite, which share chemical sub-structures with lead arsenate. About 3% of the synsets had similar misclassifications. Gender (5% of the wrongly classified synsets) denotes gender antonyms like { *actor, actress* }, but were classified together, perhaps because they were provided as examples for a hypernym, say, an artist. Though these pairs fail Leibniz's Substitution Principle, certain dictionaries' entries support their synonymy. Iffy (3% of the misclassified synsets) contains near-synonyms whose validity or invalidity is hard to assert. An example of an "abbreviation" (3% of the misclassified synsets) error are synsets such as, { *nm, nanometer, nuclear_magneton* }, where the same abbreviation for two different entity played a role in all of them getting clubbed together. Spelling (3% of the misclassified synsets) errors are all due to typographic errors in the corpus.

Clearly, drifters are a major problem synthesis of correct synonymy information, and it is clear that the primary reason for their inclusion is the lack of disambiguation between the senses of a word, as MindNet output consists of only words and not word senses. So, the next part of our research focussed on synthesizing the synsets with sense-disambiguation.

4 Latent Semantic Analysis

In this section, we focus on the Word-sense Disambiguation (WSD) step and how it can filter extracted synonymies into correct synsets. We do this WSD filtering with Latent Semantic Analysis (LSA), a statistical method of assessing words' semantic contexts (11) (1). First, we construct a word-by-document matrix for a large text corpus. Next, because of this matrix's sparseness, we extract its principal vectors via singular

value decomposition. Finally, we use this information to test the putative synonym pairs provided by MindNet: if the cosine similarity of their vectors is greater than or equal to a threshold, then they are joined into a synset. We hypothesize that the word neighborhood of plank will differ sufficiently from the word neighborhood of committee so that LSA can thereby "read" two senses of the word board. These two approaches – broad-coverage parsing and Latent Semantic Analysis – are complementary modules in that the former's syntactic approach is blind to parasentential patterns, whereas the latter's "*bag-of-words*" approach is largely blind to intrasentential patterns. In this experiment, we ran the extraction-side on the two machine-readable dictionaries already mentioned, AHD and LDOCE. In the first set of experiments, we used that any two co-occurring words in the same document are considered associated semantically. In the subsequent analysis, we tighten this assumption, by considering only a window of n words, to form semantic associations.

5 Quantitative Evaluation of Synthesized Synsets

We performed LSA on the Brown Corpus (9) to extract a 15-dimensional words space for computing similarity. The Brown corpus consists of about 1 Million words, 40,897 unique words distributed among 500 documents. The average words per document is about 2,000, indicating fairly large documents.

We used cosine similarity measure between two words to distinguish their senses, and we used threshold values between 0.8 and 0.95 to empirically study the impact of the threshold on the formation of good synsets. A threshold of 1.0 yields a degenerate solution of cleaving every synset into singletons, and hence was not considered for analysis. While a lower threshold left most good synsets intact, a higher threshold disassociated runaway and loose synsets, creating smaller units, though possibly cleaving even some of the good ones. The words covered in these synsets are exactly the same as those presented earlier in the naïve approach, but they are, understandably, organized differently.

First, we note that the number of non-singleton synsets went from nearly 2,800 to nearly 17,000, indicating that a number of large runaway synsets were broken into smaller synsets. We observe, in Figure 4, that the average words per synset (of non-singleton synsets) decreases with the threshold parameter, indicating that the synsets are getting smaller and presumably tighter (an analysis to verify this is provided in a later section).

We compared these synsets with WordNet synsets, whether the synthesized synset is identical, superset or subset of WordNet synset, purely based on the words of the synsets, and the results are shown in Figure 5.

We observe that the number of identical synsets between WordNet and MindNet increases, indicating that the LSA analysis help in building semantically tighter synsets.

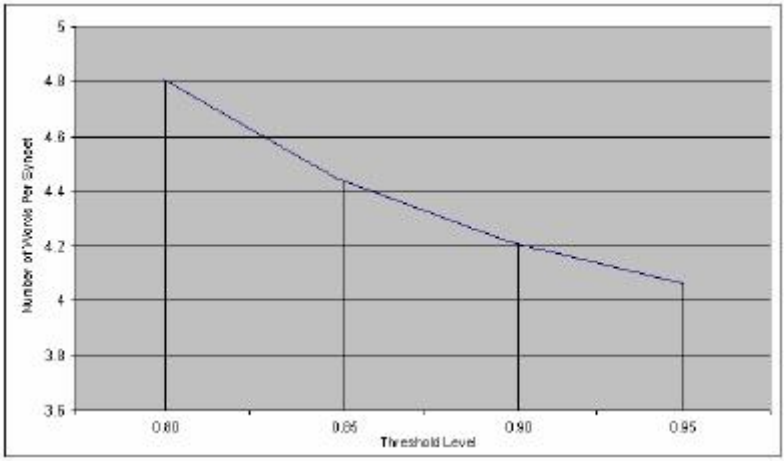


Figure 4: Average Sizes of Mindnet and LSA Synthesized Synsets

THRESHOLD	0.8	0.85	0.9	0.95
Total Synsets	16,882	17,061	17,581	18,250
Identical to WordNet Synset	1,317	1,361	1,419	1,497
Subset of WordNet Synset	3,281	3,442	3,702	3,944
Superset of WordNet Synset	478	294	156	46

Figure 5: Analysis of Mindnet Synsets Synthesized with WxD LSA on Brown Corpus

We also note the two positive trends that the number of synthesized synsets that are subsets of WordNet synsets (thus, well-formed) increase, where as those that contain WordNet synsets (thus, possibly ill-formed) decrease.

The quality of the extracted synsets will directly depend on the quality of the LSA, given that extraction side of synonymy relationships are fixed; hence we experimented with a larger corpus, in order to capture the semantic relationships between different words in a statistically significant way. Hence we chose Microsoft Encarta, a corpus with about 17 Million words, comprising of 173,807 unique words distributed among 42,153 documents, to see if the resulting synsets are tighter, when compared with that of WordNet synsets. In addition, the Encarta corpus has about 413 words per document, making the documents smaller, and hopefully providing more meaningful associations between words. It should be noted that the words were not stemmed, but used as

they occur in the corpus. In comparison, the Brown corpus has about 1 Million words, distributed in about 500 documents. Figure 6 lists the results of this analysis:

THRESHOLD	0.8	0.85	0.9	0.95
Total Synsets	16,904	17,316	17,795	18,217
Identical to WordNet Synset	1,328	1,385	1,435	1,469
Subset of WordNet Synset	3,399	3,551	3,746	3,885
Superset of WordNet Synset	240	139	57	14

Figure 6: Analysis of Mindnet Synsets Synthesized with WxD LSA on Encarta Corpus

Comparing the results between Figure 5 and Figure 6, we can first observe that the same pattern of variation of the parameters with the threshold. We observe that the number of extracted synsets that are identical to the WordNet synsets is more, as well as those that are subsets of WordNet synsets. The extracted synsets that are superset of WordNet synsets decrease with the threshold. The figures taken together indicate that a small improvement in quality may be achieved by using a larger corpus for LSA analysis, in line with our expectations

5.1 Word-context provided by a Window

In the subsequent set of experiments, we used a word-proximity to measure and quantify the context of a word, as intuitively, such a behaviour might be more meaningful than assuming every word in a document provides the semantic context to a given word. In this procedure, a window representing a span of n words is passed over the corpus being used for LSA analysis, and the window is assumed to provide the context of a word to capture its semantics. The width of the window can be set, but we assumed the size to be 11 (providing 5 words on either side) as the context of a given word. In addition, we assumed a weighting parameter for a context word that is inversely proportional to the number of words separating the context word from the word under consideration, in a given window. Such weights provides a strength for an association between several co-occurring words. Such a window-based context measure provides a co-occurrence matrix that has, as axes, the entire vocabulary found in the corpus. Each cell of the matrix represents the co-occurrence counts for every word pair, weighted appropriately depending on the intervening words. Please note that the word pair, in our discussion, is sensitive to the direction of association; the associations "$x \ldots y$" and the associations "$y \ldots x$" are captured in different cells of the co-occurrence matrix.

We applied the above mentioned methodology on Brown corpus (of about 1 Million words), and by using a window size of 10 words (5 to either sides of the target word),

a word by word co-occurrence matrix was created. Such matrix was used for the LSA analysis (in a very similar manner as explained earlier), with the dimensions reduced to 15, using Singular Value Decomposition. Once created, word-by-word cosine similarity measures were used to disambiguate between words. The results of the above experimentation are shown in Figure 7.

THRESHOLD	0.8	0.85	0.9	0.95
Total Synsets	17,806	18,043	18,435	18,659
Identical to WordNet Synset	1,428	1,470	1,518	1,535
Subset of WordNet Synset	3,666	3,833	4,001	4,066
Superset of WordNet Synset	282	118	27	5

Figure 7: Analysis of Mindnet Synsets Synthesized with WxW LSA on Brown Corpus

Comparing the results of word-by-document analysis (as given in Figure 5) and the results by word-by-word analysis (as given in Figure 7), we notice that the same pattern of improvement of quality in all parameters of quantita-tive evaluation. However, there is a significant improvement in quality of the synthesized synsets, even more than that are synthesized by the Encarta corpus. One may conclude that the word-by-word context captures the semantic associations (for the same corpus), and is even more effective than just using larger corpus for analysis[2].

5.2 Verification against WordNet Synsets

While the above verification methodology compared the extracted synsets as a whole, against those of WordNet, we used a second methodology to examine how well the constituents (that is, constituent words) of the extracted synsets measured against the reference synsets, namely the WordNet synsets. While we do present a third strategy (in Section 6) that examines linguistically the quality of the synsets, such a methodology is too expensive, time-wise and resource-wise, to be pursued for the entire set of extracted synsets. In this section, we present our strategy to compare quality of extracted synsets against WordNet synsets in a quantitative manner.

The reference synsets that we use for the evaluation are from WordNet, and we take the hand-crafted WordNet synsets as the gold standard (that is, we do not question the correctness or completeness of the WordNet synsets). In this methodology, we only measure how well our synthesis strategy was able to cover the WordNet synsets.

[2]We were unable to run the word-by-word analysis for ENCARTA corpus, as the resulting size of the word-by-word matrix (with about 173,000 rows and columns each) was too large for our mathematical analysis system to handle.

First we define two metrics, precision and recall, for our extracted synsets against the reference synsets, and subsequently present the two metrics for each of the above synthesis methodology. A word-pair is defined to be a doubleton from a given synset. Hence, given a synset $\{ s_1, s_2, \cdots s_n \}$, there are n^2 word-pairs in it. Note that though a synset is a set of words, there could be multiple synsets that are associated with a word, corresponding to different senses of the word. Given the above, the precision metric of an extraction of synsets is defined (along the lines of IR systems) as the ratio of the common word-pairs between the extracted synsets and the WordNet synsets, to the total number of word-pairs in the extracted synsets. Similarly, the recall of the synthesized synsets is computed as the ratio between the common word-pairs between the extracted synsets and the WordNet synsets and the total number of word-pairs in the WordNet synsets. In essence, the recall metric indicates the fraction of the information in the WordNet synsets that has been captured in our synthesis, and the precision metric indicates the amount of extraneous information (or noise) present in the extracted synsets.

	0.8	0.85	0.9	0.95
BROWN, WORD-BY-DOC				
Total Synsets	16,882	17,061	17,581	18,250
Precision	0.186	0.218	0.234	0.251
Recall	0.047	0.0461	0.045	0.045
ENCARTA, WORD-BY-DOC				
Total Synsets	16,904	17,316	17,795	18,217
Precision	0.223	0.238	0.248	0.255
Recall	0.046	0.045	0.045	0.44
BROWN, WORD-BY-WORD				
Total Synsets	17,806	18,043	18,435	18,659
Precision	0.215	0.241	0.253	0.257
Recall	0.045	0.044	0.044	0.045

Figure 8: Analysis of Precision and Recall of Synthesized Synsets

Once the metrics were specified as above, the corresponding values of these metrics for each of our extraction methodology may be com-puted automatically. Figure 8 provides the results, from which, we could infer the following: First, the recall metric is very similar in all methodologies; this is to be expected, since the extraction-side is the same for all methodologies. Hence, irrespective of the methodology that is used for synthesis of the synsets, the same words would have been used, and hence recall

is bound to be similar. Second, the recall values are small (4.5%), for all extraction methodologies, which may be due to the following reasons: First, our strategy extracted only about 18,000 synsets (compared with about 80,000 in WordNet; second, out of the extracted synsets only about 1,500 are exactly same as the WordNet synsets, and we have nearly three-times as many synsets that are proper subsets of the WordNet synsets (thus yielding less number of word-pairs than the corresponding WordNet synsets) and, third, very few of the synthesized synsets are super-sets of the WordNet synsets. The precision of synthesized synsets sense-disambiguated using larger corpus are markedly better, in-line with our intuition. Further, we see that the disambiguation is better with a context provided by words around a given word, than by the entire document. We are currently, experimenting with different definitions of precision and recall metrics, in order to arrive at intuitive ones.

6 Qualitative Evaluation of Synthesized Synsets

Next, we manually examined the synsets synthesized to ascertain the quality, using the following procedure: first, we inspected the pre-LSA synset output, tagging synsets with inference-side errors (specifically, drifters) as bad. We then looked at subsets of the good and bad synsets post-LSA (viz., the cleaved synsets). Of the good synsets, 68% remained untouched by the LSA step (i.e., perfect overlap of pre- with post-LSA), while 32% got cleaved (27% of pre-LSA are supersets of their post-LSA counterparts; 5% are partial intersects). Of the bad synsets, meanwhile, every-thing was cleaved (0% perfect overlap of pre- and post-LSA; 95% of the pre-LSA synsets are supersets of their post-LSA counterparts; 5% are partial intersects), with most of them moving to "good" category.

Next, we selected a random 10% sample of synsets that were not well-formed in the naïve approach, and examined all the synsets in the new synthesis, containing any words that are part of the selected set. The new synsets were classified as *well-formed*, *ill-formed* and *iffy*, as done in the naïve approach, and the results presented in Figure 9.

THRESHOLD	0.8	0.85	0.9	0.95
Good Synsets	50%	54%	49%	59%
Bad Synsets	37%	38%	43%	32%
Iffy Synsets	13%	8%	8%	8%

Figure 9: Classification of Synthesized Synsets

It should be noted that we did not examine any words from the well-formed synsets from the naïve approach, since any synset from naïve approach can only break into

smaller pieces, and any subset of a well-formed synset will remain well-formed. We see that as the similarity threshold increases, the percentage of good synsets increased (as expected, as the synsets get smaller, and possibly, tighter). The growth in the good synsets was mainly due to the cleavage of the drifters from the pre-LSA synsets. The fraction of synsets that were iffy remains the same, indicating that their existence may be due to the extraction side errors.

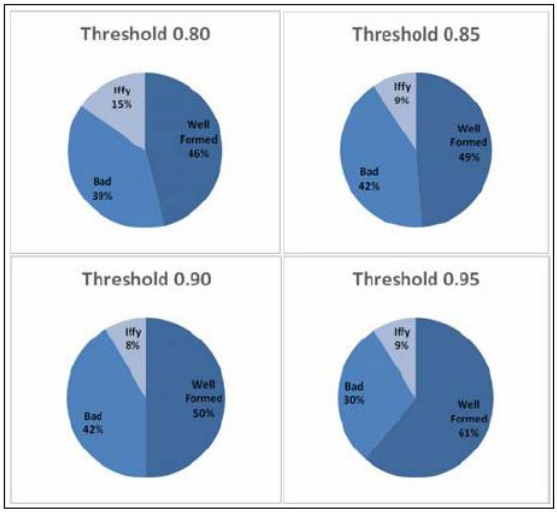


Figure 10: Word Coverage by Classification in the Synthesized Synsets

In addition, as shown in Figure 10, the words are also classified into one of the three categories, in line with that of the synsets. As expected, we see most words that are from bad synsets move into good synset category, with increasing threshold. Overall, we find that nearly half of the not well-formed synsets synthesized by naïve approach could be cleaved into smaller well-formed synsets, showing good promise in extraction of synonymy information using our methodology.

7 Conclusions and Future Work

In this paper, we presented an experiment to automatically acquire a lexical knowledge base of the same type as synonymy information represented in WordNet, using two complementary techniques – a broad-coverage parser for gleaning semantic information from a large corpus, and a word-sense disambiguation methodology to synthesize

the synsets. To validate our methodology, we conduct this experiment in English, so that the results may be compared directly with WordNet. We used the MindNet system for extracting synonymy information from a set of machine readable dictionaries, specifically the AHD and LDOCE, and construct synonymy using a naïve transitive closure approach. While this approach produced reasonable synsets, we observe the primary shortcoming that a large fraction of the synsets are drifters, that is, those that accumulate large unrelated collection of words, due to the polysemous nature of words and the lack of sense disambiguation used in synset construction. Subsequently, we used the result of Latent Semantic Analysis on a large corpus, and used the resulting basis for adding senses of a given word during the synthesis process. A manual analysis indicates that the quality of the resulting synsets improves significantly. Though our proposed methodology did not produce perfect synsets, it shows promising results in automatically extracting synsets from natural language text.

The current experiment uses a specific type of natural language text, namely, machine readable dictionaries, but this approach is not limited to dictionaries as many others have demonstrated algorithms to identify definitional text in freely occurring natural text, as in, (18) and (8). The current experiment also takes its input from *Syn* and *Hyp* relations extracted by MindNet using a broad-coverage parser. Naturally, we cannot make the assumption that a parser exists for the language for which we seek to create a WordNet resource, where we can only expect little or no resources. However, other studies have shown that the accuracy for acquiring hypernymy and synonymy using simple string patterns can be as high as 86% for dictionary text (2), and it is likely that the accuracy will be similarly high for the acquisition from text classified as definitional, using patterns such as described in (7). We used the synonyms provided by MindNet not to demonstrate that a broad-coverage parser was required, but rather to demonstrate the feasibility of combining automatically extracted synonyms with LSA to produce a lexical knowledge base similar in quality to WordNet. What remains to be shown is the size of the knowledge base we might extract in this manner for a language that might have a smaller body of available text to draw from than languages already studied. However, we anticipate that the knowledge base created can act as a seed for subsequent extensions, such as suggested by (17) and (20). In combination, these methods will pave the way for unprecedented levels of access to the under-represented languages of the world.

Acknowledgements *We thank Jagadeesh Jagarlamudi for his insightful comments and discussions, and Gary Kacmarcik for his valuable help in experimentations.*

8 References

References

Bellegarda, J. R. Exploiting Latent Semantic Information in Statistical Language Modeling. *Proceedings of the IEEE, Vol. 88, No. 8*, 2000.

Chodorow, M., Byrd, R. J. and Heidorn, G. E. Extracting Semantic Hierarchies from a Large On-Line Dictionary. *Proceedings of the ACL*, 1985.

Euro WordNet. *http://www.illc.uva.nl/EuroWordNet.*

Fellbaum, C., Ed. 1998. WordNet: An Electronic Lexical Database. *MIT Press, London*.

Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter L. A. and Lochbaum, K. E. *Proceedings of the 11th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 1988.

The Global WordNet Association. *http://www.globalwordnet.org*.

Hearst, M. A. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 1992.

Joho, H. and M. Sanderson. Retrieving descriptive phrases from large amounts of free text. *Proceedings of CIKM, pages 180-186*, 2000.

Kucera, H. and Francis, W. N. Computational Analysis of Present-Day American English. *Brown University Press, Providence RI*, 1967.

Kumaran, A., Makin, R., Pattisapu, V., Sharif, S. E., Kacmarchi, G. and Vanderwende, L. Automatic Extraction of Synonymy Information: An Extended Abstract. *Proceedings of the Ontologies in Text Technology Workshop*, 2006.

Landauer, T. K., Foltz, P. W., and Laham, D. Introduction to Latent Semantic Analysis. *Discourse Processes 25: 259-284*, 1998.

Montemagni, S. and Vanderwende, L. Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries. *Proceedings of COLING*, 1992.

Nakov, Preslav, Popova, A. and Mateev, P. Weight Functions Impact on LSA Performance. *Recent Advances in NLP*, 2001.

Oxford English Dictionary. 2nd ed. 1989 (ed. J. A. Simpson and E. S. C. Weiner). *Oxford University Press, Oxford, UK*.

Resnik, P. and Yarowsky, D. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering 5 (3): 113-133*, 2000.

Richardson, S. D., Dolan, W. B., and Vanderwende, L. MindNet: acquiring and structuring semantic information from text. *Proceedings of the COLING*, 1998.

Roark, B. and Charniak, E. Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 1998.

Saggion, H. and Gaizauskas, R. Mining on-line sources for definition knowledge. *Proceedings of the 17th International FLAIRS Conference*, 2004.

Schütze, H. Automatic word sense discrimination. *Computational Linguistics, Vol 24, Issue 1*, 1998.

Snow, R., Jurafsky, D., and Ng, A. Semantic taxonomy induction from heterogeneous evidence. *Proceedings of COLING/ACL*, 2006.

Suzuki, H., Kacmarcik, G., Vanderwende, L. and Menezes, A. MindNet / mnex: An Environment for Exploring Semantic Space). *Proceedings of the 11th Annual Meeting of the Society of Natural Language Processing*, 2005.

Vanderwende, L. Ambiguity in the Acquisition of Lexical Information. *AAAI Spring Symposium Series, No. 95/01, 174-179*, 1995.

Vanderwende, L., Kacmarcik, G., Suzuki, H. and Menezes, A. MindNet: An Automatically-Created Lexical Resource. *Proc. of HLT/EMNLP*, 2005.

Daniela Goecke, Maik Stührenberg, Tonio Wandmacher

A hybrid approach to resolve nominal anaphora

In order to resolve nominal anaphora, especially definite description anaphora, various sources of information have to be taken into account. These range from morphosyntactic information to domain knowledge encoded in ontologies. As the acquisition of ontological knowledge is a time-consuming task, existing resources often model only a small set of information. This leads to a knowledge gap that has to be closed: We present a hybrid approach that combines several knowledge sources in order to resolve definite descriptions.[1]

1 Resolving nominal anaphora

1.1 Nominal anaphora and semantic knowledge

The term nominal anaphora comprises both pronominal anaphora as well as NP or definite description anaphora (DDA henceforth). In order to resolve DDA not only morphological or syntactic knowledge is needed but also information on (lexical) semantics and domain knowledge. A large amount of work in the domain of anaphora resolution has been done in the area of pronominal anaphora achieving good results (see Mitkov, 2002, for an overview); extensive work is still done in the area of resolving definite description anaphora (Vieira and Poesio, 2000; Ng and Cardie, 2002; Versley, 2007). Many of these approaches make use of information from pre-established lexical resources (WordNet or GermaNet), or try to acquire lexical knowledge by applying automated extraction methods on large corpora (or the web, cf. Markert et al., 2003; Versley, 2007). Other approaches rely on methods that determine semantic relatedness from cooccurrence information in corpora (cf. Poesio et al., 1998).

DDA relations hold between nominal discourse entities (or referents, cf. Karttunen, 1976)[2] and can be of various types: In example (1), the antecedent is explicitly mentioned and can be resolved via the same head noun (*direct anaphora* in terms of Vieira and Poesio, 2000). In the second example the antecedent is not explicit mentioned, however

[1]The work presented in this article is a joint effort of the projects A2 and C2 of the Research Group *Text-technological modelling of information* funded by the German Research Foundation. The corpus under investigation was developed by the projects A2 and C1.

[2]Discourse entities are constants within a discourse model evoked by NPs and which can be referred to in the subsequent discourse. NPs can either evoke new discourse entities in the discourse model or can "refer to ones that are already there" (Webber, 1988).

the anaphoric relation can be resolved on the basis of the hyperonymy relation between *questionnaire* and *form*. In order to resolve examples (3) and (4) additional semantic knowledge is needed. As opposed to examples (1) and (2) the anaphoric element and its antecedent do not refer to the same referent in the latter examples. Following the terminology of Clark (1977) we will refer to examples (3) to (4) using the term *bridging relations*.

(1) a questionnaire - the questionnaire

(2) a questionnaire - the form

(3) an interview - the questionnaire

(4) an interview - the respondent

Bridging relations occur when the antecedent is not explicitly mentioned in the text but has to be inferred from the context. Cues to solve bridging relations are domain knowledge (frames, scripts or schemata, e.g. interviews are often done using questionnaires) or (lexical-)semantic knowledge encoded in lexical nets like GermaNet or WordNet. The classification of these lexical nets ranges from "terminological ontology" (Sowa, 2000) to "full ontology" (Oard, 1997). We follow the terminology introduced by Erdmann (2001, p. 72) who uses the term *light weight ontology* to define ontologies that consist primarily of a representation schema to define taxonomies as well as attributes or relations. In contrast, *heavy weight ontologies* include complex logical descriptions that are specified in more expressive logical formalisms. However, using GermaNet alone as resource for detecting semantic relations is not sufficient considering the coverage in regard to the corpus under investigation. In order to close this gap we present a hybrid approach for automatically determining semantic relatedness in order to identify the most likely antecedent from a set of antecedent candidates.

1.2 Acquiring semantic knowledge

In the past few years a variety of approaches has been presented to automatise the extraction of ontological knowledge from structured and unstructured data. The output of these systems is usually rather rudimentary and noisy. Nevertheless, this kind of information coming from automated approaches can be considered as a valuable resource for our task. Regarding the current approaches to derive ontological knowledge from unstructured data two main classes can be made up:

The first one is based on distributional or structural similarity, starting from the assumption that words being semantically similar tend to occur in similar contexts and structural settings. In this family we find completely knowledge-free approaches relying on cooccurrence only (e.g. Paaß et al., 2004); Poesio et al. (1998) showed how this kind of

information can be used to resolve nominal anaphora (using the *HAL*-model; cf. Lund and Burgess, 1996).

The second class of methods basically relies on lexico-syntactic patterns, the so-called *Hearst* patterns (Hearst, 1992). Here, a text corpus is scanned for characteristic word combinations, typically containing a semantic relation between two terms (e.g. X *such as* Y, X being a hyperonym of Y). Recently, hybrid approaches can be found using both techniques to enhance the quality of the extractions. In Cimiano and Staab (2005), nouns are first clustered by cooccurrence methods and Hearst patterns are applied afterwards to extract the most useful relations. Cederberg and Widdows (2003) go the other way around: Based on patterns they extract word pairs from text and filter them by a cooccurrence based threshold being able to raise precision by 30%, compared to a standard pattern-based approach.

1.3 Objectives and organization of the article

The objective of our approach is to increase information on semantic relatedness of terms by a combination of – amongst others – extracted relations and cooccurrence information, and to use it in an anaphora resolution system. In general, the anaphora resolution process can be subdivided into three steps: (1) For each anaphoric element, determine an antecedent candidate list (ACL) and (2) apply constraints to exclude incompatible candidates from the ACL; (3) identify the most likely antecedent.

This paper concentrates on step 3, i.e. on the identification of the most likely antecedent candidate. We use a fixed search window to collect the candidate list and we do not apply constraints to downsize the list thus leading to forced test conditions for step 3. Ongoing work focuses on the implementation of a variable search window size as well as on the implementation of constraints for step 2. The remainder of the paper is structured as follows: Section 2 introduces the corpus under investigation as well as the annotation scheme and procedure, Section 3 describes the methods applied in our approach: GermaNet lookup, Hearst patterns, recency information and a semantic similarity measure, based on Latent Semantic Analysis (LSA). Finally, section 4 discusses the results of our approach, and Section 5 presents a conclusion and ongoing work.

2 The corpus under investigation

2.1 Annotation Scheme and Procedure

The evaluation of the approach described above is based on a corpus of German scientific and newspaper articles annotated for training and evaluation of an anaphora resolution system. The subset used for the evaluation presented in this paper includes three scientific articles and one newspaper article. For the purpose of anaphora resolution

the corpus has been annotated for discourse entities (DEs) and anaphoric relations between DEs. Several annotation schemes for annotating anaphoric relations have been developed in the last years, e.g. the UCREL anaphora annotation scheme (Fligelstone, 1992; Garside et al., 1997), the SGML-based MUC annotation scheme (Hirschmann, 1997), and the XML-based MATE/GNOME scheme for anaphoric annotation (Poesio, 2004), amongst others. The annotation scheme used for our approach is based on the one presented by Holler et al. (2004) and has been adapted for the annotation of bridging relations (Goecke et al., 2007). The versions of the annotation scheme are used within our research group both for the task of hypertextualization (project B1) as well as for the task of anaphora resolution (project A2). Therefore, the annotation of anaphora and coreference is distinguished explicitly: "Although anaphoric and coreferential relations can coincide, it is not generally the case that all coreferential relations are anaphoric, nor are all anaphoric relations coreferential" (Holler et al., 2004).[3] This distinction especially holds for bridging relations that can be inferred due to semantic role assignment (*a wedding - the bride*) or the meronymy relation (*a car - the wheel*): In these examples the anaphor and the antecedent do not refer to the same discourse entity even if an anaphoric relation holds between them. For cospecification and bridging relations two types of primary relations have been defined:

- *cospecLink*: Cospecification; anaphor and antecedent refer to the same referent;
- *bridgingLink*: Bridging; associative or indirect anaphora (Clark, 1977); anaphor and antecedent do not refer to the same referent.

For each of these relations a set of secondary relation types has been defined (see Table 1).

The corpus has been preprocessed using the dependency parser *Machinese Syntax*[4] which provides lemmatization, POS information, dependency structure, morphological information and grammatical function. Based on this information, discourse entities have been detected automatically by identifying nominal heads (i.e. nouns or pronouns) and their premodifiers. The anaphoric relations are annotated using the annotation tool *Serengeti* described in Stührenberg et al. (2007)[5]. The annotation procedure is subdivided into four steps: First, it is checked for each discourse entity (DE) whether it is used anaphorically. For each anaphoric DE the correct antecedent is identified, and for each anaphor/antecedent pair (AC pair henceforth) the primary relation is chosen. As the last step, the secondary relation is chosen. Listing 1 shows a sample annotation from a german linguistic article. In this example a bridging relation holds between the

[3] The MATE scheme states the distinction between anaphoric relations and reference proper, however the distinction is not made explicit in the annotation scheme; the term *coreference* is used to denote anaphoric annotation (Poesio, 2004).

[4] `http://www.connexor.eu/technology/machinese/machinesesyntax/`

[5] `http://coli.lili.uni-bielefeld.de/serengeti/annotator.pl`

cospecLink		
ident	pronouns same head noun of anaphor and antecedent	a man – he a man – the man
namedEntity	anaphor is an NP referring to a proper noun antecedent	Peter Jones – the man
propName	anaphor is a proper noun antecedent may be either an NP or a proper noun	the CTO – Peter Jones Peter Jones – Jones
synonym	synonymy holds between head nouns	a car – the automobile
paraphrase	anaphor is a paraphrase	the HTML-editor – the web site creation tool
hyperonym	anaphor is an hyperonym of the antecedent	a horse – the animal
hyponym	anaphor is an hyponym of the antecedent	an animal – the horse
addInfo	anaphor adds further information	Peter Jones – the 67 year old CTO
bridgingLink		
possession	possessive relation	Peter – his car
meronym	anaphor is part of the antecedent	a room – the window
holonym	anaphor has the antecedent as one of its parts	the window – the room
setMember	anaphor is an element of a set	two cars – the red car
hasMember	anaphor is a set consisting of its antecedents	Paul [...] Susan – the two children
bridging	associative link (e.g. role assignment, schema)	a wedding – the bride

Table 1: Secondary relation types for cospecLink and bridgingLink

discourse entities denoted by *die Befragung* ('interview'; lines 4-6) and *der Fragebogen* ('questionnaire'; lines 27-29).

2.2 Corpus Design

The evaluation set comprises a total amount of 4196 DEs. Based on these DEs, a total amount of 1433 *cospecLinks* and 541 *bridgingLinks* could be found. In our study we focus on those relation types between anaphor and antecedent that can be found in GermaNet: synonymy, hyperonymy, hyponymy, meronymy, holonymy, bridging. The subset that contains the semantic relations under investigation comprises a total amount of 224 anaphoric links. As distance between anaphor and antecedent is a crucial point, we defined a fixed distance for our evaluation. Especially for bridging relations in scientific articles, distances between anaphor and antecedent can be extremely large. For our corpus, distances up to hundred DEs could be found, therefore, not all of the relations have been taken into account. Corpus investigation shows that limiting the distance to 15 DEs results in a reasonable subset: 50% of the cospecLinks and 55.78% of the bridgingLinks find their antecedent within this window. Thus, for each anaphoric DE a candidate list of (at most) 15 possible antecedents has been created (including

the correct one that has been marked during the annotation process).[6] This leads to an evaluation set of 115 anaphoric DEs and 1428 antecedent candidates (app. 12,5 candidates per anaphor). For the corpus study presented here we have chosen this fixed window; however one has to include more sophisticated methods in order to find suitable sets of antecedent candidates in a complete anaphora resolution system due to varying distances between anaphor and antecedent. Modelling the search space for candidate sets that cover both anaphors with small distances as well as anaphors with long distances should not be grounded solely on the linear structure of text but should be flexible in size according to structural elements, e.g. on the basis of discourse structure (cf. Cristea et al., 2000; Chiarcos and Krasavina, 2005) or logical document structure (Goecke and Witt, 2006).

3 Method

Our approach makes use of four information sources and combines them into one measure. It is a forced choice algorithm, i.e. to any input pair of anaphor and antecedent candidate a score will be assigned. In the following we describe the four single methods separately, and then we show how we combine their information.

3.1 GermaNet relations

As we have already shown, many bridging phenomena are based on synonymy, hyponymy or meronymy. These relations are encoded in a lexical resource like GermaNet, making it our first source of information, since the information being found here are very reliable and noise-free, despite of their low coverage. For each AC pair the underlying lemmas are looked up in GermaNet and – if both are included – the distance between the corresponding nodes is computed (cf. Poesio et al., 2004, for node-node distance measures using WordNet). Nevertheless, distance information does not include information on the relation holding between two lemmas, this information has to be computed from the path information separately. In our study node-node distances have been computed using the implementation provided by the project A4 of our research group (cf. Mehler et al., 2007)[7]. The resulting distance values (in terms of path length) have been normalised for each set of AC pairs belonging to a given anaphor. A value of 1 indicates the shortest path within a given set and a value of 0 indicates either maximum length or the fact that one token of the AC pair is not found in GermaNet.

[6] Only non-pronominal DEs can serve as antecedents, thus the candidate list may be shorter than 15 elements.
[7] http://www.scientific-workplace.org/

```
<chs:chs>
<chs:text>
 <cnx:token ref="w2732">In</cnx:token>
 <chs:de deID="de764" deType="nom" headRef="w2734">
  <cnx:token ref="w2733">die</cnx:token><cnx:token ref="w2734">Befragung</cnx:token>
 </cnx:de>
 <cnx:token ref="w2735">wurden</cnx:token><cnx:token ref="w2736">nur</cnx:token>
 <chs:de deID="de765" deType="nom" headRef="w2738">
  <cnx:token ref="w2737">solche</cnx:token><cnx:token ref="w2738">Kurse</cnx:token>
 </cnx:de>
 <cnx:token ref="w2739">einbezogen</cnx:token><cnx:token ref="w2740">,</cnx:token>
 <chs:de deID="de766" deType="nom" headRef="w2741">
  <cnx:token ref="w2741">die</cnx:token>
 </cnx:de>
 <cnx:token ref="w2742">bereits</cnx:token><cnx:token ref="w2743">über</cnx:token>
 <chs:de deID="de767" deType="nom" headRef="w2745">
  <cnx:token ref="w2744">gute</cnx:token><cnx:token ref="w2745">Grundkenntnisse
  </cnx:token>
 </cnx:de>
 <cnx:token ref="w2746">in</cnx:token>
 <chs:de deID="de768" deType="nom" headRef="w2749">
  <cnx:token ref="w2747">der</cnx:token><cnx:token ref="w2748">deutschen</cnx:token>
  <cnx:token ref="w2749">Sprache</cnx:token>
 </cnx:de>
 <cnx:token ref="w2750">verfügten</cnx:token><cnx:token ref="w2754">,</cnx:token>
 <cnx:token ref="w2755">da</cnx:token>
 <chs:de deID="de770" deType="nom" headRef="w2757">
  <cnx:token ref="w2756">der</cnx:token><cnx:token ref="w2757">Fragebogen</cnx:token>
 </cnx:de>
 <cnx:token ref="w2758">nur</cnx:token><cnx:token ref="w2759">auf</cnx:token>
 <chs:de deID="de771" deType="nom" headRef="w2760">
  <cnx:token ref="w2760">Deutsch</cnx:token>
 </cnx:de>
 <cnx:token ref="w2761">vorlag</cnx:token><cnx:token ref="w2762">.</cnx:token>
</chs:text>
<chs:standoff>
 <chs:semRel>
  <chs:bridgingLink relType="bridging" phorIDRef="de770" antecedentIDRefs="de764"/>
 </chs:semRel>
 <cnx:token_ref id="w2757" head="w2761" pos="N" syn="@NH" lemma="frage#bogen"
  depV="subj" morph="MSC␣SG␣NOM"/>
 <cnx:token_ref id="w2734" head="w2735" pos="N" syn="@NH" lemma="befragung"
  depV="advl" morpho="FEM␣SG␣ACC"/>
</chs:standoff>
</chs:chs>
```

Listing 1: The annotation format for anaphoric relations. Shortened and manually revised output

3.2 Relation extraction by patterns

Our second information source relies on pattern-based information. We follow the approaches of Markert et al. (2003) and Versley (2007), who look up patterns on the web. We first generate patterns of the types "*X und andere Y*", "*X wie Y*", "*X insbesondere Y*", "*X einschließlich Y*" for all AC pairs of our text corpus and submit them as queries via the *Google* API. We then compute a normalized score from the added hit counts of each pattern.

3.3 Recency information

Since linear distance between an anaphor and a potential candidate also provides valuable information, we took a closer look at the distance distribution in our corpus. We determined the distance (in DEs) between each AC pair; the (standardized) distribution is shown in Figure 1 (columns). It can be seen that the most frequent distance between anaphor and antecedent is 5 DEs. We can assume that the distances are (roughly) normally distributed after this peak. However, assuming normal distribution with the same standard deviation σ beforehand would result in an overestimation of very short distances (1-4). For this reason we apply two different σs (σ_-, σ_+) in order to best adapt to this distribution. Equation 1 displays our recency function, the curve in Figure 1 shows the developing of the function for $x = 0 - 20$.

$$Rec(x) = e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2},\ with\ \mu = 4, \sigma_- = 1\ and\ \sigma_+ = 5. \tag{1}$$

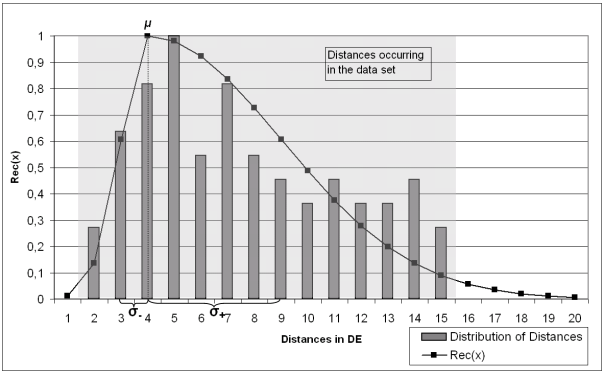


Figure 1: Graph of our recency function and distribution of distances (0-20 DEs)

3.4 LSA-based similarity

Since the early 1990s, Latent Semantic Analysis (LSA) has become a well-known technique in NLP. When it was first presented by Deerwester et al. (1990), it aimed mainly at improving the vector space model in information retrieval. Its abilities to enhance retrieval performance are remarkable; results could be improved by up to 30%, compared to a standard vector space technique (Dumais, 1995). Moreover, meaningful documents could be retrieved that did not share a single word with the query.

LSA is based on the vector space model from information retrieval (Salton and McGill, 1983). Here, a given corpus of text is first transformed into a term×context matrix A, displaying the occurrences of each word in each context. Usually, this matrix is then weighted by one of the standard weighting methods used in IR (cf. Salton and McGill, 1983). The decisive step in the LSA process is then a *singular value decomposition* (SVD) of the weighted matrix. Thereby the original matrix A is decomposed as follows:

$$SVD(A) = U\Sigma V^T \tag{2}$$

The matrices U and V consist of the eigenvectors of the columns and rows of A. Σ is a diagonal matrix, containing in descending order the singular values of A. By only keeping the k strongest (k usually being 100 to 300) eigenvectors of either U or V, a so-called semantic space can be constructed for the terms or the contexts, respectively. Each term or each context then corresponds to a vector of k dimensions, whose distance to others can be compared by a standard vector distance measure. In most LSA approaches the *cosine* measure is used.

We use a slightly different setting, close to the one described by Schütze (1998) and Cederberg and Widdows (2003), where the original matrix is not based on occurrences of terms in documents but on other cooccuring terms (term×term-matrix). We thus count the frequency with which a given term occurs with others in a predefined context window ($\pm 10 - 100$ words). After applying *singular value decomposition*, each word is represented as a vector of k dimensions, and for every word pair w_i, w_j of our vocabulary we can calculate a similarity value $Sim(w_i, w_j)$, based on the *cosine* between their respective vectors.

Treatment of compounds: As German compounds are lexicalized as a single graphical unit, they are often a tricky problem for NLP applications. Many algorithms rely at some point on string matching in order to identify lexical units in a given text; many compounds are not part of any predefined vocabulary, therefore they are neglected in further processing stages. Our LSA component, however, is able to deal with compound words, since we make the (somewhat simplifying) assumption that the meaning of a compound word is the sum of its parts. This idea is straightforward in a vectorial setting: Every time we encounter a compound which is not contained in the vocabulary, we split

it up into its parts (by partial matching) and take the vector sum of the corresponding vectors. This simple measure works surprisingly well, as can be seen in the Section 4.

3.5 Combining information sources

So far we have four information sources at hand, which can describe possible anaphoric relations: GermaNet, lexico-syntactic patterns, linear distance or recency information, and LSA similarity. We now have to combine this information into one measure in order to be able to calculate the most likely antecedent out of our candidate list. A well-known way to combine information from several sources is interpolation. We describe in the following how this can be done in our setting:

So, for a given anaphoric expression b and a set of candidates of antecedents $A = (a_1, ..., a_n)$,

1. we consult for each candidate $a_1, ..., a_n$ if a path to b can be found in GermaNet. We define a function $GN(a_i, b)$ whose values range from 0 to 1, according to the normalised path length;

2. we define a function $Pat(a_i, b)$ returning the normalized frequency score of matching candidate strings including a_i and b;

3. we determine the LSA-similarity $Sim(a_i, b)$ between a_i and b with respect to a previously calculated reference semantic space;

4. finally a recency function $Rec(a_i, b)$ determines the recency factor for the distance between a_i, b, as described in Formula 1.

Each candidate a_i then receives a score $Sc(a_i)$ by interpolating the results from the single functions defined above. The parameters λ_{GN}, λ_{Pat}, λ_{LSA}, λ_{Rec} will be set empirically. It is clear that advanced optimization techniques such as the EM algorithm could be employed here. However, since our test set is rather small, we could not assure to reach converged values, therefore we adjust the values manually.

$$Sc(a_i) = \lambda_{GN} \cdot GN(a_i, b) + \lambda_{Pat} \cdot Pat(a_i, b) + \lambda_{LSA} \cdot LSA(a_i, b) + \lambda_{Rec} \cdot Rec(a_i, b)$$

It is important to note that this function assigns a score to any pair of anaphor and antecedent. Apart from the maximum distance of 15 DEs we apply no further exclusion criteria, our algorithm is forced to make a choice among the candidates, according to their respective score, even though none of the semantic components might be able to assign a value (due to an unknown word in the pair). The choice is based on recency information only, which is necessarily rather unreliable.

4 Results

GermaNet Relations For 71% of the DE in our corpus the underlying lemma of the head noun is stored in GermaNet. For 759 out of 1428 AC pairs (53,15%) a path length could be computed.

Relations generated by patterns As described before we generated candidate strings comprising one out of 4 patterns and an AC pair each. We submitted each of the strings as a query to the *Google* API, and we summed up the total hit counts for each AC pair.[8] The summed up hit counts were logarithmized and normalized in order to have a meaningful score that can be used in the interpolation formula. As expected, most of the hit counts were 0, only for 119 out of 1428 AC pairs (8,3%) we could find at least one matching pattern.

LSA-based similarity factor Using the Infomap[9] toolkit, we calculated a term×term-cooccurrence matrix of 80.000×3.000 words over a corpus of 101 million token (from *Wikipedia* and *Tageszeitung*). This matrix was then reduced by *singular value decomposition* to 150 dimensions, giving us a vector for each of the 80.000 words. We now calculated for each of the 1428 AC pairs their LSA-similarity using the cosine distance of their respective vectors.

For compound words we calculated the normalized sum of the vectors of each component and used it instead of the word vector. This tremendously reduced blind spots in the calculation process: Only 94 out of the 1428 word pairs (6,5%) could not be assigned a similarity value, whereas this would have been the case for 910 pairs (63%) without compound treatment.

Recency function For each of the 1428 AC pairs, its recency factor was calculated, using the recency function in Formula 1 (see p. 50), with $\mu = 4$, $\sigma_- = 1$ and $\sigma_+ = 5$. We admit that, due to limited data resources, we could not estimate the parameters on a held out test set, however we would expect these parameters to be quite stable over different corpora.

Overall results To get a first impression of the effectiveness of each component, we set successively each of the four coefficients to 1, the others to 0. For the GermaNet component we get 20 right candidates, for the pattern approach we get 10. 51 of the correct candidates could be found by the LSA component only. The recency component by itself finds 17 correct candidates, however it seemed to interfere with the LSA

[8] Thanks to Henrik Dittmann, Universität Osnabrück for his help.
[9] `http://infomap-nlp.sourceforge.net/`

component: When we gave equal strength to both the LSA and the recency component ($\lambda_{LSA} = 0,5; \lambda_{Rec} = 0,5$), only 34 correct candidates could be found. The maximum number of correct candidates (57) could be found using the parameters given in the last line of Table 2.

Coefficients				# correct	# wrong
λ_{GN}	λ_{Pat}	λ_{LSA}	λ_{Rec}		
1,0	0	0	0	20	95
0	1,0	0	0	10	105
0	0	1,0	0	51	64
0	0	0	1,0	17	98
0,25	0,05	0,65	0,05	57	58

Table 2: Overall results for our test set of 115 anaphors

When we split up the results for the different relation types (cf. Table 3), we see immediately that there is an important difference between the semantic and the bridging relations: Whereas 34 out of the 56 anaphors based on a straightforward semantic

Relation type	# correct	# wrong	# total
Hypo-/Hyperonyms	1 (33,3%)	2	(3)
Mero-/Holonyms	4 (36,4%)	7	(11)
Synonyms	29 (69,0%)	13	(42)
All sem. relations	34 (60,8%)	22	(56)
Bridging	23 (38,9%)	36	(59)
Overall	57 (49,6%)	58	(115)

Table 3: Results for each of the relation types considered

relation could be resolved (61%), this was the case for only 23 out of 59 bridging anaphors (39%).

Another remarkable fact is that among the semantic relations the synonyms scored far better than the meronymic or hyponymic relations. This shows the effectiveness of the LSA to measure semantic similarity between terms, since the meaning of two synonyms will be more similar than that of mero- or hyponyms.

Regarding the N-best distribution in Figure 2, we can see that most of the correct AC pairs appear on the top of the N-best lists. When we consider the first two candidates, we find 71 correct pairs (62%), the first 4 candidates comprise already 86 (75%) and the first 6 candidates 97 correct pairs (84%). Our approach therefore seems to calculate plausible semantic relationships, however it is not precise enough in the selection process.

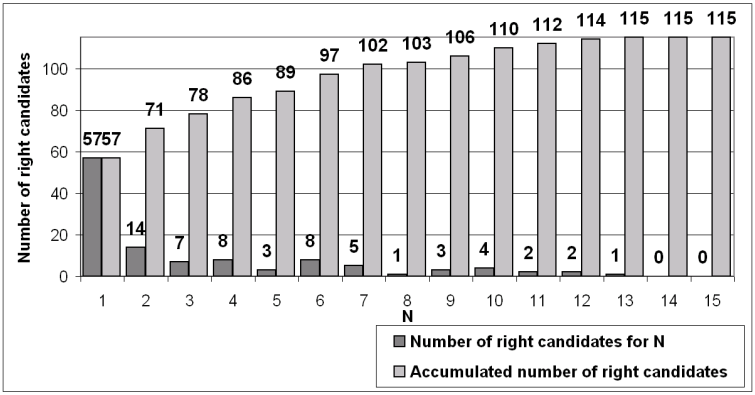


Figure 2: N-best analysis for our test set of 115 anaphors

A thorough look at the ranked lists of the candidates seems to confirm this observation: Many of the candidates are indeed ranked according to their semantic relatedness with the anaphor. Table 4 shows a typical candidate list, the candidates are ranked by their score.

correctAnte:de764 relation:bridgingLink(bridging)						Fragebogen
nbest	deID	distance	GN value	LSA	total score	text
1	**de764**	**5**	**0,4**	**0,221**	**0,294**	***Befragung***
2	de768	1	1	-0,028	0,286	*Sprache*
3	de761	8	0,6	0,027	0,203	*Unterricht*
4	de757	11	0,4	0,099	0,189	*Prüfungen*
5	de762	7	0,4	0,063	0,187	*Gruppen*
6	de767	2	0,2	0,093	0,152	*Grundkenntnisse*
7	de758	10	0,4	-0,105	0,130	*Niveaus*
8	de763	6	0,2	0,048	0,130	*Deutsch*
9	de756	12	0,4	0,015	0,128	*Vorbereitung*
10	de765	4	0,2	0,039	0,125	*Kurse*
11	de755	13	0,2	0,040	0,090	*Kurse*
12	de760	9	0	0,009	0,042	*Instituten*
13	de750	15	0	0,004	0,005	*Goethe-Institut*

Table 4: N-best list with correct antecedent found (correct antecedent in bold letters)

5 Conclusion and Outlook

This paper presents ongoing work in the domain of nominal anaphora resolution; it concentrates on the identification of the most likely antecedent from a set of antecedent candidates. Future work includes both further improvement of this component as well as work on the other two components of an anaphora resolution model: Defining the set of antecedent candidates and applying constraints to eliminate incompatible antecedent candidates from the set.

Concerning the pattern extraction component, future work focuses on the definition of more patterns and especially those extracting synonymy or meronymy relations (the results for these patterns are usually not as reliable as for the ones we used). Further experiments are needed in order to understand which patterns help and which do not.

Concerning the remaining components of the anaphora resolution system, work is done in order to define a variable search window in order to find suitable candidate sets for anaphoric items that find their antecedent at long distance. This work includes the analysis of rhetorical structure and logical document structure. Regarding the use of constraints to eliminate incompatible items from the set of candidates we assume that congruence restrictions (e.g. number agreement) might help downsizing the set of candidates and thus will help to improve the complete system; the smaller the number of elements for the semantic component the better the overall results as elements already identified as being incorrect candidates cannot interfere the LSA component.

References

Cederberg, S. and Widdows, D. (2003). Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy. In *Proc. of the Conference on Natural Language Learning (CoNNL)*.

Chiarcos, C. and Krasavina, O. (2005). Rhetorical distance revisited: a parametrized approach. In *Proceedings of Workshop on Constraints in Discourse*, pages 63–70, Dortmund, Germany.

Cimiano, P. and Staab, S. (2005). Learning concept hierarchies from text with a guided agglomerative clustering algorithm. In *Proc. of the ICML Workshop on Learning and Extending Lexical Ontologies with Machine Learning Methods*, Bonn, Germany.

Clark, H. (1977). Bridging. In Johnson-Laird, P.N. & Wason, P., editor, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, Cambridge.

Cristea, D., Ide, N., Marcu, D., and Tablan, M.-V. (2000). Discourse structure and co-reference: An empirical study. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*, Luxembourg.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Dumais, S. T. (1995). Latent semantic indexing (lsi): Trec-3 report. In Harman, D., editor, *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*, volume 500-226, pages 219–230. NIST Special Puclication.

Erdmann, M. (2001). *Ontologien zur konzeptuellen Modellierung der Semantik von XML*. Books on Demand GmbH.

Fligelstone, S. (1992). Developing a Scheme for Annotating Text to Show Anaphoric Relations. In Leitner, G., editor, *New Directions in English Language Corpora: Methodology, Results, Software Developments*, pages 153–170. Mouton de Gruyter, Berlin.

Garside, R., Fligelstone, S., and Botley, S. (1997). Discourse Annotation: Anaphoric Relations in Corpora. In Garside, R., Leech, G., and McEnery, A., editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 66–84. Addison-Wesley Longman, London.

Goecke, D., Stührenberg, M., and Holler, A. (2007). Koreferenz, kospezifikation und bridging: Annotationsschema. Interne Reports der DFG-Forschergruppe 437 "Texttechnologische Informationsmodellierung".

Goecke, D. and Witt, A. (2006). Exploiting logical document structure for anaphora resolution. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

Hirschmann, L. (1997). MUC-7 Coreference Task Definition (version 3.0). In Hirschman, L. and Chinchor, N., editors, *Proceedings of Message Understanding Conference (MUC-7)*.

Holler, A., Maas, J.-F., and Storrer, A. (2004). Exploiting coreference annotations for text-to-hypertext conversion. In *Proceeding of LREC*, volume II, pages 651–654, Lisbon, Portugal.

Karttunen, L. (1976). Discourse referents. *Syntax and Semantics: Notes from the Linguistic Underground*, 7:363–385.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments and Computers*, 28(2):203–208.

Markert, K., Modjeska, N., and Nissim, M. (2003). Using the web for nominal anaphora resolution. In *Proc. of the EACL Workshop on the Computational Treatment of Anaphora*.

Mehler, A., Waltinger, U., and Wegner, A. (2007). A formal text representation model based on lexical chaining. In *Proceedings of the KI 2007 Workshop on Learning from Non-Vectorial Data (LNVD 2007)*, pages 17–26, Osnabrück. Universität Osnabrück.

Mitkov, R. (2002). *Anaphora resolution*. Longman, London.

Ng, V. and Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111.

Oard, D. W. (1997). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval*. American Association for Artificial Intelligence.

Paaß, G., Kindermann, J., and Leopold, E. (2004). Learning prototype ontologies by hierarchical latent semantic analysis. In *Knowledge Discovery and Ontologies*, Pisa, Italy.

Poesio, M. (2004). The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proceedings of SIGDIAL*, Boston.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to resolve bridging references. In *Proceedings of the ACL*, pages 143–150.

Poesio, M., Schulte im Walde, S., and Brew, C. (1998). Lexical clustering and definite description interpretation. In *Proc. of the AAAI Spring Symposium on Learning for Discourse*, pages 82–89, Stanford, CA.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124.

Sowa, J. (2000). *Knowledge Representation: Logical, Philosophical and Computational Foundations*. Brooks Cole Publishing Co., Pacific Grove, CA.

Stührenberg, M., Goecke, D., Diewald, N., Mehler, A., and Cramer, I. (2007). Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of the Linguistic Annotation Workshop (Merger of NLPXML 2007 and FLAC 2007)*, Prague, Czech Republic.

Versley, Y. (2007). Using the web to resolve coreferent bridging in german newspaper text. In Rehm, G., Witt, A., and Lemnitzer, L., editors, *Data Structures for Linguistic Resources and Applications. Proceedings of GLDV-2007*, pages 253–261, Tübingen. Gunter Narr Verlag.

Vieira, R. and Poesio, M. (2000). An empirically-based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Webber, B. L. (1988). Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistic (ACL–88)*, pages 113–122, State University of New York at Buffalo. June 27-30 1988.

Pablo Gamallo Otero, Gabriel Pereira Lopes, Alexandre Agustini

Automatic Acquisition of Formal Concepts from Text

This paper describes an unsupervised method for extracting concepts from *Part-Of-Speech* annotated corpora. The method consists in building bi-dimensional clusters of both words and their lexico-syntactic contexts. The method is based on *Formal Concept Analysis* (FCA). Each generated cluster is defined as a *formal concept* with a set of words describing the extension of the concept and a set of contexts perceived as the intensional attributes (or properties) valid for all the words in the extension. The clustering process relies on two concept operations: *abstraction* and *specification*. The former allows us to build a more generic concept by intersecting the intensions of the merged concepts and making the union of their extensions. By contrast, specification makes the union of the intensions and intersects the extensions. The result is a concept lattice that describes the domain-specific ontology underlying the training corpus.

1 Introduction

The pervasive and explosive proliferation of information systems requires a better understanding, control, and management of the conceptual structure underlying information. Solutions to represent conceptual structures are emerging in the form of *ontologies*, i.e., computer-based repositories of formal concepts about application domains (Reinberger et al., 2003). It is broadly assumed that, not only database schemas or semi-structured data, but also textual sources play an important role to extract concepts and learn ontologies. Recent work in ontology learning has started to develop methods for the automatic construction of conceptual structures (Philipp Cimiano, 2005). This is typically done in an unsupervised manner on the basis of text corpora relevant for the domain of interest. We have opted for extraction techniques based on unsupervised learning methods since these do not require specific external domain knowledge such as thesauri, and the portability of these techniques to new domains is much better.

This paper describes an unsupervised method for extracting concepts from *Part-Of-Speech* annotated corpora. The method consists in building bi-dimensional clusters of both words and their lexico-syntactic contexts. Each cluster, which represents a concept such as "entities in danger" is the result of either merging or unifying their constituents (i.e., words and contexts). In the last step of the method, we will identify prototypical

constituents from the generated clusters. These prototypes will be used as concept centroids in the last step of our method: word classification.

The basic intuition underlying our corpus-based approach is that similar concepts can be aggregated to generate either more specific or more generic concepts, without inducing odd associations between contexts and words. A new concept is generated by specification if we make the union of the constituent contexts (intension expansion) while the words are intersected (extension reduction). A new concept is generated by abstraction if the lexico-syntactic contexts are intersected (intension reduction), while we make the union of the constituent words (extension expansion). Intersecting words and contexts in an accurate way allows us to generate tight clusters with prototypical constituents. The theoretical background of our work is based on is *Formal Concept Analysis* (FCA). The clusters we acquired have all the features of "formal concepts" in FCA. Figure 1 shows a cluster of words and lexico-syntactic contexts learnt by our system. The cluster represents a formal concept with a word extension and a descriptive intension. The clustering algorithm only selects those contexts that can co-occur with all words in the extensional set.

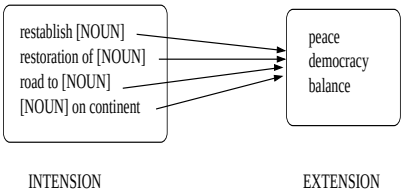


Abbildung 1: *A bi-dimensional cluster generated by our method*

2 Related Work

Local syntactic contexts have been largely used to extract classes (or concepts) of semantically similar words. Yet, approaches differ in the way they define word similarity. Some of them assume that two words are similar if they co-occur with a number of identical local contexts (Grefenstette, 1994; Lin, 1998). Semantic similarity is then computed by using the whole set of local contexts associated with each word. Unfortunately, the contexts of a word are usually very heterogeneous and multidimensional. They impose different selection restrictions and then select for different semantic facets or senses of a word. For instance, the noun *organisation* appears, at least, in two different types of contexts: those selecting for temporal events (*organisation* of the party, to finish the *organisation*, etc.) and those requiring institutions (hired by the *organisation*, the president of the *organisation*, etc.). Given such a contextual diversity, this word can be semantically

associated to a list of very heterogeneous nouns: *procedure*, *action*, *company*, *ministry*,.... This "absolute" view of semantic similarity leads to collapsing heterogeneous contextual information onto a single axis.

In order to induce semantically homogeneous lists of words, other approaches do not compare the semantic similarity between words, but between $< context, word >$ pairs and sets of those pairs. These sets are perceived as lexico-semantic concepts (also called "classes" or "selection types") (Pereira et al., 1993; Roth, 1995). Given two vocabularies, W and $CNTX$, which represent respectively the set of words and the set of local contexts, a class ou concept is defined as a pair $< CNTX', W' >$, where $CNTX' \subseteq CNTX$ and $W' \subseteq W$. In this model, the same word or context can in principle belong to more than one concept. So, the positive side of these approaches is that they try to take into account linguistic polysemy. Some difficulties arise, however, in the process of class generation. Those approaches propose a clustering algorithm in which each concept is represented by the centroid distributions of all of its members. This is in conflict with the fact that many words and local contexts can significatively involve more than one semantic dimension. As a result, the clustering method turns out to be too greedy since it overgenerates many wrong associations between words and local contexts. For instance, the work by Roth induced a particular concept containing the association between verbs (viewed as local contexts) such as *cost, play, spend, be, ...* and nouns like *money, role, fund, part*, etc. See Figure 2. This concept contains several wrong association pairs: for instance, $< cost[N], role >$, $< play[N], fund >$, etc. Besides that, there also are too broad-sense words (*be, use, part, time, ...*), which may belong to almost any concept.

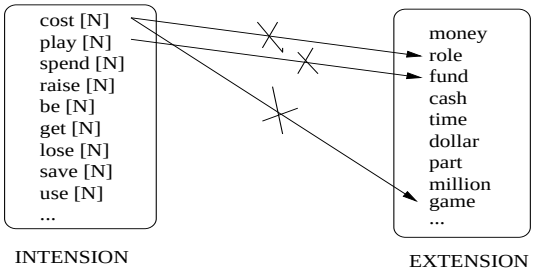


Abbildung 2: *An excerpt of a bi-dimensional cluster appearing in (Roth, 1995)*

To avoid these drawbacks, a more recent approach tried to limit the information contained in the centroids by introducing a process of "clustering by committee" (Pantel and Lin, 2002). The centroid of a cluster is constructed by taking into account only a subset of the cluster members. This subset, called "committee", contains the more representative members (prototypes) of a concept. So, the main and more difficult task

of such an approach is to first identify a list of committees, i.e., a list of semantically homogeneous clusters. Committees represent basic linguistic concepts of similar words and are useful for word classification.

Other approaches also try to identify homogenous clusters representing basic semantic concepts. The main difference with regard to the former method is that each basic cluster is constituted, not by similar words, but by a set of similar local contexts (Faure, 2000; Pantel and Lin, 2000; Allegrini et al., 2003; Reinberger and Daelemans, 2003; Gamallo et al., 2005). The method is focused on computing the semantic similarity between lexico-syntactic contexts. Words are no more seen as entities to be clustered but as features of contexts. These are taken as the main objects in the clustering process. As lexico-syntactic contexts turn out to be less polysemic than words, these approaches assume that searching for concepts of homogeneous contexts is easier and more efficient than to find tight concepts of semantically related words. The main problem, however, is that the basic clusters of contexts identified in the first step tends to be very small and specific. The average size of a basic cluster is only two members. In order to generate larger concepts, most of these approaches require a second step with a greedy clustering process. Unfortunately, this greedy clustering step tends to overgenerate many context-word associations.

The method proposed in this paper is close to the last type of approaches described in the previous paragraph. Our main contribution is the use of very restrictive operations (specification and abstraction) in the process of building tight clusters. Thanks to these constraints on clustering, we try to solve the overgeneration problem. A tight cluster will be defined as a bi-dimensional entity consisting on both a set of words and a set of contexts, if only if each word is semantically associated to all contexts of the cluster.

3 Theoretical Background: Formal Concept Analysis

Formal Concept Analysis (FCA) (Hereth et al., 2003; Priss, 2006) is a particular method of data analysis and knowledge representation based on *Galois lattice* (also called *concept lattice*). In this framework, a concept is defined as a dual unit consisting of two parts: a set of objects (the extension of the concept) and a set of attributes or properties valid for all the objects in the concept (its intension). The family of these concepts obeys the mathematical axioms defining a lattice.

The main idea underlying FCA is to argue that a concept lattice is an efficient tool for several applications, such as lexical database design, ontology learning (Philipp Cimiano, 2005), knowledge acquisition, or conceptual clustering. In this paper, our contribution is to use a concept lattice to design a particular strategy of conceptual clustering.

3.1 Formal Concepts

To define formal concepts, FCA starts with the notion of formal context. A *formal context* is a triple $\Bbbk := (O, A, R)$, where O is a set of objects, A, a set of attributes, and $\Re$ a binary relation between O and A, i.e. $\Re \subseteq O \times A$. A concept lattice of $\Bbbk$ is a partial order over all pairs of the form (E, I), where $E \subseteq O$, $I \subseteq A$ s.t.:

$$\begin{aligned} E &= \{o \in O \,|\forall a \in I, oRa\} \\ I &= \{a \in A \,|\forall o \in E, oRa\} \end{aligned} \tag{1}$$

The relationship oRa (which belongs to $\Re$) is read "the object o has the attribute a". The pair (E, I) is called a *formal concept*, where E is the extension of the concept (i.e., the set of objects it comprises), and I is its intension, i.e., the set of attributes shared by all members of the concept's extension. Partial order is defined as follows: if (E_1, I_1) and (E_2, I_2) are formal concepts, we define a partial order $\leq$ by saying that $(E_1, I_1) \leq (E_2, I_2)$ whenever $E_1 \subseteq E_2$. Equivalently, $(E_1, I_1) \leq (E_2, I_2)$ whenever $I_1 \subseteq I_2$. Every pair of concepts in this partial order has a unique greatest lower bound (meet) and a unique least upper bound (join), so it satisfies the axioms defining a lattice. The greatest lower bound of (E_1, I_1) and (E_2, I_2) is the concept with objects $E_1 \cap E_2$ and attributes $I_1 \cup I_2$. The least upper bound of (E_1, I_1) and (E_2, I_2) is the concept with attributes $I_1 \cap I_2$ and objects $E_1 \cup E_2$.

3.2 A Toy Example

The cross table 1 depicts a small formal context. The elements on the left side are objects while the elements at the top are attributes (or properties) of the objects. The relationship between them is represented by crosses. In this toy example, the objects are some states and the attributes describe whether they have a president, prime-minister, or a king, whether they belong to the European Union or whether they are ruled by Islamic principles.

Figure 3 represents the concept lattice of the formal context in Table1. In the diagram, each node represents a formal concept, consisting of a set of objects noted below (the extension) and a set of attributes appearing above (the intension). A concept c_1 is a subconcept of a concept c_2 if only if there is a path of descending edges from the node representing c_2 to the node representing c_1. The label of an object o is always attached to the node representing the smallest concept with o in its extension. In Figure 3, the label "Iran" is in the concept with extension {'I', 'PK'} and intension {'ir', 'pr'}. There is no smaller concept with 'I' in the extension. Conversely, the label of attribute a is always attached to the node representing the largest concept with a in its intension. For instance, the label "kingdom" is in the concept with extension {'B', 'A'} and intension {'k'}. There

Tabelle 1: A formal context of "states"

	president (pr)	prime-minister (pm)	european union (eu)	kingdom (k)	islamic rules (ir)
Belgium (B)		X	X	X	
Portugal (P)	X	X	X		
Pakistan (PK)	X	X			X
Iran (I)	X				X
Saudi Arabia (A)				X	X

is no larger concept with 'k' in the intension set. The top and bottom concepts in a concept lattice are special. The top concept is the largest one since it has all objects in its extension. Its intension is often empty but does not need to be empty. The bottom concept is the smallest one and has all attributes in its intension. Its extension is empty when there are at least two attributes that are mutually incompatible. For instance, "being a kingdom" ('k') and "to have a President" ('pr'). The top concept can be considered as the "universal" concept and the bottom one the "null" concept of a formal context.

A central notion of a concept lattice is the duality of concepts. This duality implies that if one makes the sets of extensions larger, they correspond to smaller sets of intensions, and vice versa. In Figure 3 those nodes with larger extensions (at the top) tend to have only one attribute. On the bottom, nodes with larger intensions have only one object. However, in the middle of the diagram, we find more balanced nodes, i.e., concepts with a similar number of elements in both the extension and the intension. In our toy example, these balanced concepts represent useful notions to describe some political systems of states. For instance, the concept characterised by the properties "to have a President" and "to have a Prime-Minister" represents those states that are standard republics. Islamic republics, on the other hand, can be represented by the concept containing the properties "islamic rules" and "to have a President". We claim that balanced concepts tend to be significant and meaningful nodes in any ontology, terminology, or lexical database.

3.3 Building a Concept Lattice by Clustering of Words and Contexts

Following the ideas introduced above, we can build a lattice of formal concepts consisting of two linguistic dimensions. One dimension is the intension definition, i.e., a set of similar lexico-syntactic contexts with the same selection restrictions. The other one is its extension, i.e., the set of words appearing in such contexts and satisfying their semantic requirements. When the intension is very specific because it contains a large set of contexts, then the extension tends to be small. A lexico-syntactic context can be defined

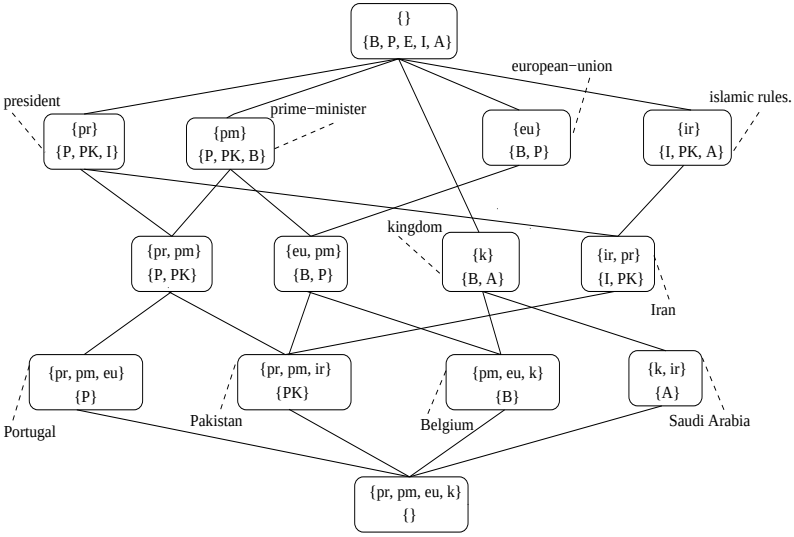


Abbildung 3: *A concept lattice from the formal context depicted in Table 1*

as a linguistic pattern constituted by a lexical word, a syntactic relation, and a morpho-syntactic position. For instance, "president of [NOUN]" is the lexico-syntactic context of nouns such as "Portugal", "Belgium", "Real Madrid", "republic", or "company", i.e. nouns denoting institutions with a president. In this particular application, co-occurrence in a corpus turns out to be the specific binary relationship between extensions and intensions. So, within a formal concept, each word in the extension "co-occurs" with each lexico-syntactic context in the intension.

New formal concepts are generated by means of a clustering process endowed with two complementary operations: *specification* and *abstraction*. If two similar formal concepts, FC_1 and FC_2, defined respectively as the pairs $< CNTX_1, W_1 >$ and $< CNTX_2, W_2 >$, are aggregated into a new concept, we can opt for two different operations:

specification: $FC_1 \Theta FC_2$, which represents a more specific concept whose intension is the set of contexts $CNTX_1 \cup CNTX_2$, and the extension the word set $W_1 \cap W_2$.

abstraction: $FC_1 \Phi FC_2$, which represents a more generic concept whose intension is the intersection $CNTX_1 \cap CNTX_2$, and the extension the union $W_1 \cup W_2$.

The clustering method we will describe in the following section makes use of these two operations. The resulting concepts generated by such operations give rise to a

concept lattice. The more balanced concepts in that lattice will be the startpoint (i.e., centroids) of a further process: word classification.

4 The Method

Our method consists of 4 steps. In Step I, we describe the linguistic process allowing us to create context vectors. Step II introduces a clustering algorithm relying on the specification operation. The aim is to identify a list of balanced concepts. In Step III, these concepts are merged by a hierarchical clustering and the abstraction operation. As a result, we build a concept lattice with several unrelated abstract formal concepts at the top level. The specific information involved in the definition of each top abstract concept will be used in the following classification step. Finally, in Step IV, further words are classified and assigned to the appropriate formal concepts.

4.1 Step I: Building Context Vectors

In this step, lexico-syntactic contexts will be represented as vectors of word lemmas. The basic value of each vector position is the co-occurrence frequency between the context and the corresponding lemma. The whole vector space can be perceived as the *Formal Context* from which we will extract formal concepts.

To create the vector space, we first need to identify lexico-syntactic contexts from texts. We start by POS tagging the input corpus. Then, we use basic pattern matching techniques to identify potential binary dependencies. From each binary dependency, two complementary lexico-syntactic contexts are selected. Table 2 shows some representative examples. A lexico-syntactic context defines a set of semantically related words. Given a binary dependency:

to (threat, health) ,

two templates are selected: < danger to [NOUN] >, which represents the set of nouns that can appear after "danger to", for instance, "health", "peace", "stability", etc. On the other hand, < [NOUN] to health > represents the set of nouns appearing before "to health": "danger", "access", "threat", etc. We follow the notion of *co-requirement* introduced in (Gamallo et al., 2005).

Note that *lobj* represents the relationship between a verb and the noun immediately appearing at its left; *robj* is the relationship between a verb and the noun appearing at its right. On the other hand, *modAdj* is the relationship between a noun and its adjective modifier and *modN* is the relation between two nouns: the head and its modifier.

Finally, each lexico-syntactic context is associated to its co-occurring words to build the vector space.

Binary Dependencies	Contexts
to (threat, health)	< threat to [NOUN] > < [NOUN] to health >
of (import, sugar)	< import of [NOUN] > < [NOUN] of sugar >
robj (approve, law)	< approve [NOUN] > < [VERB] law >
lobj (approve, president)	< president [VERB] > < [NOUN] approve >
modAdj(legal, document)	< legal [NOUN] > < [ADJ] document >
modN(area, protection)	< protection [NOUN] > < [NOUN] area >

Tabelle 2: Some binary dependencies and their corresponding lexico-syntactic contexts.

4.2 Step II: Extracting Balanced Concepts

4.2.1 Filtering Concepts

We start by filtering out lexico-syntactic contexts that are sparse in the training corpus. A context is sparse if it has high word dispersion. Dispersion is defined as the number of different word lemmas occurring with a lexico-syntactic context divided by the total number of different word lemmas in the training corpus. So, the vector space is only constituted by those lexico-syntactic contexts whose word dispersion is lower than an empirically set threshold.

4.2.2 Context Similarity

Then, for each context with low dispersion, we compute its top-k similar ones, where $k = 5$, using a Dice coefficient as similarity measure (Frakes, 1992).

Similarity between two lexico-syntactic contexts $cntx_1$ and $cntx_2$ is computed as follows:

$$Dice(cntx_1, cntx_2) = \frac{2 * \sum_i min(f(cntx_1, w_i), f(cntx_2, w_i))}{F(cntx_1) + F(cntx_2)} \quad (2)$$

where $f(cntx_1, w_i)$ represents the number of times $cntx_1$ co-occurs with the word lemma w_i. $F(cntx_i$ stands for the absolute frequency of $cntx_1$. This is the similarity score used to build the top-5 lists of similar contexts. For instance, Table 3 shows a list with the 5 most similar contexts to "threat to [N]", according to the information extracted from the corpus *Europarl*.

Tabelle 3: The 5 most similar contexts to "threat to [N]"

{threat to [N]}	{risk to [N]}	0.213
{threat to [N]}	{endanger [N]}	0.191
{threat to [N]}	{[N] aspect}	0.172
{threat to [N]}	{damage [N]}	0.171
{threat to [N]}	{guarantee of [N]}	0.155

Tabelle 4: The top-5 concepts built around the context "threat to [N]"

00231	{threat to [N], risk to [N]}	{health, environment, security, price, peace, stability}
00232	{threat to [N], endanger [N]}	{whole, democracy, peace, life, health, environment, security, stability}
00233	{threat to [N], [N] aspect}	{welfare, safety, employment, health, security}
00234	{threat to [N], damage [N]}	{employment, integrity, peace, life, health, environment, fishing, stability}
00235	{threat to [N], guarantee of [N]}	{safety, democracy, peace, job, freedom, security, stability}

4.2.3 Basic Concepts (input of clustering)

The basic concepts used as input of the clustering process are extracted from these ranked lists. Given the top-5 list associated to a lexico-syntactic context (and the set of word lemmas it classifies), we build 5 basic concepts by aggregating that context to each one in the list. The words in the extension are those co-occurring with both contexts. Table 4 shows the five basic concepts associated to the context "threat to [N]" that were extracted from the ranked list in 3. These basic concepts are quite generic since their intension has only two attributes (2 contexts). They will be the input of the process of clustering by specification.

4.2.4 Clustering by Specification

The first basic concept, 00231, is taken as the centroid since it is constituted by the top-1 similar context to "threat to [N]" (see again Table 4). The clustering process consists in aggregating the remaining concepts together around the identified centroid if only if they share more than 50% of the word lemmas. All aggregations are made using the operator of "specification" since each generated concept is obtained by intersecting the two word sets of each aggregated concept. As a result, we obtain:

FC_{37}	{endanger [N], damage [N], threat to [N], risk to [N]}	{health, environment, peace, stability}

which is the result of two specification operations:

$$FC_{37} = 00231\ \Theta\ 00232\ \Theta\ 00234$$

Here, clustering involves the centroid, 00231, and two concepts, 00232 and 00234, which satisfy the similarity condition (share at least 50% of words). Note that the specification operation allows us to build concepts with a more balanced relationship between the extension and the intension. This process is repeated for the other top-5 lists of similar contexts extracted from the corpus. The set of balanced concepts generated at the end of the process is the input of the following clustering step.

4.3 Step III: Generating Abstract Concepts by Hierarchical Clustering

A standard hierarchical clustering takes as input the specific and balanced concepts built in the previous step to generate more generic ones. For this purpose, we make use of an open source software: Cluster 3.0[1]. In this step, we use the operation of abstraction to build the successive aggregations. So, each generated concept is constituted by both the union of word sets and the intersection of contexts. Table 5 illustrates the concept lattice organising the information around $NODE_{77}$. This top-level concept is obtained from two successive abstractions:

$$NODE_{77} = FC_{37}\,\Phi\,NODE_{30}$$
$$NODE_{30} = FC_{420}\,\Phi\,FC_{202}$$

Words and contexts organised around $NODE_{77}$ seem to characterise the abstract concept of "entities in danger". Note that the concepts we are able to learn (e.g., entities in danger) do not try to represent word senses as the synsets do in WordNet. Rather, they characterise top-level concepts of an upper-level ontology. In our notation, concepts labeled as $NODE_i$ stands for those generated by abstraction, whereas those labeled with FC_i represent concepts generated by specification. Figure 4 depicts the diagram representation of Table 5. This is another visualisation of the same lattice sample.

In our framework, the same word lemma can belong to the extension of different top-level concepts. For instance, *environment*, which is a member of $NODE_{77}$, is also a member of another concept aggregating nouns such as *agriculture*, *interior*, *justice*, *culture*, and *finance*, by their association with contexts like "minister of [N]", "ministry of [N]", or "minister for [N]".

[1] *http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm*

Tabelle 5: Hierarchical construction of the generic formal concept $NODE_{77}$

$NODE_{77} : NODE_{30} \Phi FC_{37}$	{endanger [N]}	{health, life, patient, environment, peace, stability, quality}
$NODE_{30} : FC_{202} \Phi FC_{420}$	{endanger [N], risk to [N]}	{health, life, patient, environment, quality}
FC_{202}	{*endanger [N]*, risk to [N], expense of [N]}	{*health*, life, patient, environment}
FC_{420}	{*endanger [N]*, risk to [N], plant [N]}	{*health*, life, quality}
FC_{37}	{damage [N], *endanger [N]*, risk to [N], threat to [N]}	{*health*, environment, peace, stability}

Finally, if we observe more carefully Table 5 and Figure 4, we find out that *health* and "*endanger [N]*" are the only elements appearing in the three specific bottom-level concepts. They be considered as the prototypical or more representative constituents of these concepts with regard to the training corpus (they are in italic in the table). Prototypical elements will play an important role in the following step: word classification.

4.4 Step IV: Word Classification

So far, the generated clusters have been loosing relevant information step by step, since they were aggregated using intersecting operations. Besides that, the intersecting aggregations did not allow us to infer context-word associations that were not attested in the training corpus. As has been mentioned above, our objective was to design a very restrictive clustering strategy so as to avoid overgeneralisations.

In order to both reintroduce lost information and learn new context-word associations, the last step aims at assigning more word lemmas to the balanced concepts generated in the first clustering process. A word is assigned to one or more concepts in the following way:

We start by identifying the centroids used for classification. Given a concept, the representative centroid is constituted by the word lemmas and contexts that were considered as prototypes in the abstraction process (Step 2). For instance, the centroid extracted from concepts FC_{420}, FC_{202}, and FC_{37}, during the construction of $NODE_{77}$ is: $< \{endanger[N]\}, \{health\} >$. If a lemma fills the *classification conditions* imposed by this centroid, then it is assigned to the three balanced concepts in the example.

The classification conditions that a candidate lemma must fill are two: First, it must be *similar* to those word lemmas appearing in the centroid. Second, it must co-occur in the training corpus with the contexts of the centroid.

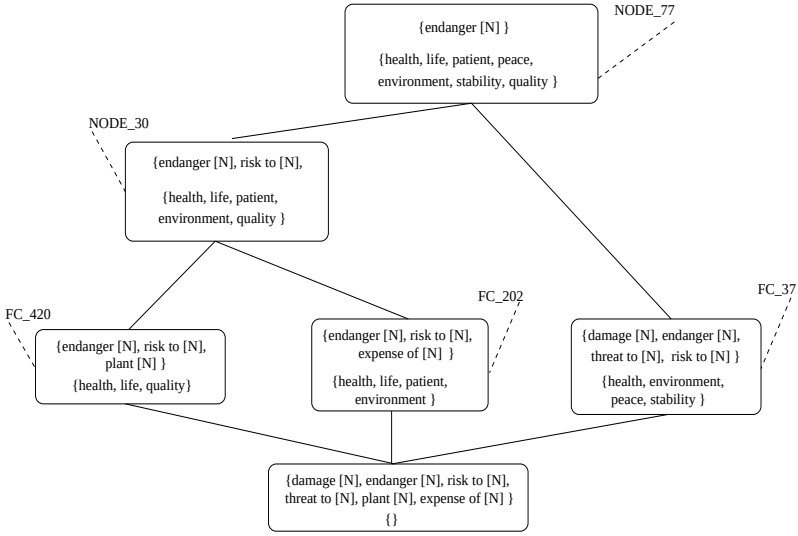


Abbildung 4: *A diagram representation of the concept lattice depicted in Table 5*

To measure similarity between word lemmas, we used the same coefficient as for context similarity: Dice score. In addition, each lemma was provided with a list containing its top-5 most similar ones. So, two word lemmas, w_i an w_j, are considered to be similar if only if w_i is in the top-5 list of w_j, or conversely, if w_j is in the top-5 list of w_i.

At the end of the classification step, our system was able to assign "security", "democracy" , "growth", and "energy" to the concepts organised around the top-level concept of *entities in danger*. Note that the acquired formal concepts refer to domain-dependent classes.

5 Experiments and Evaluation

Experiments have been carried out using two different text corpora. A Portuguese corpus with 10 million tokens extracted from the general-purpose journal *O Público*, and an English excerpt (3 million tokens) of the European Parliament Proceedings (*EuroParl*). The Portuguese corpus was *POS* tagged with TreeTagger[2], using our own training corpus and lexicon[3]. The English corpus was tagged with an open source analizer: Freeling (Carreras et al., 2004).

[2] *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html*
[3] Portuguese parameters can be downloaded in *http://gramatica.usc.es/∼gamallo/tagger.htm*

Tabelle 6: Corpus Data

	Balanced Concepts	Abstract Concepts	Classif.	Accuracy of Classif.
Público	264	91	492	92%
EuroParl	227	68	226	94%

Table 6 depicts the number of balanced and abstract concepts extracted from each corpus, as well as the number of word classifications. Let's remember that balanced concepts were the output of Step II and abstract concepts the one of Step III. The extraction was only focused on nouns and nominal contexts. Note that not many abstract classes were learnt. This is in accordance with the basic ideas underlying formal ontology.

Measuring the correctness of the acquired lexico-semantic classes is not an easy task. We are not provided with a gold standard to which results can be compared. As the acquired concepts are corpus-dependent and do not represent word senses, there is no pre-existing ontology nor thesaurus containing the type of information our system is able to learn. Indeed, most concepts we learnt refer to domain-dependent knowledge. For instance, the class of world regions with internal conflicts and genocides: *Kosovo, Balcans, Serbia, Colombia, Chechnya, East Timor, Sierra Leona, region*. These word lemmas appear in contexts such as "conflict in [N]", "war in [N]", and "genocide in [N]". Another domain-dependent concept we learnt is the class of Portuguese towns with Bishop: *Viseu, Braga, Lisboa, Beja, Coimbra, Leiria, Guarda*. These names of towns co-occur with contexts such as "bispo de [N]", "diocese de [N]", "distrital de [N]", and "distrito de [N]"[4].

Other acquired concepts represent more heterogeneous classes and consist of open sets of words. For instance, we extracted an open set of entities in danger ($NODE_{77}$ above), a set of different forces that can be involved in a process (*threat, obstacle, access, impetus, contribution, ...*), a set of negative actions (*expulsion, terror, cleansing, genocide, massacre, destruction, atrocity, fighting, terrorism, ...*), different types of statements (*remarque, comment, observation, word, point, statement, recommendation, suggestion, argument, request,* ...), and so on.

To evaluate the quality of the formal concepts we have acquired, we set a subjective evaluation protocol focused on the accuracy of word classification. Each word assignment to a concept was judged as correct or incorrect by a human evaluator. An assignment was considered as correct if the assigned word lemma is *semantically required* by all the lexico-syntactic contexts defining the concept. The 4th column of Table 6 shows the

[4]These contexts can be translated as follows: "Bishop of [N], "diocese of [N]", "District of [N]", and "District of [N]", respectively.

accuracy score. In fact, this evaluation measures the amount of overgeneration produced by the system. Overgeneration is about 8% in *O Público* and 6% in *EuroParl*.

In further research, we intend to develop a process of context classification. In this process, each formal concept will be assigned lexico-syntactic contexts that were not involved in the previous clustering steps. This way, we will be able to learn better intensional definitions of concepts.

6 Acknowledgements

This work has been supported by Ministerio de Educación y Ciencia of Spain, within the project GaricoTerm, ref: BFF2003-02866.

Literatur

Allegrini, P., Montemagni, S., and Pirrelli, V. (2003). Example-based automatic induction of semantic classes through entropic scores. *Linguistica Computazionale*, pages 1–45.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). An open-source suite of language analyzers. In *4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Faure, D. (2000). *Conception de méthode d'aprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. PhD thesis, Université Paris XI Orsay, Paris, France.

Frakes, W. (1992). *Information Retrieval. Data Structures and Algorithms*. Prentice Hall.

Gamallo, P., Agustini, A., and Lopes, G. (2005). Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA.

Hereth, J., Stumme, G., Wille, R., and Wille, U. (2003). Conceptual knowledge discovery - a human-centered approach. *Journal of Applied Artificial Intelligence*, 17(3):288–301.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal.

Pantel, P. and Lin, D. (2000). An unsupervised approach to prepositional phrase attachment using contextually similar words. In *ACL'00*, pages 101–108, Hong Kong.

Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.

Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 30th Annual Meeting of the Association of Comptutational Linguistics*, pages 183–190, Columbos, Ohio.

Philipp Cimiano, Andreas Hotho, S. S. (2005). Learning concept hierarchies from text corpora using formal concept anaylsis. *Journal of Artificial Intelligence Research (JAIR)*, 24:305–339.

Priss, U. (2006). Formal concept analysis in information science. *Information Science and Technology*, 40:521–543.

Reinberger, M.-L. and Daelemans, W. (2003). Is shallow parsing useful for unsupervised learning of semantic clusters? In *4th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-03)*, pages 304–312, Mexico City.

Reinberger, M.-L., Spyins, P., Daelemans, W., and Meersman, R. (2003). Mining for lexons: applying unsupervisded learning methods to create ontology bases. *Lecture Notes in Computer Science*, 2888:803–819.

Roth, M. (1995). Two-dimensional clusters in grammatical relations. In *Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity (AAAI 1995)*.

Autorenverzeichnis

Alexandre Agustini
Computer Science Faculty
PUCRS University
Av. Ipiranga, 6681, Prédio 32, sala 505
90619-900 Porto Alegre - RS
Brazil
alexandre.agustini@pucrs.br

Maja Bärenfänger
Institut für Germanistik
Justus-Liebig Universität Gießen
Otto-Behaghel-Str. 10 D
35394 Gießen
Maja.Baerenfaenger@germanistik.uni-giessen.de

Christian Chiarcos
Institut für Linguistik
Universität Potsdam
Karl-Liebknecht-Str. 24-25
14476 Golm
chiarcos@uni-potsdam.de

Daniela Göcke
Fakultät für Linguistik und
Literaturwissenschaft
Universität Bielefeld
Postfach 10 01 31
33501 Bielefeld
daniela.goecke@uni-bielefeld.de

Mirco Hilbert
Institut für Germanistik
Justus-Liebig Universität Gießen
Otto-Behaghel-Str. 10 D
35394 Gießen
Mirco.Hilbert@germanistik.uni-giessen.de

Kai-Uwe Kühnberger
Institut für Kognitionswissenschaft
Universität Osnabrück
Albrechtstraße 28
D-49076 Osnabrück
kkuehnbe@uos.de

A Kumaran
Multilingual Systems Research
Microsoft Research India
196/36 2nd Main Sadashivnagar
Bangalore 560 080 India
A.Kumaran@microsoft.com

Henning Lobin
Institut für Germanistik
Justus-Liebig Universität Gießen
Otto-Behaghel-Str. 10 D
35394 Gießen
Henning.Lobin@germanistik.uni-giessen.de

Gabriel Pereira Lopes
Computer Science Department
New University of Lisbon
Quinta da Torre
2829 -516 Caparica
Portugal
gpl@di.fct.unl.pt

Harald Lüngen
Institut für Germanistik
Justus-Liebig Universität Gießen
Otto-Behaghel-Str. 10 D
35394 Gießen
Harald.Luengen@germanistik.uni-giessen.de

Ranbeer Makin
Microsoft Research India
196/36, 2nd Main Road, Sadashivnagar
Bangalore 560080 India
t-ranbm@microsoft.com

Uwe Mönnich
Seminar für Sprachwissenschaft
Theoretische Computerlinguistik
Universität Tübingen
Wilhelmstraße 19
D-72074 Tübingen
um@sfs.uni-tuebingen.de

Pablo Gamallo Otero
Department of Spanish Language
University of Santiago de Compostela
Campus Universitario Norte
1572 Santiago de Compostela
Spain
pablogam@usc.es

Vijay Pattisapu
Microsoft Research India
196/36, 2nd Main Road, Sadashivnagar
Bangalore 560080 India
t-vijpat@microsoft.com

Shaik Sharif
Microsoft Research India
196/36, 2nd Main Road, Sadashivnagar
Bangalore 560080 India
t-shaiks@microsoft.com

Maik Stührenberg
Fakultät für Linguistik und
Literaturwissenschaft
Universität Bielefeld
Postfach 10 01 31
33501 Bielefeld
maik.stuehrenberg@uni-bielefeld.de

Lucy Vanderwende
Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98075, USA
lucyv@microsoft.com

Tonio Wandmacher
Institute of Cognitive Science
University of Osnabrück
Albrechtstr. 28
49076 Osnabrueck
twandmac@uos.de