Supports

OntoWeb
http://www.ontoweb.org/

ELSNET
http://www.elsnet.org/

Sponsors

Co-operating Organisations

The Workshop Programme

08:40 - 09:45 - Opening

08:45 - 09:10 - Melina Alexa, Bernd Kreissig, Martina Liepert, Klaus Reichenberger, Lothar Rostek, Karin Rautmann, Werner Scholze-Stubenrecht, Sabine Stoye. *The Duden Ontology: an Integrated Representation of Lexical and Ontological Information*

09:10 - 09:35 - Bernardo Magnini, Manuela Speranza. *Merging Global and Specialized Linguistic Ontologies*

09:35 - 10:00 - Dietmar Rösner, Manuela Kunze. *Exploiting Sublanguage and Domain Characteristics in a Bootstrapping Approach to Lexicon and Ontology Creation*

10:00 - 11:00 - Christiane Fellbaum. *Parallel Hierarchies in the Verb Lexicon*

11:00 - 11:20 - Coffee break

11:20 - 11:45 - Aldo Gangemi, Nicola Guarino, Alessandro Oltramari, Stefano Borgo. *Restructuring WordNet's Top-Level: The OntoClean based approach*

11:45 - 12:10 - Maarten Janssen. *EuroWordNet and Differentiae Specificae*

12:10 - 12:35 - James Pustejovsky, Anna Rumshisky, José Castaño. *Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics*

12:35 - 14:00 - Lunch

14:00 - 14:25 - Roberto Navigli, Paola Velardi. *Automatic Adaptation of WordNet to Domains*

14:25 - 14:50 - Wim Peters. *Self-enriching Properties of Wordnet: Relationships between Word Senses*

14:50 - 15:15 - Sandiway Fong. *On the Ontological Basis for Logical Metonomy: Telic Roles and WordNet*

15:15 - 15:40 - Anthony R. Davis, Leslie Barrett. *Relations among Roles*

15:40 - 16:40 - Yorick Wilks. *To be announced*

16:40 - 17:00 - Coffee break

17:00 - 18:00 - Discussion: *Distinctions between Lexical and Ontological Knowledge*

Workshop Organisers

Kiril Simov
Linguistic Modelling Laboratory, CLPP, Bulgarian Academy of Sciences, Bulgaria
and
OntoText Lab. Sirma AI Ltd, Bulgaria
E-mail: kivs@bgcict.acad.bg

Nicola Guarino
National Research Council, LADSEB-CNR, Italy
Email: Nicola.Guarino@ladseb.pd.cnr.it

Wim Peters
NLP group, Department of Computer Science, University of Sheffield, England
Email: W.Peters@dcs.shef.ac.uk

Workshop Programme Committee

Table of Contents

Author Index

The Duden Ontology:

An Integrated Representation of Lexical and Ontological Information

Melina Alexa[1], Bernd Kreissig[1], Martina Liepert[1*], Klaus Reichenberger[2], Lothar Rostek[3], Karin Rautmann[1], Werner Scholze-Stubenrecht[1], Sabine Stoye[2]

[1] Bibliographisches Institut und F.A. Brockhaus AG (BIFAB), Mannheim, Germany
{Melina.Alexa, Bernd.Kreissig, Martina.Liepert, Karin.Rautmann, Werner.Scholze-Stubenrecht}@bifab.de
[2] intelligent views, Darmstadt, Germany
{k.reichenberger, s.stoye}@i-views.de
[3] FhG-IPSI, Darmstadt, Germany
rostek@ipsi.fhg.de

Abstract

We report on a data model developed for the representation of lexical knowledge for the Duden ontology. The model is the result of a cooperation between the publishing house Duden and the software company *intelligent views*. Our general aim is to create an asset pool in which all the information present in the Duden dictionaries is integrated in order to support reusability for different print and electronic products, provide solutions for language technology applications as well as support the efficient maintenance of the Duden dictionary data.

1. Introduction

In this paper we describe the data model developed for the representation of lexical knowledge for the Duden ontology. Duden is a well-known publisher of language reference products in both print and electronic form as well as products for language technology for the German language. It belongs to the publishing house Bibliographisches Institut und F.A. Brockhaus AG (BIFAB). The model described here is the result of a cooperation project between Duden and the software company *intelligent views*, which is a spin-off company of the Fraunhofer Integrated Publication and Information Systems Institute (IPSI).

Our general aim is to create a rich computational resource in which all the information present in the Duden dictionaries is integrated in order to support

- the reusability for both print and electronic products,
- the development of language technology applications as well as
- the efficient maintenance of the Duden dictionary data, for example the ten volume Duden dictionary (Duden, 1999) or the Duden spelling dictionary (Duden, 2000).

Two further considerations have been important in developing this model:

- it should be flexible enough to adjust to new emerging requirements with regards to both the dictionary structure itself as well as the production of different titles and different types of dictionaries, and
- it should at a later stage allow the representation of encyclopedic information.

Note that a significant requirement has been that the Duden print dictionaries can be produced from the constructed computational resource at least as efficiently as is currently the case.

Furthermore, an important prerequisite has influenced the modeling of the data a great deal: the computational resource to be created should not only be useful for the production of print and electronic (both on- and off-line) dictionaries. It should also be useful for solving problems such as lexical and semantic ambiguity and reference resolution for knowledge intensive and real natural language applications such as, for example, a question answering system for German, for which broad-coverage of the morphological, grammatical and semantic information of the language is necessary.

1.1. Motivation

Although the majority of the Duden dictionary data are in SGML format, the markup of each dictionary is strongly print oriented rather than content oriented. For each of the SGML-based dictionaries there is a Document Type Definition (DTD) according to which the lexicographers maintain their data. Corrections or other modifications of existing lemmas and their properties as well as addition of new lemmas take place separately for each Duden title. This means that if, for example, a lexicographer modifies a lemma for the Duden dictionary *Duden – Fremdwörterbuch* (Duden, 2001a), the reference volume for the correct spelling of foreign words in German, each entry for the modified lemma in other Duden dictionaries, e.g. the Duden spelling dicitionary (2000) or Duden (2001b), has to be modified or updated manually. This is not only inefficient with regard to time but it is also prone to errors and inconsistencies. In contrast, the formal explicit representation of the Duden dictionary entries in a single knowledge base supports the administration and maintenance of dictionary data in an efficient, consistent and systematic manner.

[*] Since April 2002 at SFS, Universität Tübingen, Germany.

A further aspect concerns the additional possibilities offered by an explicit representation of all information relevant to each dictionary entry of the Duden data: depending on the quality of the data model it will be possible to generate different 'sub-lexicons' from a single data pool. These are, in principle, nothing more than different 'views' of the knowledge stored in the data pool. Examples of such sub-lexicons may be a list of all compounds in the Duden dictionaries, or a differentiated system of lexemes with their morphological (e.g. part of speech, gender), grammatical (e.g. subcategorization) and semantic (e.g. synonyms) information.

1.2. Related work

The work described in this paper relates to research on knowledge representation for lexical and semantic as well as for ontological information for the purposes of dictionary production and for natural language applications. It has to be emphasized, though, that it is our particular needs as publisher, our abilities and the tools supporting our work which guide the reported work in the first instance and not theoretical considerations. For this reason our main focus is not to construct *the* most expressive model for the representation of lexical and semantic representation but rather the construction of a large scale resource to be used for the efficient production of our dictionaries and for NLP applications.

Unlike Wordnet (Fellbaum, 1998), EuroWordNet (Vossen, 1998) and GermaNet (Hamp & Heldweg, 1997; Kunze, 2000) the Duden Ontology integrates extensive morphosyntactic properties of denotations with ontological information about their senses (see section 2). With regard to morphosyntactic information, this is represented in an extensive manner in the Duden Ontology, whereas WordNet and WordNet-like systems use elementary part-of-speech information and sub-categorization frames.

In contrast to the project WiW - Wissen über Wörter (Müller-Landmann, 2000; 2001), instead of a relational model we have opted for an object-oriented approach, which is advantageous for factorizing common information and supports inheritance of relations and attributes. A further point which distinguishes our work from the WiW-project is that we make use of the existing dictionary assets of Duden and therefore do not start from scratch. This allows us to build a comprehensive resource within a relatively short time and even more importantly to evaluate the expressiveness and suitability of the implemented model for our needs.

There are similarities between our approach and that of the Mikrokosmos project (Mahesh & Nirenburg, 1995): We too make a clear distinction between the representation of language-specific and language-neutral information. In our terminology language specific information is represented by *term* objects, whereas *concept* objects are used for representing language-neutral information (see section 2.1). One of the differences between the two projects is that the Duden Ontology integrates both kinds of information within a single resource, whereas the Mikrokosmos project uses two apparently separate databases, one for the lexicon and one for the ontology, for storing denotations and denotation-neutral concepts.

There are parallels of our work with the TransLexis conceptual schema (Bläser, 1995) with the distinction between lemma, homograph and sense. TransLexis is based on a relational model and has been driven by requirements for multilingual terminology management.

Currently, the Duden Ontology does not include an automatic classifier for classifying defined concepts on the basis of formal concept definitions, as for example the GALEN ontology and its related technology does (Rogers et al., 2001, Rector et al. 1998). With the exception of simple inference mechanisms, such as inheritance or relation path definition, the Duden Ontology does not feature a full-fledged inference engine.

2. Data model for the Duden Ontology

The Duden data model is based on a concept-oriented representation which offers the possibility of defining semantic relations between the concepts. In addition, it provides the hook for an integration of encyclopaedic data as well as for the representation of factual knowledge at a later phase.

To this end, the vocabulary of the Duden volumes is classified in a rigid manner according to a generic hierarchy relation. This is similar to WordNet where the synsets play the role of the concepts. In order to provide the hook for representing facts an explicit distinction between individuals and concepts (word senses) is necessary, which results in the creation of an ontology. According to our definition there are two essential features of an ontology:

- a classification of concepts according to a rigid generic hierarchy relation (SUBCONCEPT_OF relation) and
- the distinction between individuals and concepts, whereby an individual is related to a concept by means of an INSTANCE_OF relation.

Individuals in our data model are representations of concrete persons, geographical places, organizations, institutions, events etc. For example, 'Immanuel Kant', 'EU', 'Gran Canaria', 'Olympic Games 2004 in Athens' are all denotations of individuals.

2.1. Lemma-Term-Concept: roles of words in the language game

An ontology offers a formal method to structure sets of individuals with a set of individuals being an extension of a concept. Concepts are related to other concepts by means of a rigid hierarchy relation. This supports the factorizing of common information (see section 2.2.1) to more abstract levels.

Our idea is to represent the words of a language formally as *individuals*, called *lemmas* within our model. We consider morphosyntactic and word usage classes, e.g. information about the part-of-speech class of a word, its subcategorization frame, pragmatic usage, etc., formally as *concepts* and use them to group and classify the lemmas. This results in a further ontology, a kind of 'morphosyntactic ontology' about the 'world of words', which may be considered as a kind of further dimension of the first ontology described above, representing word senses and real world objects.

We bridge the two ontologies by using a denotation relation for connecting lemmas to one or more senses.

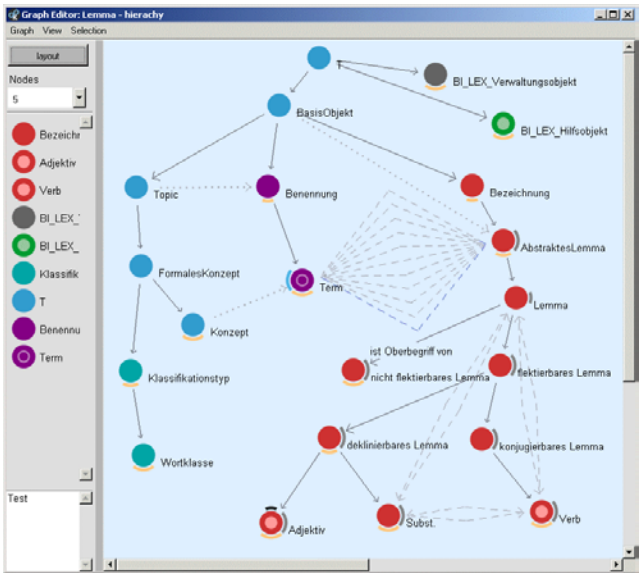


Figure 1: Top level of the Duden Ontology

Each sense of a lemma can be considered a role that this lemma plays in the language game, whereby each role played is represented by a single object, which we call *term*. In general, a lemma has more than one sense and thus a single lemma has more than one term assigned to it. Each sense of a lemma is represented by a single concept object.

On the other hand, a concept can be related to more than one term and thus to more than one lemma. This establishes the synonymy relation: Two lemmas are synonyms, if one of their corresponding terms points to the same concept.

We illustrate this in Figure 1: the top level concepts of the Duden Ontology are shown with the concept "Topic" being the root of the first ontology and "Bezeichnung" (denotation) being the root for the morphosyntactic ontology. The gap between the lemmas and concepts is bridged by means of a specific object class, i.e. "Term". All common information to all three object types is factorized at the "BasisObjekt".

2.2. Granularity gains

2.2.1. Factorizing of common information

One of our goals is to support the lexicographer in avoiding redundancy as this is one of the most important means for efficient maintenance, data consistency and multiple usage of the data. The means to avoid redundancy is the factorizing of common information: all information common to all objects should be stored in some more general object; when more general information is needed by the more specific object, this can be inherited (during runtime) from the more abstract one. Note that redundancy free storage does not hinder a redundant presentation of the data. The latter is not only useful for the lexicographer, but it is also advantageous for electronic products for which space restriction is not as rigid as in print products.

Obviously, it is not always possible to achieve a completely redundancy free data representation. Redundancy may, however, increase error-proneness in lexicography work. It has to be noted though that if redundant storage is required as a means for improving system performance, redundancy should be maintained by the system itself and be completely hidden from the user. The question of redundant storage is therefore "simply" a matter of the concrete implementation and not relevant to the model.

In our data model, the lemma is where the word-related information common to all its terms is factorized. A concept factorizes the meaning-related information common to all its synonymous lemmas. A term, though, may overwrite factorized information inherited by its corresponding lemma. In this way, we represent grammar and usage exceptions of particular lemmas, e.g. that a lemma in a particular sense may have no plural form.

2.2.2. Fine-grained relations

By representing terms as separate objects we gain granularity for the relations. In particular, we can link usage examples and citations for the dictionary entries to terms and not just to lemmas. By doing this, we disambiguate the meaning of the lemma in the usage example. Since we import data from the Duden dictionaries, the usage examples are already assigned to the particular meanings of a dictionary entry (for details see section 4.2.). With such information formally represented, one may get all usage examples of a concept simply by the union of all usage examples of all its terms.

In a similar manner, the representation of the decomposition of compound nouns on a term level and not only on a lemma level brings gains in granularity. This is advantageous when using such a resource for parsing or information retrieval tasks as the components of compounds are already disambiguated.

2.3. Concrete example

We explain the above model by means of an example from the Duden dictionary. The word "Bar" has three separate entries in the ten volume Duden dictionary (Duden, 1999):

[1]Bar, die; -, -s [engl. bar, urspr. = Schranke, die Gastraum u. Schankraum trennt < afrz. barre, Barre]: **1. a)** *intimes [Nacht]lokal, für das der erhöhte Schanktisch mit den dazugehörigen hohen Hockern charakteristisch ist:* eine B. besuchen, aufsuchen; in einer B. sitzen; **b)** *barähnliche Räumlichkeit in einem Hotel o. Ä.* **2.** *hoher Schanktisch mit Barhockern:* an der B. sitzen; Monsieur de Carrière lud mich ein, mich zu ihnen an die B. zu setzen (Ziegler, Labyrinth 258).

[2]Bar, das; -s, -s <aber: 3 Bar> [zu griech. báros = Schwere, Gewicht]: *Maßeinheit des [Luft]drucks;* Zeichen: bar (in der Met. nur: b).

[3]Bar, der; -[e]s, -e [H. u.]: *regelmäßig gebautes, mehrstrophiges Lied des Meistergesangs.*

There are three lemmas for "Bar" in the sense of (1) pub or bar, (2) measurement unit for (air) pressure and (3) a special form of song. The first entry, **[1]Bar,** has three senses (pub, hotel bar and counter) whereas **[2]Bar** and **[3]Bar** each have only one sense. Although all three lemmas are nouns, each lemma belongs to a different gender and declination class shown in the entry with the article and the genitive and plural form suffixes, e.g. "**[1]Bar**, die; -, -s" is feminine and forms the plural with a final 's'.

For each of the five senses there exists a separate term and a corresponding (separate) concept. Each sense definition, e.g. "*intimes [Nacht]lokal, ...*" for 1(a), is stored at the concept level. The usage examples and citations, e.g. „an der B. sitzen" (*English translation*: sitting at the bar) and „Monsieur de Carrière lud mich ein, mich zu ihnen an die B. zu setzen (Ziegler, Labyrinth 258)." (*English translation*: Monsieur de Carrière invited me, to join them at the bar (Ziegler, Labyrinth 258)), are connected to the term **[1]Bar (2)**.

Only the lemma, **[2]Bar** is synonymous to the lemma "**[2]bar**" as well as to the meteorological use of the sign "b". If we wish to extract all usage examples for say the concept "night bar" only those examples of the lemma "Bar" belonging to the term **[1]Bar (1a)** will be extracted. All other usage examples belong to terms, whose concepts are either hyponyms of the concept "night bar" or the concept "night bar" itself.

3. Tools and implementation

3.1. Ontology as a knowledge network

The data model is implemented with the *intelligent views* software system K-Infinity, which offers broad support for object-oriented knowledge modeling as well as for the creation, maintenance and use of a knowledge network. The software distinguishes between concepts and individuals and allows for the definition of relations and attributes both of which are inherited via the concept hierarchies.

The way we define ontology in our model fits well with the definition of a knowledge network in K-Infinity. The cornerstone of a knowledge network is a collection of concepts that structure information and allow the user to view it. The concepts are organized into hierarchies where each concept is related to its super- and subconcepts. This forms the basis for inheriting defined attributes and relations from more general to more specific concepts.

Concepts, individuals, attributes and relations are central to the construction of the knowledge network. A means for handling multiple inheritance are the so-called extensions or roles, the terms, which we use to represent the different senses of a lemma.

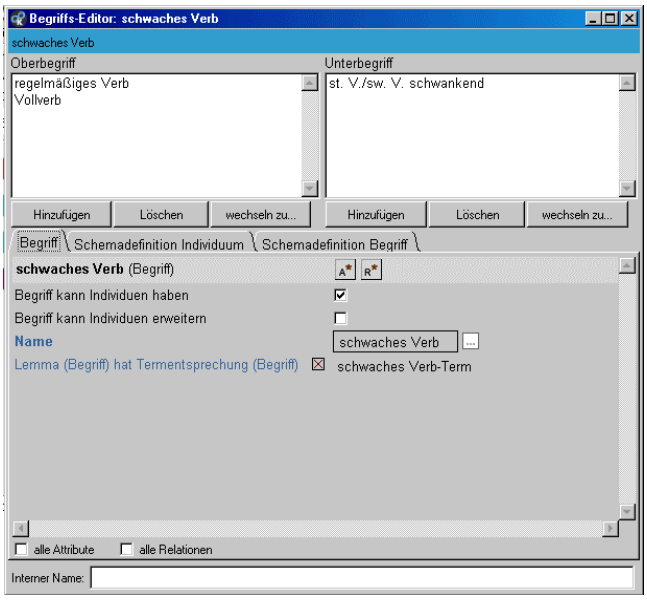


Figure 2: Concept Editor

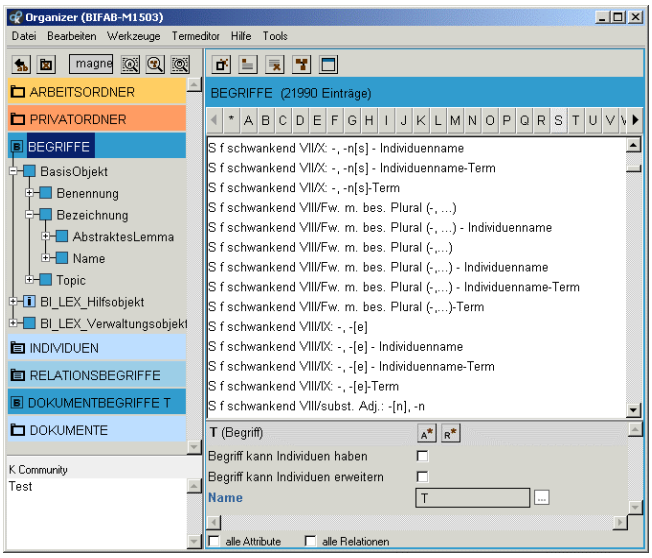


Figure 3: K-Organizer

3.2. K-Infinity Tools

The Knowledge Builder is K-Infinity's main component. It allows knowledge engineers and lexicographers to create, delete, rename and edit both objects and relations, as well as to relate objects to each other according to defined relations. This can be done in two different workspaces:

- The Graph Editor (shown in Figure 1) provides a graphical view of the network of objects and the relations between them. The network may be expanded according to the defined model. The Graph Editor supports the monitoring of the data by means of implemented consistency rules. One of the Editor's basic functions is an interactive network layout algorithm for the exploration of the knowledge network.
- The Concept Editor (see Figure 2) allows the user to focus on one object and its semantic links to neighboring objects. It is a supplement to the Graph Editor in that it allows the user to survey links and their attributes in detail, and to modify them if necessary.

Along with the tools for editing the knowledge network, there is the K-Organizer which supports administration, navigation, search and query formulation. The K-Organizer (Fig. 3) can be used to classify and group objects, either manually or by using existing object properties: for example, to organize all objects created before a certain date or all superconcepts with more than 10 subconcepts into a single folder.

Given the work context of the particular project, namely dictionary maintenance, an additional tool has been developed as a special extension for viewing and editing network objects from the perspective of a dictionary entry, called Term Editor. The Term Editor displays a lemma together with its associated terms and concepts in a single window in a comprehensive and compact way.

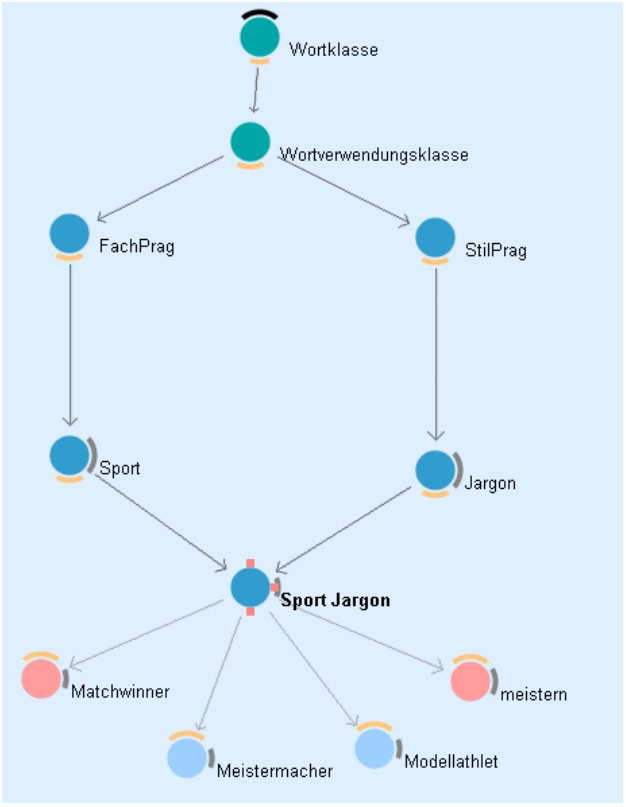


Figure 4: Example of pragmatic classes

3.3. Defined classes

There is a set of ca. 290 defined grammar classes, e.g. "noun which has a plural form", "masculine noun with declination type X", etc., ordered in a polyhierarchy. From these there are 160 classes which are assigned to lemmas; all the other classes are used to complement the poly-hierarchy as a means for flexible navigation and access.

Moreover, there are ca 1000 pragmatic classes, which are also ordered in a polyhierarchy, of which ca 250 are "basic pragmatic classes". The rest are combinations of pragmatic classes, such as for example, the class "Sport Jargon" shown in Fig. 4, which is a subclass of both "sport" and "jargon" classes. The class "jargon" is a subclass of "style" (StilPrag in Fig 4) whereas the super-class of "sport" is the pragmatic class "domain" (FachPrag). All in all there are at the moment over 200 relations defined in the model.

The defined grammar classes represent various aspects of the morphosyntactic nature of words. Starting from the general distinction of non-inflected and inflected word classes we divide the latter into conjugatable and declinable classes such as pronoun, article, adjective and noun and proceed to organize them extensively, which is necessary due to the rich morphology of German.

The noun hierarchy, shown in Figure 5, includes some abstract classes such as "noun by gender", "noun by type of declension", "noun with plural", "noun without plural", "noun derived from adjective", to classify the concrete noun classes such as the noun class the word "Aubergine" belongs to, namely, "feminine noun, declension type IX".

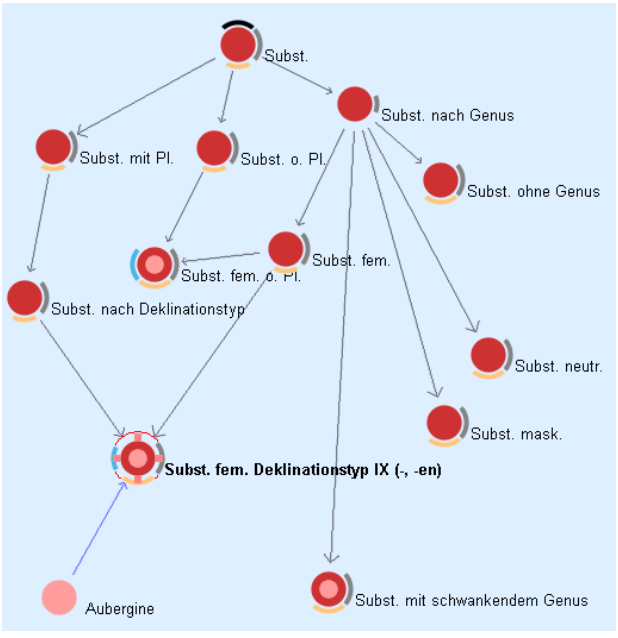


Figure 5: Example of noun classes

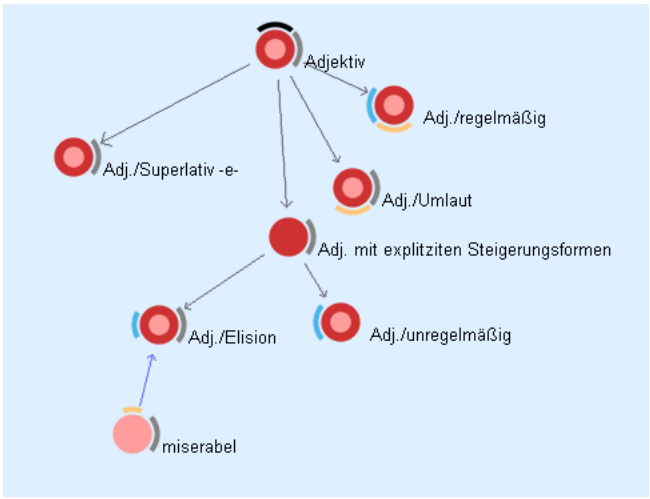


Figure 6: Example of adjective classes

As an additional example of the polyhierarchies consider the structure of the adjective classes (see Figure 6): In addition to the regular adjectives, we have defined subclasses for those with an explicit comparative form, with Umlaut and for those forming the superlative with "-e-". In the figure, the lemma "miserabel" is shown classified as an adjective belonging to the adjective subclass with an irregular comparative form, because of the elision of its -e-.

4. Import

To populate the Duden Ontology we first imported the data from the ten volume Duden dictionary (Duden, 1999), which contains ca. 200,000 lemmas, followed by the import of the entries of the Duden spelling dictionary (2000) with over 110,000 lemmas. Although there is a significant amount of overlap between the two dicitionaries, the former contains not only far more definitions than the latter, but also more grammatical,

etymological and pragmatic information. Importing and merging of further volumes are planned for the future.

The result of the complete import of the above data is a huge object network representing the information of over 200,000 entries from different dictionaries, whereby the entries themselves are decomposed into interlinked objects.

4.1. SGML dictionary data

As already mentioned, for each Duden dictionary, e.g. Duden (1999) or Duden (2000), there exists an SGML DTD. The basic structure of the dictionary articles is similar, however: Each dictionary article has a start and an end tag and each article element is divided into two parts, the head and the body. The head contains mainly information relevant to the lemma object of our data model and the body contains more detailed information concerning the senses of a lemma. The elements for phonetic, grammatical, etymological and pragmatic information are included in the head element. The body contains the substructure of the article and within this part there are elements containing definitions, examples, explanations, proverbs, idioms and idiomatic phrases. This straightforward structure is often interrupted by so called "meta-tags" which may appear anywhere within the above elements and contain some kind of text fragments. Naturally, this adds to the complexity of the import task.

There is, of course, no explicit tagging for terms and concepts, which is why a mapping from the existing mark up to the object types of our model is necessary. Because of the differences between the DTD(s) and our model it is not possible to write a simple context-free look-up table for mapping the DTD tags into the modeled object types. The content model of some elements is an iteration of a sequence of elements with optional parts, as shown in the example below for the element `defphr` (definition phrases):

```
<!ELEMENT defphr  - -
((ph?,gr?,prag?,(def|erk),erg?)?,bsp?,uew?,
rw?,spw?,iw?,(kurzf+  |  kurzw  |  abk+  |
zeich+)?)+ >
```

We map each iteration to a term, but since there is no explicit tag around this sequence of elements, the parsing process needs to exploit the contexts of the sequence in order to assign the information to the appropriate term.

4.2. Mapping

4.2.1. Creation of lemmas

Each dictionary entry is mapped to a lemma object. Typically, the homograph entries are indicated in the printed dictionary by a superscripted digit, which is also explicitly marked up as an attribute value in the article element. In this case we create different lemma objects with the same name, but with a different homograph-ID. The orthographic variants, e.g. "Photo" and "Foto", are marked up explicitly in the data. Separate lemma objects, which are related to the main lemma, are created for such variants.

Idioms and proverbs form specific lemma types which are automatically created during import.

4.2.2. Creation of terms and concepts

The different senses of an entry are structured in the dictionary by numbers or letters. We map each sense to a term and for each definition element we create an additional concept object. The usage and citation examples are assigned to the term object.

Grammatical or pragmatic information, which typically holds for the lemma, is modified in the sense description. Such modifications are stored in the corresponding term and overwrite the grammatical or pragmatic information inherited by the lemma.

The examples and definition phrases of the dictionary entries are often condensed for space reasons, e.g. the lemma appears in an abbreviated form. For instance, the entry for "Bar" in section 2.3 contains the phrase "an der B. sitzen" the complete form of which is "an der Bar sitzen". We expand such abbreviated forms during import and store the full form. Moreover – if necessary – we can generate the condensed form for export purposes.

4.2.3. Cross-references

During import we take care that no information necessary for the export of the data for the production of the dictionaries, such as the cross-references, is lost. The dictionary data contain explicit SGML elements for cross-referencing. We use the attribute values for the target article number and the subsection (the sense) in order to link the source and the target at the term level. We further check whether the subsection for the target lemma exists and whether the content of the cross-reference element can match the target lemma. In this way, we introduce an additional control for checking the correctness of cross-references, which is obviously advantageous for the quality of the constructed pool.

Due to the fact that the SGML data were originally created by an automatic conversion several thousands of the 80,000 cross-references solely refer to a subsection and have no reference to the article-ID. To resolve the missing cross-references we lemmatise the content of the cross reference elements and generate a list of target candidates, which is proofread by the lexicographers.

4.3. Enriching

Our aim is to populate the network with semantic relations, such as synonymy, hyperonymy, PART_OF or INSTANCE_OF relations. The SGML data contain no explicit mark up for such relations and a fully automated acquisition of semantic relations is not possible. We thus depend on maximal exploitation of our dictionary data in order to acquire semi-automatically semantic knowledge of this kind. For instance, the structure of the definition texts – which are stored at the concept level – is sometimes indicative for a synonymy relation holding between a given dictionary entry and its definition. As an example consider the dictionary entry "Yellow Press" in Duden (1999):

Yel||low Press ['jɛloʊ 'prɛs], die; - - (auch:) **Yel||low|press,** die; - [engl. yellow press, eigtl. = gelbe Presse] (Jargon): *Regenbogenpresse:* Längst ist die Witwe, von deren Auftritten einst die Y. P. profitierte, ruhiger geworden (FR 2. 1. 99, 9).

The word "Regenbogenpresse" (literary translation: "rainbow press") is marked up as definition text of the

term “Yellow Press”. We establish a synonymy relation between the two terms “Regenbogenpresse” and “Yellow Press” and their corresponding lemmas by assigning the same concept object to both terms.

We further plan to exploit the definition texts in combination with the cross-references to acquire hyperonymy and INSTANCE_OF relations.

A further method for extraction of hyperonyms is to automatically analyse compound words with the aim of extracting the heads of the compounds as these are in most cases the hyperonyms of the compounds.[1] For example, by analysing the compound “Volkstanz” (folk dance) we can infer that it is a hyponym of the word “Tanz” (dance).

For the representation of the morphological decomposition we define two relations and an attribute: *hat_Bestimmungswort* (has_modifier), *hat_Grundwort* (has_head) and the attribute *hat_Fuge* (has_join_morpheme). These relations are defined for both terms and lemmas. This is necessary since we cannot acquire all information we need in a single step. Rather we proceed iteratively to achieve a decomposition at the term level. In a first step all compound words of the dictionary are automatically morphologically analysed with the morphological analysis tool MPRO (Maas, 1996) to generate their components. As the decomposition of compounds is not always unambiguous, we disambiguate the analysis output by rejecting those compound analyses which have at least one component which is not a dictionary lemma. To illustrate this, there are two possible decompositions of the word “Medizinaldirektorin” (medical director) when automatically analysed:

medizinal – direktorin (medical – director)
medizin – aldi – rektorin (medicine – Aldi – rector)

The second analysis is nonsensical: Aldi is the name of a well-known German supermarket chain. The second analysis is thus rejected on the basis that there is no lemma for the the name Aldi. This strategy, however, does not always work, for example, consider the automatic analysis of the word “Marineuniform”:

marine – uniform (navy – uniform)
marine – uni – form (navy – university – form)

Again, the second decomposition is nonsensical, but in this case all three components are proper dictionary lemmas. The rule for selecting the correct decomposition is here a different one: the candidates for the right decomposition are those with the minimal number of components.

This way we fill in the lemma relations for the components of compounds[2]. If the lemmas which are components of a compound have only one sense, we have also achieved a decomposition at the term level. This is only possible, however, for a small number of compounds. Further investigation is required to determine a method to support the decomposition of compounds at the term level.

5. Conclusions and future work

In constructing the Duden Ontology our aim is not to build a general ontology of the world, but rather to create a computational resource which both supports efficient dictionary production and aids real world NLP applications. The creation of the Duden Ontology has been driven by our products and needs as well as by the abilities within the context of our work and the tools chosen.

This approach is guided by practical needs and has practical advantages for the lexicography work: by means of such an approach it is possible to maintain the dictionary data in a homogenous manner within a single data pool, something which was not previously possible for the Duden data.

With regard to the data model presented here, we believe that this kind of integrated model of semantic and grammatical information helps to avoid redundancy in storage and to maintain data without losing the ability to filter different sets of data and to generate various views of them with different granularity. The implementation of the data model is such that it allows modifications and further extensions, such as for example the definition of further semantic relations.

The next steps of our work concern the enrichment of the ontology with subcategorization information as well as with further semantic information. In particular, we plan to exploit the definition texts in combination with the cross-references to acquire hyperonymy and INSTANCE_OF relations.

For the future we plan to model further semantic relations to embed factual knowledge and encyclopedic information.

6. Acknowledgements

Special thanks go to Annette Klosa, Elke Siemon and Jan Schümmer for their valuable contribution to the project.

The work reported in this paper has been in part supported by the BMBF (German Ministry for Education and Research) grant 08C5885 for the research project „Lexikonbasierte Wissenserschließung: Natürlich-sprachige Suche und 3D-Wissensnavigation“.

7. References

Bläser. B. (1995). TransLexis: An Integrated Environment for Lexicon and Terminology Management. In P. Steffens (Ed.): *Machine Translation and the Lexicon, Third Internationals EAMT Workshop, Heidelberg, Germany, April 26-28, 1993, Proceedings*. Heidelberg: Springer Verlag, pp. 159-173.

[1] Note that ca 50% of the dictionary entries are compounds, which is attributable to the productivity of compounding in German.

[2] It is interesting to add that compound analysis at the lemma level is also important to determine the grammatical class for the compound word. Due to space reasons the single grammatical information coded for compound words in e.g. the ten volume Duden dictionary (1999) is gender. Whereas this is not problematic for a user of a dictionary, for automatic processing the missing information about the grammatical class of the compound is necessary. The grammatical class of the compound is determined by the class of the compound head.

Duden (1999). Duden - Das Große Wörterbuch der deutschen Sprache in 10 Bände. Mannheim: Dudenverlag, 3rd Edition.

Duden (2000). Band 1 – Die deutsche Rechtschreibung. Mannheim: Dudenverlag, 22nd Edition.

Duden (2001a). Duden Band 5 – Das Fremdwörterbuch. Mannheim: Dudenverlag, 7th Edition.

Duden (2001b). Duden Band 10 – Das Bedeutungswörterbuch. Mannheim: Dudenverlag, 2nd Edition.

Fellbaum, Ch. (Ed.) (1998). *Wordnet: An Electronic Lexical Database.* , Cambridge, MA: The MIT Press.

Hamp, B. & H. Feldweg (1997). GermaNet a Lexical-Semantic Net for German. In *Proceedings of the ACL/EACL-97 Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP applications*, Madrid, pp. 9-15.

Kunze, C. (2000). Extension and Use of GermaNet, a Lexical-Semantic Database. In *Proceedings of LREC 2000 Workshop on Lexicon: Semantic and Multilingual Issues* , Athens.

Mahesh, K. & S. Nirenburg, 1995. A situated ontology for practical NLP. *Proceedings of IJCAI '95 Workshop on Basic Ontologies Issues in Knowledge Sharing*. Montreal.

Maas, H-D. (1996). MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In Roland Hausser (Ed.): *Linguistische Verifikation, Dokumentation zur ersten Morpholympics.* Tübingen: Max Niemeyer Verlag.

Müller-Landmann, S. 2001. Wissen über Wörter. Die Mikrostruktur als DTD. Ein Beispiel. In H. Lobin (Ed.), *Proceedings der GLDV-Frühjahrstagung*, Universität Gießen, 2001, pp. 31-40.

Müller-Landmann, S. 2000. Design eines Internet-Lexikons zwischen Recherche und Rezeption. In U. Heid, S. Evert, E. Lehmann & C. Rohrer (Eds.): *Proceedings of the Ninth EURALEX International Congress,* Universität Stuttgart, Vol. I, pp. 97-105.

Rector, A.L., P.E. Zanstra, W.D. Solomon, J.E. Rogers, R. Baud, W. Ceusters, A.M.W. Claassen, J. Kirby, J. Rodrigues, A. Rossi Mori, E.J. Van der Haring & J. Wagner (1998). Reconciling user' needs and formal requirements: issues in developing a reusable ontology for medicine. *IEEE Transactions on Information Technology in Biomedicine, 2* (4), pp. 229-241.

Rogers, J., A. Roberts, D. Solomon, E. van der Haring, Ch. Wroe, P. Zanstra & A. Rector (2001). GALEN ten years on: Tasks and supporting tools. In Patel, V., R. Roger & R. Haux (Eds.) (2001): *MEDINFO 2001. Proceedings of the 10th World Congress on Medical Informatics*. Amsterdam: IOS, pp. 256-260.

Vossen, P. (Ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks* (reprinted from Computers and the Humanities, 32:2-3, 1998). Dordrecht: Kluwer Academic Publishers.

Relations among Roles

Anthony R. Davis*, Leslie Barrett†

* StreamSage, Inc.
1818 N St. NW, Washington, DC 20036, USA
davis@streamsage.com

†TransClick, Inc.
250 W. 57 St., New York, NY 10019, USA
leslie@transclick.com

Abstract

In this paper we discuss the possible types of relationships between participant roles in related situation types. We first discuss principles that might determine which roles are present in one type of situation, given the roles present in a related type of situation. While no simple general rules seem to exist, there are useful rules for particular cases. In addition, we discuss how relationships between roles themselves parallel relations between other elements in ontologies. Apart from the subrole relation, we consider relations analogous to meronymy and antonymy, which are rare in the domain of roles, and a complementarity relation between roles, which is fairly common.

1. Introduction

How are participant roles in one type of situation related to the roles in another? What implications do the relations between roles in different situation types have for the relations between other elements of ontologies? We will focus here on two topics that bear on these issues. First, we discuss which principles might determine, given the set of roles appropriate for one situation type, which subset of those roles are appropriate for a second, related situation type. Second, we examine the extent to which relationships between roles parallel those we find between other kinds of elements in ontologies.

2. Goals and assumptions

The way we have described the two topics above presupposes certain characteristics of an ontology (or related resource such as WordNet). We first briefly present these assumptions here, and pose the more detailed questions that we will address in this paper.

2.1. Participant roles, situation types, and hierarchies

We view participant roles for present purposes as relations between an entity and a situation.[1] Thus we will often refer to the entity as a participant in the situation.[2]. We also assume that for each role there can be type restrictions on the kinds of entities and situations that are appropriate arguments for that role. A *perceiver* role, for instance, is restricted to sentient entities in perception situations. This in turn rests on the assumption that situations and entities can be grouped into types, a strategy that has proven fruitful as a central organizing principle in many ontologies (e.g., the Cyc ontology, the SENSUS ontology, Mikrokosmos, the ontology developed in Sowa (2000), and numerous more specialized ontologies). Here we assume that types of entities and situations are hierarchically arranged, with multiple inheritance permitted (multiple inheritance is pervasive in the Cyc ontology but rare in WordNet).[3]

2.2. Subroles

In part because roles have type restrictions on the entities and situations that can serve as their arguments, it is reasonable to talk of subroles. One advantage of structuring roles in this way is that we can provide for arbitrarily specific roles for situation types anywhere in the situation-type hierarchy, while maintaining very general roles, which prove useful, for example, in stating the linguistic regularities in linking from semantic roles to syntactic arguments of predicators. The type restrictions on a subrole's arguments must be at least as restrictive as those on its super-roles. In addition, it is natural to assume that a participant playing a role in a situation also plays all of its super-roles:

(1) $R \subset R'$ implies $\forall x,e$: $R(x,e) \rightarrow R'(x,e)$

This entails a homomorphism under subsumption from the hierarchy of situation types to the hierarchy of roles, and from the hierarchy of entity types to the hierarchy of roles. The reverse implication—that all roles R and R' for which the condition on the right hand side of (1) holds are in the subrole-/super-role relation—is less obvious. This is an issue we briefly touch on below.

2.3. Role projectability between situation types

In section 3, we examine the problem of role projectability; that is, what principles and structures in an ontology determine, given the set of roles appropriate for one situation type, the set of roles of a related situation type. For example, are there any general statements we can make about the roles in subsituations, given the roles

[1] We use the term *situation* to speak of events and states. An event type is merely a situation type whose instances are events.

[2] We will not delve into the question of whether the entity actually must exist, or if it does, must temporally and spatially overlap the situation.

[3] Examples from the Cyc ontology are from the OpenCyc release of April, 2002, which can be examined or downloaded at www.opencyc.org. Version 1.7 of WordNet can be obtained from www.cogsci.princeton.edu/~wn/.

in a situation? This is particularly important in cases where it is debatable whether to analyze one situation type as a subtype of another or the second as a subsituation of the first (for instance, is *eating a meal* a subtype of *eating*, or is it better analyzed as containing a subevent of eating, along with other subevents such as serving oneself, and if the latter is the preferred analysis, how do the roles of the eating subevent project to the roles of eating a meal?). Another set of cases involves groups of similar situations, such as a group of *walking* events, for which we might wish to project some roles but not others. Finally, we also consider the interaction of role projectability and multiple inheritance.

We can classify projectability issues along three dimensions, as shown in the following table:

	situations	**entities**
type-level	sub/super-types of situations	sub/super-types of entities
individual level	sub-situations/ super-situations	sub-entities/ super-entities

Thus we can examine, for example, whether roles appropriate for a particular type of situation are appropriate for any subtypes of it, or we can examine whether a role played by a particular entity in a situation is also a role played by entities of which it is a part. Type-level projection is concerned with generalization and specialization relations between types (or hypernymy and hyponomy relations in lexical resources like WordNet), while individual-level projection is concerned with mereological (or meronymic) relations between individual situations and entities. One straightforward case is that of roles projecting from situation types to their subtypes, which is entailed by (1). We will not examine all of the possible options for projecting roles here (some of them are highly implausible in any case), but the table helps to situate the issues we examine in sections 3 and 4.

2.4. Parallelism in relations between roles and relations in other elements of ontologies

Often independently of concerns about the relations between concepts, scholars in linguistics and philosophy have been concerned with determining and classifying the roles of participants in situations. When situations and entities are arranged in type hierarchies, it is natural to inquire whether participant roles can be similarly arranged (see, among others, Parker-Rhodes (1978), Ostler (1979), Somers (1987), Lehmann (1997), and Sowa (2000)). In addition to subtype-supertype relations, however, we also find other types of relationships frequently modeled in ontologies. This leads to our second objective, which is to compare the structures of the participant role hierarchy to the other two. To what extent does the role hierarchy parallel the others, and which relationships commonly posited among situation and object types are applicable to roles as well? This will be discussed in section 4.

3. Some cases of role projectability

In this section we consider three cases of role projectability between situations and subsituations. The first concerns the case of a situation that can be regarded as composed of a group of situations of some common type. We suggest that roles of the group situation can be systematically related to those in the subsituations; the latter are subroles of the former. We next examine a more general case motivated by Lehmann's (1998) discussion of situations and roles, in which multiple inheritance in the situation-type hierarchy is pervasive. We argue that freely allowing this kind of multiple inheritance creates complications for the role system and should probably be constrained more than Lehmann envisions, or recast as a form of embedding the parent situation types as subsituations in another type rather than as multiple inheritance. Finally, we note the case of related telic and atelic situation types, which seems to require projection of roles from situations to subsituations, rather than inheritance from situation types to subtypes.

3.1. Groups of events of a common type

One frequent case of situations and subsituations is that of a group of situations of a given types, treated as a group, which itself can be regarded as a situation. This kind of operation is frequently represented in ontologies; Cyc's *GroupFn* is one example.[4] What can we conclude about the roles in the group situation, given the roles in its elements? One possibility is that they are identical. But this seems problematic. Suppose that the role *R* is defined for the situation type of the group's elements, and that in the group event *g*, the participant playing *R* is the mereological sum of all the participants playing role *R* in each of the elements of *g*. We will write this as *R(y,g)*, where *y* is the sum of the individuals playing this role in each of the elements. This is simply a case of the *cumulativity* or *summativity* property of roles (Krifka, 1992, 1998). While for some roles this is a reasonable move, it vitiates the definition of others. It may not cause any difficulties, for example, to regard a group of children running around a playground as the collective *agent* in a group *running* event. However, the *path* role in such an event is a discontinuous set of trajectories, while for a single child running, and for motion of a single body in a continuous time interval generally, the trajectory is continuous. This property of paths is important to maintain; Krifka's (1998) analysis of telicity in motion event relies on it, for example. Another example concerns *source* and *goal* roles in groups of motion events. It is useful to have a rule that either the *source* and *goal* of a motion event are distinct locations, or that the *moving object* has not changed its position (if the motion is a complete revolution in a circle, for instance). But this rule will not apply to groups of motion events; two runners might exchange places, each ending up in the other's starting location. The *source* and *goal* would then be identical in the group event.

A more palatable alternative is to assume that for every role *R* such that *R(x,e)* for some *x* in each element *e* of *g*, there is a super-role of *R*, *R'*, such that *R'(g,y)*, where *g* and *y* are the mereological sums, as above. These super-roles can have some of the properties of the original roles but need not have all of them. For example, the super-role of the *path* role could have discontinuous trajectories as its

[2] This represents some kinds of group situations adequately, but not all. Situations involving joint action or intent, for example, are not always readily decomposed into subsituations of a closely related type.

value, and those of the *source* and *goal* roles would have weaker distinctness conditions. At the same time, nothing precludes using the original roles to describe a situation that can be regarded both as a group of subsituations and as a single situation of the same type as those subsituations; in this case the same participant (a group of entities) will play the role *R*, and hence *R'*. A potential drawback to this approach is that, if we adopt the definition of subrole in (1), we are then committed to treating the type of groups of situations of type *S* as a supertype of *S* (the two types could be identical in some cases, such as at the top of the situation-type hierarchy). However, we see no obvious problems with this move, although this condition is not typically found in ontologies.

3.2. Multiple inheritance and roles

When a situation type is a child of more than one parent type, there are two possible outcomes with regard to the roles in the parent types. One is that two roles from two parent types merge, so that a single participant in an instance of the child type plays both of these roles. From two parent situation types such as *eating in a restaurant* and *eating breakfast* we can construct a type inheriting from both, *eating breakfast in a restaurant*, in which the eater and eaten roles of both parent types are merged; that is, there is a single eater participant and a single eaten participant in a situation of *eating breakfast in a restaurant*. In this case, the roles in the child type must be subroles of the roles in the parent types. This is represented graphically in Figure 1.

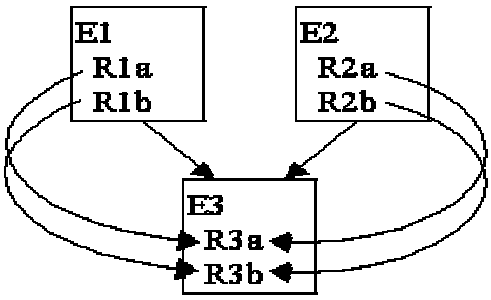


Figure 1. Merged roles in situation-type inheritance

A second possibility is that a role from one parent type does not merge with any other role, remaining distinct in the child type. In Cyc, for example, the type *CausingAnotherObjectsTranslationalMotion* is a subtype of *Movement-TranslationEvent*, which has the roles *objectMoving* and *trajectory*) and of *ActionOnObject*, with the roles *doneBy* and *objectActedOn*. The trajectory and *doneBy* roles remain distinct in the child type. As for the participant that is caused to move, it plays the roles of *objectActedOn* and *objectMoving* in an instance of *CausingAnotherObjectsTranslationalMotion*, but these two roles are not necessarily merged. There is no role reified within the Cyc system that inherits from these two roles. Instead, a rule states that the same participant plays both roles in situations of the type *CausingAnotherObjectsTranslationalMotion*. This second possibility is shown in Figure 2.

This has some implications for some of the ontological structures in Lehmann (1997). Lehmann exemplifies a situation-type hierarchy with increasingly complex types that inherit from multiple parents. For example, there are event types labeled "father gets harmed and angry child then gets revenge", a subtype of the situation types "father gets harmed" and "an angry child gets revenge".[5] Now if roles are inherited from types to their subtypes, this implies that the child type has all the roles of its two parent types. If this kind of type construction is fully productive in the situation-type hierarchy, however, it leads to the uncomfortable conclusion that roles always project from subevents to the events they are part of, since each conjunct can be considered a subevent. Consider the example of taking a trip in a car. We define event types of unlocking a car and driving a car in our hierarchy; the former type has a key as an instrument. Now we define the type of taking a trip in a car, inheriting from these two types of unlocking and then driving a car. By inheritance, this type also will have a key as an instrument, which is the undesirable situation we encountered above. This issue becomes particularly acute when there are two participants in the complex event type that are assigned the same role as a result of inheritance. Consider an event of taking dictation, where one person is reading aloud and another is copying down the words. The reader or writer in the parts of this event can both be considered agents, but we will certainly wish to distinguish these two roles in taking dictation. One solution here, of course, is to provide distinct, more specific roles, such as *reader* and *writer*. But this strategy is not always available; when two events of the same type are combined, the roles in the resulting type will be the same. As an example of this, picture a situation where two people compare versions of a text by having one read aloud and then the other, or a "call and response" situation where one person echoes another's words.

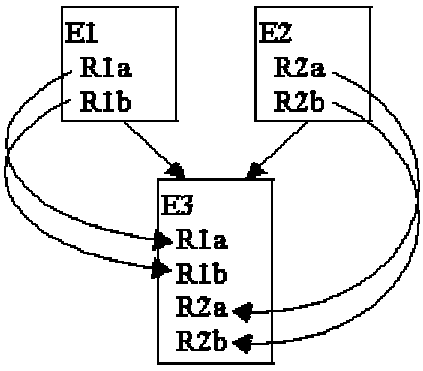


Figure 2. Distinct roles in situation-type inheritance

We could circumvent these problems in several ways. One is to postulate distinct roles for each situation type. Thus a key would play a role in the complex event type just mentioned, but it would not be the same role that it plays in the simpler subevent *unlocking a car*. This

[5] We disregard here the issue of how the temporal order of these two events in the subtype is specified. There must be some mechanism for doing so, however, since the reverse temporal order would describe a very different type of complex event.

allows us to be fully productive in creating complex situation types, at a cost of complicating our system of roles considerably. The number of roles is obviously indefinitely large, as the potential for creating successively more complex event types is unlimited, and there remains a problem of determining when two roles are necessarily filled by the same participant. For instance, a subtype of *unlocking a bicycle* is *unlocking a bicycle that is locked with a Kryptonite lock*. We can classify an instance of such an event in either fashion. This subtype has a distinct role for the key, but we want to equate the roles in the two event types, rather than worrying about whether there are two distinct keys. This representation, in which there are 4 roles, but only two participants, is shown in figure 3.

The large number of roles, and their uniqueness to individual situation types under this option, might become more palatable if we adopt a feature-based analysis of roles, along the lines of Somers (1987), Ostler (1979), Parker-Rhodes (1978), or Sowa (2000). From a linguistic standpoint, for instance, something like such features would be needed to account for regularities in the mapping from roles to syntactic arguments of verbs and nominalizations (see Dowty (1991) and Wechsler (1995) for similar accounts that can cast in a feature-based model). But in some sense we have merely shifted the problem from projectability of roles to projectability of features. If the features of a key as an instrument in unlocking a car are projected to its role in driving, why is it so odd to say that "we drove to the store with the key"?

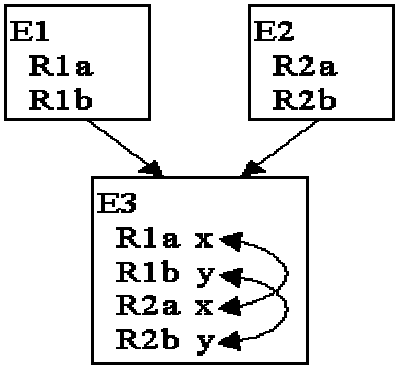


Figure 3. Distinct roles, shared by identical participants in subtype

Another option would be to structure the set of roles more richly, so that both sets of roles are inherited in a complex event, maintained as two separate structures (with additional roles potentially added as well). This option is in the spirit of feature structure representations, in which structures can be embedded recursively (Lehmann may allude to something similar when he refers to "structural specification"). A representation of this kind, in which the role-sets of the parent events are embedded within new role features in the child event, is shown in figure 4. The roles *R3a* and *R3b* within *E3* are filled by subevents; they might be relations such as *cause* and *effect*, for example. This allows roles to be inherited, albeit in a non-uniform way, which depends on how the parent situation types are combined in the child type. Furthermore, it is necessary to specify when a single participant fills roles in each part of the situation. In the type "father gets harmed and angry child then gets revenge on perpetrator", the same individual (the perpetrator) plays a role in both subevents.

Yet another approach would be to restrict the situation-type hierarchy to a set of types for which role inheritance makes sense. The trouble with this is that it seems too restrictive for many purposes. We sometimes do wish to refer to "composite" event types, like commuting to work on a bicycle, moving from one city to another, or holding a presidential election. But some kind of compromise position may be possible. We might maintain the kind of role inheritance that appears useful by designating one parent type as the "principal type", whose roles are inherited. For commuting by bicycle, the principal parent might be something like riding a bicycle, and the roles of the bicycle, the rider, the origin, and the destination would be inherited. Other, "minor" events involved in commuting, like locking and unlocking the bicycle, would not be involved in role inheritance. A subgraph of the hierarchy of situation types, filtered by "main event" or "principal type", links might be homomorphic to the role hierarchy. This approach seems reasonable for many of the situation types that we would be likely to reify in an ontology. It may apply less well to elaborate and complex events with many participants, such as political elections, which have many specialized roles, and would not necessarily inherit many of them from their parents representing their subevents. In some of these complex event types, the notion of a "main event" might not make much sense.

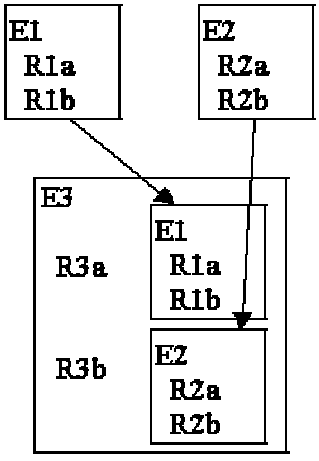


Figure 4. Embedded role-sets in subtype

From a linguistic standpoint, multiple inheritance in the situation-type hierarchy interacts with issues of linking; that is, the syntactic realization of predicates and their arguments. Subject selection is one good example; verbs denoting commercial transactions refer to situations in which there are two *agents*, as do causative verbs in the many languages that allow causativization of verbs denoting agentive situations. In each of these cases an accurate account of subject selection must appeal to more than the agentive status of a participant, since more than one participant plays an agentive role (see (Dowty 1991), (Wechsler 1995).and the Framenet system developed by

Filllmore and others for some approaches to this problem). Designating one of the subevents as the "main" or "salient" event for linguistic purposes, as in Framenet, accords well with the foregoing suggestion, although linguistic evidence is only a rough guide in these matters.

3.3. The inheritance of properties

One final issue regarding roles and subroles concerns how strictly we wish to enforce inheritance of properties. The OntoClean proposals of Guarino and Welty (2002), for example, place high importance on transmitting various properties dependably in inheritance. For instance, they argue against a pervasive characteristic of Cyc, that individual object types commonly inherit from the stuff types of which the objects are composed (thus *Ocean* is a subtype of *Water* in Cyc). When we examine the comparable situation in the realm of situations, we are led to the conclusion that telic situation types, such as eating an apple or painting a wall, should not be regarded as subtypes of atelic types such as eating or painting. The latter types are cumulative: two eating events may be combined and treated as a single eating event, but two events of eating an apple cannot be regarded as a single larger event of eating an apple.[6] This suggests that, parallel to the object and stuff types, telic and atelic event types should not be in a subtype-supertype relationship. If so, then telic event types will not inherit the roles of corresponding atelic event types. Instead, we could adopt a projectability rule that states: if *e* is an event of telic event type *T*, and *T* is "composed" of events of atelic type *A* (just as oceans are composed of water), then *e* also has those roles. In some cases, there may be no roles specified for events of type *T*, independently of type *A*. In others, such as many telic movement event types, additional roles are present, including *source* and *goal* roles.

In this case, then, we are led to a conclusion that is roughly the reverse of what we advocated in the case of groups of situations. For group situations, a consideration of roles for the group and for the subsituations comprising it led us to suggest that the group situation type is a supertype of the type of the elements. For the case of atelic and telic situation types, which might initially appear to be in a supertype-subtype relation, a re-examination of this assumption leads us to posit projection of roles from (atelic) subsituations to (telic) situations.

In sum, we see that there are unlikely to be simple general principles regulating the projection of roles between situations and their sub- or super-situations, although there do appear to be some useful, more specific principles covering some cases of interest.

4. Parallelism between roles and other elements in ontologies

In this section we explore what parallelisms may exist between the hierarchy of participant roles and other types of ontologies. Besides supertype-subtype relations, mereological relations are crucial in ontologies and in lexical resources like WordNet. Lexical resources also frequently employ an antonymy relation between words, though it is less clear that this is coherent ontological relation and ontologies emphasize this much less. In this section we will investigate to what extent these other relations can be applied to roles. In doing so, we will continue to mention issues of role projectability, this time with respect to entities and their parts.

4.1. Specialization/generalization (hyponymy/hypernymy)

Concept specialization is represented in WordNet with hyponymy links, and in Cyc with the predicate *genls* (and some extensions of it for relations). These apply both to entity types (or nouns in WordNet) and situation types (or verbs in WordNet, which then refers to this relation as "troponymy"). The comparable relationship for roles is simply the subrole relation; if one role is a subrole of another, then any participant that plays the first role in a situation necessarily also plays the second. This is the chief organizing relation for the hierarchy of roles, as it is for object and situation types.

However, we would like to remark here on one more linguistically relevant issue, since much of this same machinery is brought to bear on computational lexicons, including WordNet. Because the mapping from semantic roles to syntactic arguments is not completely semantically determined and displays some arbitrary variation, we cannot assume that hyponyms of a verb will exhibit the same mapping as that of the verb itself. In some cases, for example, an argument is incorporated in the verb (e.g., "spread butter on the bread" vs. "butter the bread", "put the money in the pocket" vs. the "pocket the money"). In others, the mapping is simply different (e.g., "eat oysters" vs. "dine/gorge on oysters"). This means that syntactic patterns are not necessarily reliable indicators of participant roles, and although hyponymy usually does imply inheritance of participant roles, corresponding roles may not occupy corresponding syntactic positions.

4.2. Partial roles (meronymy/holonymy)

Meronymy/holonymy, the lexical part/whole relation, and other mereological relations in ontologies, appear to be more complex, with several discernable subtypes. For example, Winston, Chaffin, and Hermann (1988) differentiate seven types of meronym: component-object (branch/tree), member-collection (tree/forest), portion-mass (slice/cake), stuff-object (aluminum/airplane), phase-process (adolescence/growing up), feature-activity (paying/shopping), and place-area (Baltimore/Maryland). Iris, Litowitz, and Evens (1988) acknowledge only four, however: functional part (wheel/bicycle), segment (slice/loaf), member (sheep/flock) and subset (meat/food), which is really specialization rather than meronymy. Likewise Cyc distinguishes numerous part/whole relations, including ingredients, physical and abstract parts, and subevents. The WordNet hierarchies employ just a single meronym link type, used only in the noun hierarchy. Meronymy applies just as usefully, however, to situation types (or verbs in WordNet), as we have been assuming throughout this paper. The type of meronymy called "phase-process" by Winston, Chaffin and Hermann

[6] Note that one and the same event can be regarded as both atelic or telic; eating an apple is certainly also eating. The telicity distinction is at the situation-type level, not at the individual level.

(1988) relates pairs of nouns and gerunds such as *adolescence/growing_up*. Feature-activity meronymy relates pairs of gerunds such as *paying/buying* or *steering/driving*. In short, events can be said to have component parts just as objects have them. The analogy to meronymy in the domain of participant roles is much less obvious than the specialization parallel, however.

We can begin by offering a definitions of "partial roles", as a mereololgical counterpart to the definition of subroles in (1):

(2) *R'* is a partial role of *R* iff:
R(x,e) → ∃*x,e*: *x'* is a part of *x* and *e'* is a subsituation of *e* and *R'(x',e')*

Unlike physical part and subsituation relationships, which are ubiquitous and obviously crucial to ontologies, there are relatively few instances of roles in this relationship that we are aware of, beyond the trivial case where *R* = *R'*, *x* = *x'*. and *e* = *e'*. Two cases are exemplified in the following sentences, where the participant denoted by the object of 'with' or 'by' is a part of another participant. Thus the "instrument" role is a partial role of *agent* in a. and the "body part" role is a partial role of the *grabbed* participant (or *theme*, or *affected object*) in b:

(3) a. I bumped the vase with my elbow.
b. I grabbed the iguana by the tail.

A third case of partial roles involves the *moving object* in movement events. In such events the parts of the object also move during at least some subintervals of the event, so the role *moving object* is partial to itself in a non-trivial way. In a parallel fashion, some roles in states are non-trivially self-partial. If someone owns a car for a year, that person owns the engine for the first six months, and if a beam supports a roof for a year, it is plausible to infer that a section of the beam supports a section of the roof for any period within that year.

Despite these cases, it appears that this type of part/whole relationship between roles is rare, and not particularly useful in inference. Possibly this is due to the relational character of roles, mediating between situations and their participants. We will now consider a more widespread phenomenon, the projection of role from participant entities to larger entities of which those participants are parts.

4.3. Projection of roles from entities to super-entities

We now examine the question of which roles can be projected from parts to wholes and vice versa; that is, if an object plays a role in a situation do larger objects of which it is a part and smaller objects that are parts of it also play that role in the situation? It should be clear that when this is the case, the role in question violates Krifka's uniqueness of objects property (Krifka 1992, 1998). Two kinds of roles for which this does seem to be true are roles of *source* and *goal* in motion events. For example, the following inferences seem valid:

(4) I flew from Baltimore to Boston. therefore,
I flew from Maryland to Boston. and
I flew from Baltimore to Massachusetts.

This inference has limits, in that the super-region cannot include both the origin and the destination of the trip, however, so the following are aberrant:

(5) #I flew (from the U.S.) (to the U.S.) and,
#I flew from the U.S. to Boston. and,
#I flew from Baltimore to the U.S.

The *path* role, in contrast, can be projected down to parts of the trajectory, but not to larger paths:

(6) Kim hiked (all of) the John Muir Trail. therefore,
Kim hiked the Tahoe-Yosemite Trail.

As Krifka (1992, 1998) has pointed out, we can make similar inferences from parts to wholes in the case of roles that involve contact or perception, as the following examples illustrate:

(7) John touched the door handle. therefore,
John touched the door.

(8) Kim rammed Sandy's bumper. therefore,
Kim rammed Sandy's car.

(9) The jar contacts the countertop. therefore,
The jar contacts the counter.

Note also that in situations involving both motion and contact, the contact inference is allowed even if the motion is not:

(10) I shook a link of the chain. therefore,
I touched the chain (even if I didn't shake it).

As for roles involving perception, the same pattern seems to apply, though the inference seems less solid:

(11) Fred saw the elephant's trunk. therefore,
Fred saw the elephant.

(12) Alice smelled the roasted chicken. therefore,
Alice smelled the meal.

As the story about the blind men and the elephant suggests, however, there is some uneasiness about such inferences. Perception differs from contact in this respect.

Finally, there are situation types in which one participant stands in a relationship of superiority to another, denoted by verbs such as 'exceed', 'surpass', 'dwarf', and verbs prefixed with 'out-'. In these cases, it arises virtually a matter of definition that the superior participant's role projects to objects of which it is a part, and the inferior one's role to its parts. This is exemplified in the following sentences:

(13) Nitrous oxide levels exceeded the Federal standards. therefore,
Smog levels exceeded the Federal standards.

(14) Bach outlived Vivaldi. therefore,
The Bach family outlived Vivaldi.

(15) Russia dwarfs Korea. therefore,
Russia dwarfs North Korea.

There are many roles for which projection to parts or wholes does not follow, except in some metaphorical or metonymic sense, including most roles involving agency, motion, and affectedness. In sum, "spatial" roles (including those that are appropriate for situation types whose linguistic realization is metaphorically based on spatial relationships) exhibit some projection properties that should prove useful in inference. But there is no direct parallel among roles to part/whole relationships of the type that apply ubiquitously to entities and situations.

4.4. Antonymy/opposition

Another relation in WordNet, more explicitly lexical, is antonymy, although as Miller (1998) points out, it is not a fundamental an organizing relation between nouns. True antonymy is present in the verb hierarchy, as well as among adjectives. Change-of-state verbs, for example, have antonymous counterparts quite similar to nouns, although the verb pairs don't normally share parents (e.g., 'lengthen'/'shorten' and 'strengthen'/'weaken'). Relations of opposition occur as well, where there is no common superordinate or entailed verb unique to the pair (e.g., 'give'/'take', 'buy'/'sell').

Antonymy is closely tied to lexical properties and not a coherent ontological relationship, but some aspects of it can be singled out and represented as conceptual relationships. For example, reversative actions (*zipping* and *unzipping*, *loading* and *unloading*, *arriving* and *leaving*, *creating* and *destroying*) exemplify a fairly coherent notion of opposition that bears on participant roles. We cannot say that the event types in each pair have the same roles; for example, *loading* and *arriving* both have a *goal* role, but may lack a *source*, while *unloading* and *leaving* are the opposite. But it is probably fair to say that each role of an event type has a counterpart in the corresponding reversative event type. The same may hold true for other sorts of opposites (e.g., *helping* and *hindering*, *benefiting* and *suffering*, *believing* and *doubting*), though in many of these cases we are more likely to say that the role's counterpart is itself. It seems less meaningful to posit a counterpart relationship between roles in some other types of situations sometimes thought of as "opposites" (being awake or asleep, liking and disliking, and many others), let alone antonyms in the domain of properties and objects.

4.5. Complementary roles

Another relation between two roles that seems worthwhile is what we term *complementarity*. For some situation types, we know that when one role is present, another role must be also. We then say that this second role is complementary to the first. Complementarity may unidirectional or bi-directional, but most of our examples will involve the latter case. Some examples of such roles are *buyer* and *seller*, *buyer* and *payment*, *moving object* and *path*, *driver* and *vehicle*, and *perceiver* and *perceived*. One application of a complementarity relation in inference should be fairly clear; it allows us to postulate the existence of a participant filling a role when the participant playing the complement role is known to be present. However, this sort of inference is probably equally simply performed with reference to situation types, as long as they specify which roles are necessary and appropriate. The complementarity relation bears some resemblance to meronymy and to the "partial role" relations; it could even be considered a type of partial role relation applied to situations, disregarding the requirement in (2) that participants playing each role be in a part-whole relationship. Complementarity certainly has counterparts in the entity and situation domains. The existence of a hole depends on the existence of a cavity wall, and the two transfer subevents of a commercial transaction seem complementary in much the same way that the roles are.

Roles that are complementary and that, in a given situation type, are entailed to be filled by the same participant, may violate the reverse of the implication in (1). That is, if *R'* is a complementary role of *R*, and a situation type is constrained so that *R(x,e)* and *R'(x,e)* for any situation *e* of that type, then a bi-directional interpretation of (1) would treat *R* as a subrole of *R'*.[7] There may be legitimate grounds, however, to distinguish two participant roles in such situations. For example, someone who is talked into performing an action is both a *addressee* and a *performer* (of the action). It is possible in such cases to create a role specific to that type inheriting from the two roles *R* and *R'*, but it does not always seem desirable to do so. We leave this question open.

5. Conclusions

We have seen that an examination of the relations among roles can be fruitful in illuminating other aspects of ontologies and lexical resources. Considering the question of role projectability has shown that permitting multiple inheritance to operate without constraint in the situation-type hierarchy is problematic, and that other mechanisms do not cause the same difficulties for inheritance of roles from situation types to their subtypes. We have also seen how role inheritance interacts with two particular cases of situations and subsituations: a group of like situations and telic situations composed of atelic subsituations. In these two cases, role projectability reveals interesting relationships among situation types in ontologies.

Roles parallel situation and entity types in constituting a hierarchy, but, perhaps because of their inherently relational nature, the parallelism beyond that is limited. While we can formulate coherent definitions of relationships between roles that parallel the mereological relations that are so pervasive among situations and entities, their usefulness is less apparent. In contrast, the complementarity relation between roles is widespread and its utility in inference clear.[8]

6. References

Chaffin, R., Hermann, D.J. and Winston, M.E., 1988. An empirical taxonomy of part-whole relations: Effects of the part-whole relation type on relation identification. *Language and Cognitive* Proceses, 31, pp. 17-48.

[7] If the two roles are complementary to each other, then each would be a subrole of the other, hence they would be the same role.

[8] We would like to thanks the referees for comments which have improved this paper.

Dowty, D., 1989. On the Semantic Content of the Notion 'Thematic Role'. In G. Chierchia, B. Partee, and R. Turner (eds.), *Properties, Types, and Meaning*. Dordrecht: Reidel.

Dowty, D., 1991. Thematic Proto-Roles and Argument Selection. *Language* 67:3, pp. 547-619.

Fellbaum, C. (ed.) 1998. *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press

Guarino, N., and Welty, C., 2002. Evaluating Ontological Decisions With OntoClean. *Communications of the ACM (CACM)*, 45:2, pp. 61-65.

Iris, M.A., Litowitz, B.E. and Evans, M., 1988. Problems of the part-whole relation. In M. Evans (ed.) *Relational models of the lexicon*, Cambridge: Cambridge University Press.

Krifka, M., 1992. Thematic Relations as Links between Nominal Reference and Temporal Constitution. Iin I. Sag and A. Szabolcsi (eds.), *Lexical Matters*. Stanford, CA: CSLI Publications.

Krifka, M., 1998. The Origins of Telicity. In S. Rothstein (ed.), *Events and Grammar*. Dordrecht: Kluwer Academic Publishers.

Ladusaw, W., and Dowty, D., 1988. Toward a Nongrammatical Account of Thematic Roles, In W. Wilkins (ed.), *Syntax and Semantics 21: Thematic Relations*. San Diego: Academic Press.

Lehmann, F., 1997. Big Posets of Participatings and Thematic Roles. In P. Eklund, G. Ellis, and G. Mann (eds.), *Conceptual Structures: Knowledge Representation as Interlingua*. Heidelberg: Springer.

Miller, G.A., 1998. Nouns in Wordnet. In C. Fellbuam (ed.), *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ostler, N., 1979. *Case Linking: A Theory of Case and Verb Diathesis Applied to Classical Sanskrit*. Ph.D. dissertation, MIT.

Parker-Rhodes, A., 1978. *Inferential Semantics*. Atlantic Highlands, NJ: Harvester Press.

Pustejovsky, J., 1995. *The Generative Lexicon: A theory of computational semantics.* Cambridge, MA: MIT Press.

Somers, H., 1987. *Valency and Case in Computational Linguistics*. Edinburgh: Edinburgh University Press.

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

Wechsler, S., 1995. *The Semantic Basis of Argument Structure*. Stanford, CA: CSLI Publications.

Restructuring WordNet's Top-Level: The *OntoClean* approach

Alessandro Oltramari[(1)], Aldo Gangemi[(2)], Nicola Guarino[(1)], Claudio Masolo[(1)]
[(1)]LADSEB-CNR*, Padova, Italy: {Nicola.Guarino, Alessandro.Oltramari, Claudio.Masolo}@ladseb.pd.cnr.it
[(2)]ISTC-CNR, Rome, Italy: gangemi@ip.rm.cnr.it

Abstract

In this paper we propose an analysis and a rearrangement of *WordNet's* top-level taxonomy of nouns. We briefly review WordNet and identify its main semantic limitations, in the light of the ontology evaluation principles lying at the core of the *OntoClean* methodology. Then we briefly present a first version of the *OntoClean Top* (OCT) ontology, and show how WordNet can be aligned with it. The result is a "cleaned-up" WordNet, which is meant to be conceptually more rigorous, cognitively transparent, and efficiently exploitable in several applications.

1 Introduction

The number of applications where WordNet is being used more as an ontology than just as a lexical resource seems to be growing more and more. To be used as an ontology, however, some of WordNet's lexical links need to be interpreted according to some formal semantics, which tells us something about "the world" and not (just) about the language. One of such links is the hyponym/hypernym relation, which corresponds in many cases to the usual subsumption (or IS-A) relation between concepts. An early attempt at exploring the semantic and ontological problems lying behind this correspondence is described in (Guarino, N., 1998).

In the recent years, we developed a methodology for testing the ontological adequacy of taxonomic links called OntoClean (Guarino, N. & Welty, C., 2002; Guarino, N. & Welty, C., 2002), which was used as a tool for a first systematic analysis of WordNet's upper level taxonomy of nouns (Gangemi, A. *et al.*, 2001). The first version of OntoClean was based on an ontology of properties (unary *universals*), characterized by means of meta-properties. We are now extending OntoClean with an ontology of *particulars* called OCT (OntoClean Top ontology), which is presented here in some detail, although still in an informal way. The OCT will be the first module of a minimal library of *foundational ontology* that we shall develop within the *WonderWeb*[1] project.

This paper is structured as follows. In the next section we present an extension of our FOIS paper (Gangemi, A. *et al.*, 2001), concerning some ontological inadequacies of WordNet's taxonomy of nouns. Then we introduce the most recent version of our OntoClean Top ontology, and discuss the preliminary results of an alignment work aimed at improving WordNet's overall ontological (and cognitive) adequacy, and facilitate its effective deployment in practical applications.

2 WordNet's Preliminary Analysis

2.1 Experiment Setting

We applied our methodological principles and techniques to the noun synsets taxonomy of WordNet 1.6.To perform our investigation, we had to adopt some preliminary assumptions in order to convert WordNet's databases[2] into a workable knowledge base. At the beginning, we assumed that the hyponymy relation could be simply mapped onto the subsumption relation, and that the synset notion could be mapped into the notion of concept. Both subsumption and concept have the usual description logics semantics (Woods, W. A. & Schmolze, J. G., 1992). In order to work with named concepts, we normalized the way synsets are referred to lexemes in WordNet, thus obtaining one distinct name for each synset: if a synset had a unique noun phrase, this was used as concept name; if that noun phrase was polysemous, the concept name was numbered (e.g. window_1). If a synset had more than one synonymous noun phrase, the concept name linked them together with a dummy character (e.g. Equine$Equid).

Firstly, we created a Loom[3] knowledge base, containing, for each named concept, its direct super-concept(s), some annotations describing the quasi-synonyms, the gloss and the synset topic partition, and its original numeric identifier in WordNet; for example

```
(defconcept Horse$Equus_Caballus
    :is-primitive Equine$Equid
    :annotations ((topic animals)
    (WORD |horse|)
    (WORD |Equus caballus|)
    (DOCUMENTATION "solid-hoofed herbivorous quadruped domesticated since prehistoric times"))
    :identifier |101875414|)
```

noun entries	116364
equivalence classes: synonyms, spelling variants, quasi-synonyms	50337
noun synsets (with a gloss and an identifier for each one)	66027
nouns	95135
monosemous nouns	82568
polysemous nouns	12567
one-word nouns	70108
noun phrases	25027

Table1: Elements processed in the Loom WordNet kb

The elements processed in the Loom WordNet knowledge

*In the process of moving to ISTC-CNR, Rome, Italy.

[1] http://wonderweb.semanticweb.org/

[2] We used the Prolog WordNet database, the Grind database, and some others from the official distribution.

[3] Loom is a knowledge representation system that implements a quite expressive description logic (MacGregor, R. M., 1991).

base are reported in Table 1. We report in Figure 2 an overview of WordNet's noun top-level as translated in our Loom knowledge base. The nine Unique Beginners are shown in boldface.[4]

2.2 Main problems found

Once the Loom WordNet was created, we systematically applied the OntoClean methodology to the upper taxonomy of noun senses. Let us discuss now the main ontological drawbacks we found after applying this cleaning process.

2.2.1 Confusion between concepts and individuals

The first critical point was the confusion between concepts and individuals. For instance, if we look at the hyponyms of the Unique Beginner Event, we'll find the synset Fall - an individual - whose gloss is “the lapse of mankind into sinfulness because of the sin of Adam and Eve”, together with conceptual hyponyms such as Social_Event, and Miracle.[5] Under Territorial_Dominion we find Macao and Palestine together with Trust_Territory. The latter synset, defined as "a dependent country, administered by a country under the supervision of United Nations", denotes a general kind of country, rather than a specific country as those preceding it. If we go deeper in the taxonomy, we find many other examples of this sort. For instance, the hyponyms of Composer are a mixture of concepts and instances: there are classes corresponding to different special fields, such as Contrapuntist, or Songwriter, and examples of famous musicians of the past, such as Bach, and Beethoven.

Under Martial_Art, whose top hypernym is Act, we find Karate, and Kung Fu, but these synsets do not stand for concepts, they represent individuals, namely particular examples of martial arts.

If we look through Organization, under the branch whose root is Group, we find conceptual hyponyms such as Company, Alliance, Federation, Committee, together with instances like Irish_Republican_Army, Red Cross, Tammany Society[6], and so on.

We face here a general problem: the concept/individual confusion is nothing but the product of an “expressivity lack”. In fact, if there was an INSTANCE-OF relation, we could distinguish between a concept-to-concept relation (subsumption) and an individual-to-concept one (instantiation).

2.2.2 Confusion between object-level and meta-level: the case of Abstraction

The synset Abstraction_1 seems to include both object-level concepts, such as Set, Time, and Space, and meta-level concepts such as Attribute and Relation. From the corresponding gloss, an abstraction “is a general concept formed by extracting common features from specific examples”. An abstraction seems therefore intended as a psychological process of generalization, in accordance to Locke's position ((Lowe, E. J., 1998), p.211). This meaning seems to fit the latter group of terms (Attribute, Relation, and possibly some hyponyms of Quantity), but not to the former. Moreover, it is quite natural to consider attributes and relations as meta-level concepts, while set, time, and space, seem to belong to the object domain.

2.2.3 OntoClean constraints violations

A core aspect of OntoClean is the analysis of subsumption constraints induced by the identity, rigidity, and unity meta-properties. In our analysis, we only found rigidity violations. We suspect that there are two reasons why we didn’t observe other kinds of violation: on one hand, we limited our analysis to the upper levels, where the criteria of identity and unity are very general; on the other hand, WordNet tends, notoriously, to multiply senses, so the chances of conflict are relatively limited.

The most common violation we have registered is bound to the distinction between roles and types. A role cannot subsume a type. Let's see an important clarifying example.

In its first sense, Person (which we consider as a type) is subsumed by two different concepts, Organism and Causal_Agent. Organism can be conceived as a type, while Causal_Agent as a formal role. The first subsumption relationship is correct, while the second one shows a rigidity violation. We propose therefore to drop it.

Someone could argue that every person is necessarily a causal agent, since ‘agentivity’ (capability of performing actions) is an essential property of human beings. Causal_Agent should therefore be intended as a synonym of ‘intentional agent’, and considered as rigid. But, in this case, it would have only hyponyms denoting things that are (essentially) causal agents, including animals, spiritual beings, the personified Fate, and so on. Unfortunately, this is not what happens in WordNet: Agent, one of Causal_Agent hyponyms, is defined as: "an active and efficient cause; capable of producing a certain effect; (the research uncovered new disease agents)". Causal_Agent subsumes roles such as Germicide, Vasoconstrictor, Antifungal. Instances of these concepts are not causal agents essentially. This means that considering Causal_Agent as rigid would introduce further inconsistencies.

These considerations allow us to add a pragmatic guideline to our methodology: when deciding about the formal meta-property to attach to a certain concept, it is useful to look at all its children.

2.2.4 Heterogeneous levels of generality

Going down the lower layers of WordNet's top level, we register a certain ‘heterogeneity’ in their intuitive level of generality. For example, among the hyponyms of Entity there are types such as Physical_Object, and roles such as Subject. The latter is defined as “something (a person or object or scene) selected by an artist or photographer for graphic representation”, and has no hyponyms (indeed, almost any entity can be an instance of Subject, but none is necessarily a subject)[7].

For Animal (subsumed by Life_Form) this heterogeneity becomes clearer. Together with classes such as Chordate, Larva, Fictional_Animal, etc., we find out more specific concepts, such as Work_Animal, Domestic_Animal,

[4] Note that the sense numeration reported in our Loom kb is different from the WordNet's original one. Nevertheless, the reader will easily recognize the synsets we are referring to.

[5] In the text body, we usually do not report all the synonyms of a synset (or their numeration), but only the most meaningful ones.

[6] “A political organization in New York city (late 1800’s early 1900’s) seeking political control by corruption and bossism”.

[7] We can draw similar observations for relation_1 and set_5 with respect to abstraction_1, etc.

Mate_3, Captive, Prey, etc. We are induced to consider the formers as types, while the latters as roles.
Although problematic on the side of ontological distinctions among event-classes, the hyponyms of Phenomenon_1 represent another meaningful example of heterogeneity. At the same taxonomic level there are "reasonably" general synsets like Natural_Phenomenon and Process together with a specific concept like Consequence, which could be modeled as anti-rigid (every event can be a consequence of the occurring of a previous event, but we could assume that this is not the essential characteristic of the event itself[8]).
In short, intuitively some synsets sound too specific when compared to their siblings. Look at them from the formal point of view we are developing, we can pinpoint their "different generality" by means of the distinction between types and roles.

3 The OntoClean Top Ontology

Before presenting our (still preliminary!) OCT ontology, a couple of clarifications may be useful. First of all, we do *not* intend this as a candidate for a "universal" standard ontology. Rather, we support the vision of a *library* of *foundational ontologies*, reflecting different commitments and purposes. In our opinion, the most important challenge today is not so much the agreement on a monolithic set of ontological categories, but rather the careful isolation of the fundamental ontological options and their formal relationships. If general ontologies reflecting different commitments and purposes are described in terms of these formal notions, then we can hope they will form a library of "foundational" ontologies accessible in a modular way, keeping the necessity of largely shared ontological commitments to the very minimum, and making the rationales and alternatives underlying the different ontological choices as explicit as possible. This is one of the goals of the *WonderWeb* project, where the OCT ontology will be linked to other foundational ontologies.
A second clarification concerns the general attitude underlying our ontological choices. The OCT ontology has a clear *cognitive bias*, in the sense that we aim at capturing the ontological categories lying behind natural language and human commonsense. Hence, we do not claim that our categories have "deep" metaphysical implications related to the intimate nature of the world: rather, they are thought of as "conceptual containers" useful to describe ontologies as cognitive artifacts ultimately depending on human perception, cultural imprints and social conventions. So, especially with respect to natural language, our attitude is more "descriptive" than "revisionary" (Strawson, P. F., 1959; Loux, M. J., 1998).
Finally, we have to point out that the ontology presented here is an ontology of *particulars*. Properties and relations are therefore not part of its domain. Some proposals for a ontology of properties have been made in (Guarino, N. & Welty, C., 2000). We are not aware of any systematic work on the ontology of relations.

[8] For instance, the extinction of dinosaurs could have be the consequence of the impact of an asteroid on the Earth, or of a sudden glaciation, or of a mortal epidemic – scientists are not sure about this – but in terms of ontology of events, it is a conclusive event, at most an annihilation event, and there is no need (and here no possibility) to model it as a consequence.

3.1 General notions

Before introducing the OCT categories, let us first introduce the general notions we shall use to characterize them. Some of these notions (like rigidity and unity) have already been defined in previous papers (respectively, (Guarino, N. & Welty, C., 2002) and (Gangemi, A. *et al.*, 2001)), and will not be discussed here. So we shall limit ourselves to the basic distinction between *enduring and perduring* entities, and the varieties of dependence relationships involving particulars.[9] We shall keep the discussion to an informal, introductory level; a rich axiomatization will be presented in a forthcoming paper.

3.1.1 Enduring and perduring entities

A fundamental distinction we assume in the OCT ontology is that between *enduring* and *perduring* entities. This is almost identical, as we shall see, to the distinction between so-called *continuants* and *occurrents* (Simons, P., 1987), which is still being strongly debated both in the philosophical literature (Varzi, A., 2000) and within ontology standardization initiatives[10]. Again, we must stress that this distinction is motivated by our cognitive bias: we do not commit to the fact that both these kinds of entity "really exist", and we are indeed sympathetic with the recent proposal made by Peter Simons, that enduring entities can be seen as equivalence classes of perduring entities, as the result of some kind of abstraction mechanism (Simons, P., 2000).
But let us see what this distinction is about. The difference between enduring and perduring entities (which we shall also call *endurants* and *perdurants*) is related to their behavior in time. Endurants are always *wholly* present (i.e., all their proper parts are present) at any time they are present. Perdurants, on the other hand, just extend in time by accumulating different temporal parts, so that, at any time they are present, they are only *partially* present, in the sense that some of their proper parts (e.g., their previous phases) may be not present. For instance, the piece of paper you are reading now is wholly present, while some temporal parts of your reading are not present any more. Philosophers say that endurants are entities that *are in time*, while lacking however temporal parts (so to speak, all their parts travel with them in time). Perdurants, on the other hand, are entities that *happen in time*, and can have temporal parts (all their parts are fixed in time).
This different behavior affects the notion of change in time. Endurants can "genuinely" change in time, in the sense that the very same whole endurant can have incompatible properties at different times; perdurants cannot change in this sense, since none of their parts keeps its identity in time. To see this, suppose that an endurant has a property at a time *t*, and a different, incompatible property at time *t':* in both cases we refer to the whole object, without picking up any particular part. On the other hand, when we say that a perdurant has a property at *t*, and an incompatible property at *t'*, there are always two different parts exhibiting the two properties.
We have already mentioned that endurants and perdurants can be taken as synonyms of the more common terms

[9] In the OntoClean taxonomy evaluation methodology only dependence between properties is used.
[10] See for instance the extensive debate about the "3D" vs. the "4D" approach at www.suo.org.

continuants and *occurrents*. We prefer however the adopted terminology, because the continuants/occurrents distinction is sometimes considered only within so-called *concrete* entities, while, as we shall see, we take it as spanning the whole domain of particulars, including abstracts that we shall consider as endurants. Finally, we shall take *occurrence*, and not *occurrent*, as synonym of *perdurant*, since it seems natural to use *occurrent* to denote a type (a *universal*), whose instances are occurrences (*particulars*).

The endurants/perdurants distinction evidences the general necessity of temporally indexing the relationships within endurants. This means that, in general, it is necessary to know *when* a specific endurant bears a certain relation to other endurants. Consider for instance the classical example of Tibbles the cat (Simons, P., 1987): Tail is part of Tibbles before the cut but not after it, i.e. we have to "temporalize" the part relation: *P*(Tail, Tibbles, *before*(cut)) and ¬*P*(Tail, Tibbles, *after*(cut)).

With respect to a temporalized relation *R*, we can distinguish *R-constant* endurants from *R-variable* endurants. An endurant *e* is called *R-constant* iff, when $R(x_1, \dots, x_n, e, t)$ holds for a temporal interval *t*, then $R(x_1, \dots, x_n, e, t')$ also holds whenever *e* is present at *t'*.

We can also strengthen this definition introducing the modal notion of an *R-invariant* endurant. An endurant *e* is called *R-invariant* iff, if it is possible that $R(x_1, \dots, x_n, e, t)$ then necessarily $R(x_1, \dots, x_n, e, t)$ holds whenever *e* is present at *t'*.

For the purpose of characterizing the OCT categories, the property of being constant (or invariant) with respect to the parthood relation (*mereologically constant* (*invariant*)) has a special relevance. For example, we usually take ordinary material objects as mereologically variable, because during their life they can lose or gain parts. On the other hand, amounts of matter are taken as mereologically invariant (all their parts are *essential part*), and so on.

3.1.2 Dependence

Let us now introduce informally some useful definitions based on the notion of dependence, adapted from (Thomasson, A. L., 1999). We focus here on *ontological dependence* (holding primarily between particulars, and only by extension between properties), to be distinguished from *notional dependence*, which only holds between properties).

A particular *x* is *specifically constantly dependent* (SCD) on another particular *y* iff, at any time *t*, *x* can't be present at *t* unless *y* is also present at *t*. For example, a person might be specifically constantly dependent on its brain.

A particular *x* is *generically constantly dependent* (GCD) on a property ϕ iff, at any time *t*, *x* can't be present at *t*, unless a certain instance *y* of ϕ is also present at *t*. For example, a person might be generically constantly dependent on having a heart.

1.2 The OntoClean Top Categories

The most general kinds of particulars assumed in the OntoClean Top ontology are described in Figure 1. They are assumed to be mutually disjoint, and covering the whole domain of particulars. They are also considered as *rigid* properties, according to the OntoClean methodology that stresses the importance of focusing on these properties first.

Quality
Quality Region
Aggregate
- Amount of matter
- Arbitrary collection

Object
- Physical Object
  - Body
  - Ordinary object
- Mental Object

Feature
- Relevant part
- Place

Occurrence
- State
- Process
- Accomplishment

Abstract

Figure 1: Onto Clean Top Categories.

1.2.1 Qualities and quality regions

'Quality' is often used as a synonymous of 'property', but this is not the case in the OCT ontology: qualities are particulars, properties are universals. According to our view, every entity comes with certain qualities, which exist exactly as long as the entity exists. These qualities belong to different *quality types* (like color, size, smell, etc.), and are characteristic (*inhere to*) specific individuals: no two particulars can have the same quality. So we distinguish between a quality (e.g., the color of a specific rose), and its "value" (e.g., a particular shade of red). The latter is called *quale*, and describes the "extension" (or "classification") of an individual quality with respect to a certain *conceptual space* (called here *quality space*) (Gärdenfors, P., 2000), So when we say that two roses have the same color their two colors are classified in the same way wrt the color space (they have the same *color quale*), but still they have two numerically distinct qualities.

The reason of this distinction between qualities and qualia, which is inspired to the theory of tropes (with some differences that can't be discussed here[11]), is mainly due to the fact that natural language – in certain constructs – seems often to make a similar distinction. For instance, when we say "the color of the rose turned from red to brown in one week" or "the room's temperature is increasing" we are not speaking of a certain shade of red, or a specific thermodynamic status, but of something else that changes its properties in time while keeping its identity. This is why we assume that qualities are endurants.

On the other hand, when we say that "red is opposite to green" or "red is close to brown" we are not speaking of qualities, but rather of regions within quality spaces. The specific shade of red of our rose – its color quale – is therefore an atom in the color space.[12]

[11] An important difference is that standard tropes theories explain a qualitative change in terms of a substitution of tropes (an old trope disappears and a new one is created). We assume instead that qualities are a sort of "enduring tropes".

[12] The possibility of talking of qualia as particulars rather than reified properties is another advantage of our approach.

Each quality type has an associated quality space with a specific structure. For example, lengths are usually associated to a metric linear space, and colors to a topological 2D space. The structure of these spaces reflects our perceptual and cognitive bias.
Under this approach, we can explain the relation existing between 'red' intended as an adjective (as in "this rose is red") and 'red' intended as a noun (as in "red is a color"): the rose is red because its color is located in the red region within the color space (more exactly, its color quale is a part of that region).
As a final remark, we note that qualities are assumed to be as specifically constantly dependent on the entities they *inhere to*.

1.2.1.1 Location

In the OCT ontology, space and time are considered as quality types like color, weight, etc. The spatial (temporal) individual quality of an entity is called *spatial* (*temporal*) *location*, while its quale is called *spatial* (*temporal*) *region* and it belongs to the associated quality space (respectively geometric space and temporal space). For example, the spatial location of a physical object is just one of its individual qualities: it belongs to the quality type *space*, and its quale is a region in the geometric space. Similarly for the temporal location of an occurence. This allows an homogeneous approach that remains neutral about the properties of the geometric/temporal space adopted (for instance, one may assume a circular time).
Notice that quality regions can have qualities themselves (for instance, the spatial location of a certain object can have a shape), in particular we assume that all quality regions are temporally located, and that their temporal qualia coincide with the temporal universe, i.e. quality regions are always present.

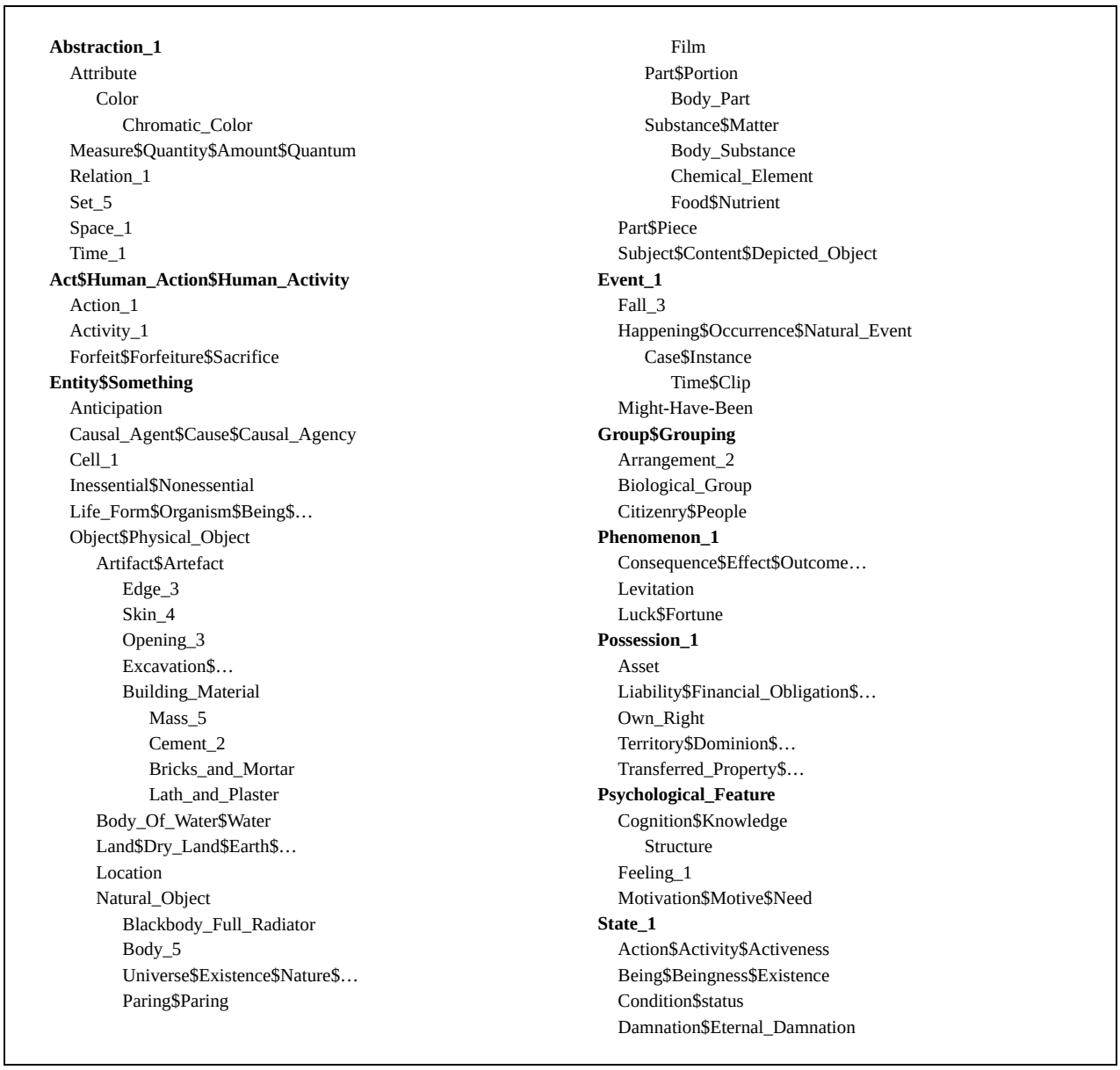


Figure 2: WordNet's top Level

1.2.2 Aggregates

The common trait of aggregates is that they are endurants and none of them is an essential whole. We consider two kinds of aggregates: *Amounts of matter* and *Arbitrary collections*. The former are mereologically invariant, in the sense that they change their identity when they change some parts. The latter are defined as "mere mereological sums" of essential wholes which are not themselves essential wholes (like the sum of a person's nose and a computer keyboard). They are essentially mereologically *pseudo-constant*, in the sense that they change their identity when a member (i.e. a special part of a collection, see (Gangemi, A. *et al.*, 2001)) is changed, while a change in the non essential parts of a member is allowed. We may have called arbitrary collections *groups*, or perhaps *sets*; but we prefer to use *set* for abstract entities, and *group* for something having an intrinsic unity.

1.2.3 Objects

The main characteristic of objects is that all of them are endurants and essential wholes. They have no *common* unity criterion, however, as different subtypes of objects may have different unity criteria. Often objects (indeed, all endurants) are considered ontologically independent from occurrences (discussed below). But, if we admit that every object has a life, it is hard to exclude a mutual ontological dependence between the two. Nevertheless, we can use the notion of dependence to distinguish between objects that are not specifically constantly dependent on other objects and have a spatial location (*physical objects*) and objects that are generically constantly dependent on *persons* (that are also objects) and do not have a spatial location (*mental objects*). Among physical objects, we further distinguish between *bodies* and *ordinary objects*. Bodies are mereologically invariant, and then they are material objects in the sense of physics.[13]. Ordinary objects (and mental objects even more) have a more cognitive nature, as they are admitted to change some of their parts while keeping their identity: they can have therefore *temporary parts*. Among mental objects, we could distinguish between purely *subjective mental objects*, i.e. objects depending on a singular person (like an intention, or a competence), and *intersubjective mental objects*, i.e. objects depending on a community of persons (like a project, a legal norm, a moral value, an aesthetic notion).

1.2.4 Features

Typical examples of features are "parasitic entities" such as holes, bumps, surfaces, or stains, which are generically constantly dependent on physical objects[14] (their *hosts*). All features are essential wholes, but no common unity criterion may exist for all of them. However, typical features have a topological unity, as they are *singular* entities. Features may be *relevant parts* of their host, like a bump or an edge, or *places* like a hole in a piece of cheese, the underneath of a table, the front of a house, which are not parts of their host.

[13] Notice that differently from the amounts of matter they are essential whole.

[14] We may think that features are specifically constantly dependent on their host, but an example like "a whirlpool" is very critical in this sense. Notice that we are not considering as features entities that are dependent on mental-objects.

1.2.5 Occurrences

Occurrences are synonymous of perdurants. They comprise what are variously called events, processes, happenings, and states. Occurrences can have temporal parts or spatial parts. For instance, the first movement of (the execution of) a symphony is a temporal part of it. On the other side, the play performed by the left side of the orchestra is a spatial part. In both cases, these parts are occurrences themselves. Clearly objects can't be parts of occurrences, rather they *participate* to them.

Within occurrences, we consider two main ontological dimensions of distinction: homeomery and relationality. The first dimension has been introduced by Parsons, Cresswell, and Mourelatos (see (Casati, R. & Varzi, A., 1996)): intuitively, we say that an occurrence is homeomeric iff all its temporal parts can be described *in the same way* used for the whole occurrence: for instance, every temporal part of "my sitting here" for an hour is still a "sitting here of mine". But if we consider "Messner's ascent to Everest" (intended in the complete sense), no parts of it are a "Messner's ascent to Everest". To formalize this notion, we need to refer to a certain property that holds for all the temporal parts of a certain occurrence *o*. We individuate this property by considering the most specific *occurrent* of *o*, i.e. the most specific occurrence type *o* is instance of. Then we can say that *o* is homeomeric iff all its temporal parts are instances of the same most specific occurrent.

The second dimension takes inspiration mainly from (Smith, B., 1982). An occurrence is said *non-relational* when only one object participates to it, while it is *relational* when it has two or more objects as participants. Occurrences involving qualities varying in time (i.e., which can change their qualia in time) are prototypical examples of non-relational occurences: the change of color of a rose has only one object as a participant (there may be other participants, such as the rose's color, but this is a quality and not an object).

In our proposal, homeomery seems to be enough to account for the distinctions proposed in the literature (especially (Mourelatos, A., 1996)) among *states*, *processes*, and *accomplishments*. It is easy to see that states are homeomeric occurrences (e.g., "the air smelling of jasmine"), while *accomplishments* are non-homeomeric (e.g. "the sunset"). Processes can be characterized as *weakly non-homeomeric*, in the sense that *some* temporal parts of them are instances of the same most specific *occurrent*, and some are not. For instance, in the case of "running", if you consider that instantaneous temporal part of your running through the park in which your right foot touches the ground while your left foot does not (think about photo-finish in a race), this sub-event is no more a "running". Together, processes and accomplishments are often described as *dynamic events*, just because of an (apparent) change of some of their properties across their different temporal parts.

In any case, we can further divide each of these categories into relational and non-relational occurrences.

1.2.6 Abstracts

Like mental-object and their qualities, abstracts are enduring entities that do not have a spatial location (indeed they do not have any "physical quality"). Differently from mental-object and their qualities, abstracts are independent from objects (and in particular from persons). Exam-

ples of abstracts are *sets, symbols, propositions, structures*, and *physical laws.*

4 Mapping WordNet into the OCT ontology

Let us consider now the results of integrating the WordNet top concepts into our top-level. According to the OntoClean methodology, we have concentrated first on the so-called *backbone taxonomy*, which only includes the rigid properties. Formal and material roles have been therefore excluded from this preliminary work.

Comparing WordNet's unique beginners with our ontological categories, it becomes evident that some notions are very heterogeneous: for example, Entity looks like a "catch-all" class containing concepts hardly classifiable elsewhere, like Anticipation, Imaginary_Place, Inessential, etc. Such synsets have only a few children and these have been already excluded in our analysis.

The results of our integration work are sketched in Table 2. Our categories are reported in the first column; the second column shows the WordNet synsets that are *covered* by such categories (i.e., they are either equivalent to or included by them); the third column shows some hyponyms of these synsets that were rejected according to our methodology. Finally, the last column shows further hyponyms that have been appended under our categories, coming from different places in WordNet. The problems encountered for each category are discussed below.

4.1 Aggregates, Objects, and Features

Entity is a very confused synset. As sketched in the table, a lot of its hyponyms have to be "rejected": in fact there are roles (Causal_Agent, Subject_4), unclear synsets (Location[15]) and so on. This Unique Beginner maps partly to our Aggregate and partly to our Object category. Some hyponyms of Physical_Object are mapped to our new top concept Feature.

By removing roles like Arrangement and Straggle, Group$grouping becomes a partition of the Ordinary Object category. In fact, hyponyms like Collection, Social_Group, Biological_Group, and so on, are nothing but plural objects, supporting a clear unity criterion.

Possession_1 is a role, and it includes both roles and types. In our opinion, the synsets marked as types (Asset, Liability, etc.) should be moved towards lower levels of the ontology, since their meanings seem to deal more with a specific domain - the economic one - than with a set of general concepts (except some concepts that can be mapped to Mental Object, such as Own_Right). This means that the remainder branch is also to be eliminated from the top level, because of its overall anti-rigidity (the peculiarity of roles).

4.2 Abstracts and Qualities

ABSTRACTION_1 is the most heterogeneous Unique Beginner: it contains abstracts such as Set_5, mental objects such as Chromatic_Color (an example of quality space[16]), qualities (mostly from the synset Attribute) and a hybrid concept (Relation_1) that contains mental objects, concrete entities (as Substance_4[17]), and even meta-level categories (see §2.2.2). Each child synset has been mapped appropriately.

Psychological_feature contains both mental objects (Cognition[18]) and events (Feeling_1). We consider Motivation as a material role, so to be added to lower levels of the taxonomy of mental objects.

The classification of qualities deals mainly with adjectives. This paper focuses on the WordNet database of nouns; nevertheless our treatment of qualities foreshadows a semantic organization of the database of adjectives too, which is a current desiderata in the WordNet community (see (Fellbaum, C., 1998), p. 66).

4.3 Occurences

Event_1, Phenomenon_1, State_1 and Act_1 are the Unique Beginners of those branches of WordNet denoting events. WordNet does not support the distinction between relational and non-relational occurrences, so first of all, in order to restructure this partition of the top level, we need to separate the hyponyms of the above-mentioned four synsets by means of our defined first dimension. We see, for example, that State_1 maps in part to non-relational state (condition$status, cognitive_state, existence, death_4, degree, skillfulness…), in part to relational state (medium_4, relationship_1 and relationship_2, disorder, order, hostility, conflict…). We register a similar behavior for the children of Process (a subclass of Phenomenon_1): decrement_2, increment and shaping could be seen as kinds of process involving a single main participant, while chelation, economic_process, execution and some hyponyms of Natural_Process (a direct hyponym of Process) seem to denote relational occurrences. Under Act_1 we find in general events of two kinds: processes (see activity_1 and its hyponyms) and accomplishments (see the homonymous synset under action_1). For sake of simplicity, we consider the hyponyms of Act_1 as being both relational and non-relational, depending on the context in which they are used. Event_1 has a too much generic composition in order to be partitioned clearly in terms of our approach (see, for instance, the beginning of §2.2.1): to a great extent, however, its hyponyms could be added to lower levels of the taxonomy of occurrences.

5 Conclusions

The final results of our integration effort are sketched in Figure 3. Our results show that a serious taxonomy rearrangement is needed. The blind application of OntoClean's taxonomy evaluation methodology provides a first guideline, but stronger ontological commitments seem to be unavoidable in order to get a "disciplined" taxonomy. In our opinion, strong (and explicit) ontological distinctions do also reduce the risk of classification mistakes in the ontology development process, and simplify the update and maintenance process.

Our research is still in progress: we hope we have paved

[15] Referring to Location, we find roles (There, Here, Home, Base, Whereabouts), instances (Earth), and geometric concepts like Line, Point, etc.).

[16] By looking to the corresponding hyponyms, it becomes clear that this synset could also be viewed as denoting a quality (by means of this we decide to append it both under Quality and Quality Region top concepts).

[17] "The stuff of which an object consists".

[18] "The psychological result of perception, and learning and reasoning".

the way for future work and possible cooperation.

6 Acknowledgements

We would like to thank Stefano Borgo and Luc Schneider for the fruitful discussions and comments on the earlier version of this paper. This work was jointly supported by the Eureka Project IKF (E!2235, Information and Knowledge Fusion), the IST Project 2001-33052 *WonderWeb* (Ontology Infrastructure for the Semantic Web) and the National project TICCA (Cognitive Technologies for Communication and Cooperation with Artificial Agents).

7 References

Casati, R. & Varzi, A. (eds.) (1996). Events. Aldershots, USA, Dartmouth.

Fellbaum, C. (ed.) (1998). WordNet - An Electronic Lexical Database. , MIT Press.

Gangemi, A. *et al.* (2001). Understanding top-level ontological distinctions: In Proceedings of IJCAI-01 Workshop on Ontologies and Information Sharing (26-33). Seattle, USA, AAAI Press.

Gangemi, A. *et al.* (2001). Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top-Level. In C. Welty & S. Barry (Eds.), Formal Ontology in Information Systems. Proceedings of FOIS2001 (285-296). , ACM Press.

Gärdenfors, P. (2000). Conceptual Spaces: the Geometry of Thought. Cambridge, Massachussetts, MIT Press.

Guarino, N. (1998). Some Ontological Principles for Designing Upper Level Lexical Resources. In A. Rubio *et al.* (Eds.), Proceedings of First International Conference on Language Resources and Evaluation (527-534). Granada, Spain, ELRA - European Language Resources Association.

Guarino, N. & Welty, C. (2000). A Formal Ontology of Properties. In R. Dieng & O. Corby (Eds.), Knowledge Engineering and Knowledge Management: Methods, Models and Tools. 12th International Conference, EKAW2000 (97-112). France, Springer Verlag.

Guarino, N. & Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. Communications of the ACM, 45(2), (61-65).

Guarino, N. & Welty, C. (2002). Identity and subsumption. In R. Green *et al.* (Eds.), The Semantics of Relationships: an Interdisciplinary Perspective . , Kluwer (in press).

Loux, M. J. (1998). Metaphysics, a Contemporary Introduction. , Routledge.

Lowe, E. J. (1998). The possibility of metaphysics. Oxford, Clarendon Press.

MacGregor, R. M. (1991). Using a Description Classifier to Enhance Deductive Inference: In Proceedings of Seventh IEEE Conference on AI Applications (141-147).

Mourelatos, A. (1996). Events, Processes, States. In R. Casati & A. Varzi (Eds.), Events (457-476). Aldershot, Dartmouth Publishing Company.

Simons, P. (1987). Parts: a Study in Ontology. Oxford, Clarendon Press.

Simons, P. (2000). How to Exist at a Time When You Have No Temporal Parts. The Monist, 83(3), (419-436).

Smith, B. (ed.) (1982). Parts and Moments: Studies in Logic and Formal Ontology. München, Philosophia Verlag.

Strawson, P. F. (1959). Individuals. An Essay in Descriptive Metaphysics. London and New York, Routledge.

Thomasson, A. L. (1999). Fiction and Metaphysics. Cambridge, Cambridge University Press.

Varzi, A (2000). Foreword to the special issue on temporal parts. The Monist, 83(3).

Woods, W. A. & Schmolze, J. G. (1992). The KL-ONE family. In F. W. Lehmann (Ed.) Semantic Networks in Artificial Intelligence (133-177). Oxford, Pergamon Press.

OCT Top Categories	**Covered Synsets**	**Rejected Hyponyms**	**Imported Hyponyms**
Quality	Attribute*	Trait, Ethos, Inheritance, …	
Temporal Location	Time_interval$interval*	Eternity, Greenwich_Mean_Time, Present, Past, Future	
Spatial Location	Position$place		
Color …	Chromatic_color		
Quality Region	Attribute*	Trait, Ethos, Inheritance, …	
Time Region	Time_1, Time_interval$interval*	Eternity, Greenwich_Mean_Time, Present, Past, Future	
Space Region	Space_1*	Subspace, …	
Color Region …	Chromatic_color		
Aggregate	Aggregate_2 (!)		
Amount of Matter	Substance$Matter*	Bedding_Material, Ballast, Atom, …	Mass_5, Cement_2, Substance, …
Arbitrary Collection			
Object	ENTITY$SOMETHING*	Anticipation, Causal_Agent, Imaginary_Place, Substance	
Physical Object			
Body	Natural_Object*	Dead_Body, Constellation, Stone, Nest, …	
Ordinary Object	Physical_Object*, Group*	Finding, Catch, Vagabond; Arrangement, Social_Group, …	
Mental Object	PSYCHOLOGICAL_FEATURE*	Feeling_1, Motivation_1	Own_Right (!), Social_Group
Feature			
Relevant Part	Part$portion*, Fragment	Substance_4	Edge_3, Skin_4, Paring$Parings, …
Place			Opening_3, Excavation$hole_in_the_Ground, …
Occurrence	STATE_1*, PHENOMENON_1*, ACT*	Utopia, Dystopia, Nature, Consequence, Stay_1, …	
State	STATE_1*	Utopia, Dystopia, Nature	
Non-relational	Condition$status, Cognitive$State, Existence, Death_4, Degree, …		
Relational	Medium_4, Relationship_1, Relationship_2, Order, Disorder, Hostility, Conflict, …		
Process	Process, Activity_1		
Non-relational	Decrement_2, Increment, Shaping		
Relational	Chelation, Execution, …		
Accomplishment	Accomplishment$achievement		
Non-relational			
Relational			
Abstract			Statement_1, Cognition, Arrangement_2,
Proposition	Proposition_1		
Set …	Set_5		

Table 2: Synsets marked with '*****' are heterogeneous (some of their children are to be moved elsewhere, some are roles, or some are instances); those marked with '(**!**)' have no hyponyms; those in upper case are WordNet Unique Beginners.

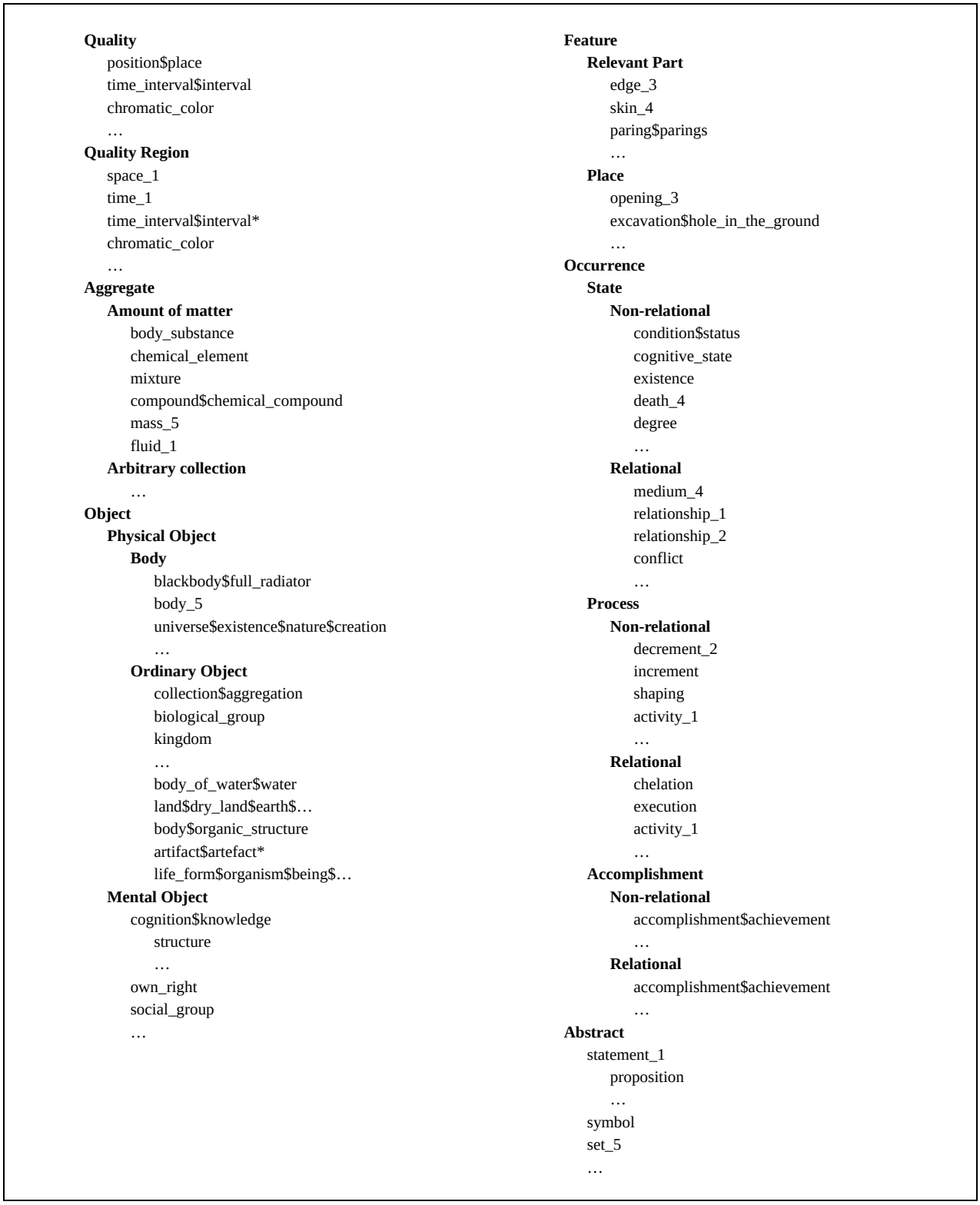


Figure 3: WordNet cleaned up: mapping WordNet into the OntoClean top-level.

Parallel Hierarchies in the Verb Lexicon

Christiane Fellbaum

Cognitive Science Laboratory, Department of Psychology
Princeton University, Princeton, NJ 08544, USA

Abstract

We discuss semantically heterogeneous manner-relations in the verb component of a lexical database. To make verb hierarchies more consistent while at the same time including instances of links among verbs that are based on expectancy instead of logical necessity, we propose to augment the lexical database with a parallel relation among hierarchically organized verbs. Possibilities for identifying instances of para-troponymy in corpora are outlined and the advantages of an enriched lexical database for NLP are briefly discussed.

1. Introduction and Background

It has been pointed out that the noun hierarchies in WordNet are built on heterogeneous subsumption relations (Gangemi et al., 2001; Gangemi et al., 2002; Guarino and Welty, 2001). The most common violation of the subsumption relation is the failure to distinguish Types and Roles (Guarino and Welty, 2002). Thus, WordNet lists as subordinates of the synset *dog, domestic_dog, Canis_familiaris* such synsets as *poodle, poodle_dog, Newfoundland*, and *corgi, Welsh_corgi* along with synsets like *cur, mongrel, mutt, lapdog, hunting_dog,* and *working_dog*. (Gangemi et al., 2001; Gangemi et al., 2002) propose eliminating from WordNet violations of strict subsumption (Type) relations and moving Roles like *student* to lower levels of the taxonomy.

Some of WordNet's verb hierarchies exhibit heterogeneous kinds of subordinates that seem intuitively similar to the Type/Role distinction among the nouns. For example, among the manner-subordinates of *clean*, we find *steam-clean* along with *brush, sweep*, and *wipe*. One of our goals here is to examine the heterogeneous manner-of relations in WordNet's verb component. Referring to work in progress, (Gangemi et al., 2002) briefly outline a clean ontology of events, categorizing them on the basis of criteria such as aspect and intentionality. Their examples are all complex events, such as *conducting a symphony* and *running a 100-meter race*. The number and nature of the event's participants as well as its spatial and temporal parts provide criteria for the ontological status of the events.

WordNet's verb entries are for the most part simple lexical items and do not include the kinds of complex events cited in (Gangemi et al., 2002). To the extent that WordNet is an ontology, it is a strictly lexical ontology whose entries are limited to concepts that are lexicalized in English[1]. WordNet resembles a traditional dictionary or thesaurus in that it does not explicitly account for aspectual or argument-taking properties of verbs (though verbs that are hierarchically related frequently share the same valency and aspectual properties). Therefore, the criteria for a clean ontology of events outlined by (Gangemi et al., 2002) are not applicable, and, indeed, may be complementary to the present discussion. Our treatment of simple verbs must necessarily be less ambitious, though we hope, no less interesting.

Besides offering some theroretical reflections, this paper attempts to outline how the different manner relations among the verbs could be constructively exploited and how corresponding links might be added to WordNet. Distinguishing and introducing a second manner relation parallel to the existing one would not only ensure semantically consistent relations but also yield a richer and more tighly interconnected network with a greater potential for NLP applications.

2. Hierarchies in WordNet's Verb Lexicon

In WordNet (Fellbaum, 1998), a word's meaning is represented by its membership in a group of cognitively synonymous words (a synset), and labelled pointers among the synsets that stand for semantic relations such as hyponymy, meronymy, and opposition.

The semantic relation that organizes most of the verbs in WordNet is the manner relation, or troponymy (Fellbaum, 1998). This relation allows one to build hierarchical structures akin to those found in the noun lexicon. Similar to the hyponymy relation expressable by the formula "X is a kind of Y", the formula for troponymically related verbs is (1):

(1) to X is to Y in some manner/way

For example, *stammer*, *lisp*, and *whisper* are among the many manner subordinates of *speak*, as the statement "to stammmer/lisp/whisper is to speak in some manner" shows.

Thus, WordNet expresses (part of) the meaning of verb X in terms of the meaning of its superordinate, Y. And the meaning of verb Y is expressed, in part, as the sum of the meaning of its subordinates (troponyms), such as X.

The manner relation is highly polysemous, as (Fellbaum, 1998) notes. Depending on the semantic domain, the differentiae distinguishing the superordinate from the more specific subordinate may be dimensions like speed (*walk-run*), direction (*move-rise*), volume (*talk-scream*), or intensity (*persuade-brainwash*). Despite these differences, the formula given in (1) seems to fit thousands of English verb senses and could be used to construct WordNet's extensive net, which currently includes well over 13,000 verb synsets.

3. Heterogenous Troponymy Relations

Most verbs fit neatly into a given hierarchy and can be assigned to a clearly identifiable superordinate (following

[1] WordNet's verb component contains a few non-lexicalized nodes that are arguably occupied by lexical gaps. See (Fellbaum and Kegl, 1989) for discussion.

an initial stage of identifying and coding top-level concepts, WordNet was constructed bottom-up). But if one examines specific hierarchies, it becomes clear that the relation is not just polysemous along the dimensions referred to above, but semantically heterogeneous.[2]

For example, *exercise* has subordinates like *jog*, *swim*, and *bike*. But these are clearly also manners of *moving/travelling*[3]. Both the following statements are true:

(2) to jog/swim/bike is to exercise in some manner

(3) to jog/swim/bike is to move in some manner

But clearly, there is a difference. The relation between *jog*, *swim*, *bike* and *exercise* is defeasible: Not every jogging/swimming/biking event is necessarily an exercising event. By contrast, every jogging/swimming/biking event is necessarily a moving event:

(4) She jogged/swam/biked but did not exercise

(5) *She jogged/swam/biked but did not move

The concept *exercise* is definable only by means of subordinates like *swim*, *jog*, and *bike* that are shared with another subordinate, *move*. But *move* has many subordinates that are not shared with *exercise*, such as *fly* and *drive*.

The relation of *jog*, *swim* and *bike* to their superordinates *move* and *exercise* is similar to that between, e.g., *dog*, *cat*, and *goldfish* to *animal* on the one hand and to *pet* on the other hand:

(6) A dog/cat/goldfish is a kind of pet.

(7) A dog/cat/goldfish is a kind of animal.

(8) That's my dog/cat/goldfish, but it is not a pet.

(9) *That's my dog/cat/goldfish, but it is not an animal.

Just as one can recognize dogs, cats, and goldfish as animals, but not (necessarily) as pets (Guarino, 1999), so one can recognize instances of biking, swimming, jogging as moving events, but not (necessarily) as exercising events. Unlike moving, the exercise component of biking, swimming, and jogging does not supply an identity criterion and is notionally dependent. Applying the terminology of (Guarino and Welty, 2001; Guarino and Welty, 2002) for nouns to verbs, we could say that *moving* is a rigid property, and *exercising* is an anti-rigid property of a biking/swimming/jogging event. Thus, verbs like *exercise* are similar to role nouns like *pet*, and *move* is similar to type nouns like *animal*.

[2]Some of the examples discussed here are not in fact coded in the current version WordNet, 1.7.

[3]For the sake simplification, we omit other nodes that may intervene; e.g., *jog* is linked to *move* via *run*.

3.1. Consequences for a Lexical Database

(Gangemi et al., 2002) propose an important criteria for "cleaning up" an ontology like WordNet: An anti-feature cannot subsume a feature. Thus, anti-rigidity cannot subsume rigidity. (Gangemi et al., 2002) advocate eliminating all violations of this principle found among WordNet's nouns. This would cut out hierarchical links between synset pairs like *animal* and *fictitious_animal*, while leaving intact the relation between pairs like *animal* and *horse*.

3.2. Arguments for Including Heterogeneous Troponymy Relations

The verb component of WordNet contains (perhaps many) cases of heterogeneous subsumption relations, and these must be recognized and distinguished. But we argue for retaining the corresponding pointers and, in fact, for coding more instances. Our arguments are grounded largely in a pragmatic view of WordNet as an NLP tool, rather than as an ontology that is perfectly consistent with strict logical principles.

First, if links between verbs like *bike* and *exercise* were eliminated in favor of links such as between *bike* and *move, travel*, important and potentially valuable information would be lost. In some cases, the semantic relation between words that are not conforming to strict subsumption principles is more salient than between words that are properly linked. This point will be discussed further later on.

Second, lexical databases that are useful for NLP gain from a tight network of relations. Word sense disambiguation, anaphor resolution, and applications relying on measures of textual cohesion can benefit from links such as between *bike* and *exercise*.

Finally, a random search in the WordNet shows up a fair number of subsumption violations of the *jog/swim/bike* as a manner of *exercise* kind. They are not simple lexicographic errors, as demonstrated by the goodness of the formula *to jog/bike/swim is to exercise in some manner*. But at present, we don't know how common such relations are, nor whether they are distributed evenly throughout the lexicon. Eliminating them when found would preclude a systematic study of the range, variety, and distribution of these relations and a better understanding of the structure of the lexicon.

4. Representing Different Kinds of Verb Hyponymy

Various possibilities exists for representing links between *bike, swim, jog* and superordinates like *move* on the one hand and *exercise* on the other hand.

First, each verb could be linked to multiple parents by means of the same labelled "manner" pointer. However, this "tangled hierarchy" approach is clearly unsatisfactory, as it implies that every jogging/swimming/biking event is both an exercising and a moving event, when in fact only the latter is true.

The second possibility is to posit two distinct senses each for verbs like *swim*, *bike* and *jog*, each sense with a different superordinate, here *move* and *exercise*. Some traditional dictionaries take this route; for exampe, *jog* is

represented in the *American Heritage Dictionary* as having distinct *running* and *exercising* senses. But this solution has the undesirable effect of increasing polysemy. More seriously, positing two distinct senses misses the fact that is every instance of jogging-as-exercise is necessarily also an instance of moving.

A better way to capture the relevant semantic facts is to introduce two distinct kinds of super-/subordinate relation linking a single verb to two superordinates. In addition to strict hyponymy, there would be a parallel hyponymy relation with the appropriate properties.

4.1. Para-troponymy

(Cruse, 1986) proposes a relation dubbed *para-hyponymy* for organizing nouns like *dog* and *pet* hierarchically. Like regular hyponymy, para-hyponymy admits the formula *Xs and other Ys*, where X is the subordinate and Y the superordinate: Both *roses and other flowers* and *dogs and other pets* are good. This formula can easily be adopted for verbs, and fits both strict hyponymy and para-hyponymy:

(10) Biking/swimming/jogging and other manners of moving/travelling

(11) Biking/swimming/jogging and other manners of exercising

To distinguish strict hyponymy from para-hyponymy among nouns, (Cruse, 1986) cites the *but*-test:

(12) It's a dog, but it's not a pet

This test shows that the hyponymy relation between *pet* and *dog* is first, expected, and second, defeasible.

Para-hyponymy can easily be applied to concepts expressed by verbs. The pairs *walk* and *exercise*, *jog* and *exercise*, *bike* and *exercise* etc. are all good in the *but not* frame:

(13) It's a walking/jogging/biking event but it's not an exercising event.

To distinguish this relation in the verb lexicon from para-hyponymy among nouns, we will call it para-troponymy. Our proposal for WordNet or a similar lexical database designed for NLP applications then is to include among the verb relations both strict troponymy and para-troponymy.

Other examples of verbs related by para-troponymy are listed below[4]. *Brush, wipe, sweep* are para-troponyms of *clean* and troponyms of *rub*; by contrast, *steam-clean, dry-clean* are strict troponyms of *clean*. *Nod, wink, scowl, frown, pout* are para-troponyms of *gesture*, *communicate* and troponyms of *move [a specific bodypart]* (omitting several intervening nodes).

[4]The examples of para-troponyms that we have found so far intuitively suggest a similarity to the telicity of Role nouns in para-hyponymic hierarchies; para-troponyms refer to events with a specific purpose or goal, as noted in (Fellbaum, 2002)

5. Expectation

(Cruse, 1986) notes that para-hyponymy is defined not by logical necessity but by "expectation." While intuitively convincing, this notion immediately raises several questions. How can expectation be characterized? Can it be quantified? How can pairs of verbs related by para-troponymy identified in the lexicon? And how do we know whether, say, a verb token *jog* in a corpus refers to an exercising event or (merely) to a running event?

To begin with, expectation is often context-dependent rather than inherent in the concept. In some contexts, a given verb's interpretation as a para-troponym is more salient, whereas in other context, its reading as a strict troponym of another superordinate is more appropriate:

For example, *move* is more salient in (14), but *exercise* is more salient in (15):

(14) a. The boat capsized and we had to swim to the shore.

(14) b. My car is in the repair shop so I'll bike to work.

(14) c. It started to rain heavily so she ran into the library.

(15) He swims/bikes/runs 3 miles every morning before work.

Some contexts allow for an underspecified reading:

(16) He jogged to the store.

More specifically, the nature of the verb's argument projection may play a role in setting up the expectation and the appropriate reading in some cases. *Clear dishes from the table*, where the Locatum entity is the direct object, seems to favor the *remove* reading (the strict superordinate) rather than the *clean* reading (the para-superordinate); *clear the table of dishes*, with the Location entity in direct object position, appears to favor the *clean* interpretation.

Second, the degree of expectation may differ across verbs independently of specific contexts. For some verbs, the para-relation is stronger than the strict relation, and the reverse may be true for other verbs. For example, *jog* intuitively is more strongly associated with its para-superordinate *exercise* than with its logical superordinate *run, move*. This is reflected in the fact that some dictionaries have distinct running and exercising senses for *jog*, as noted earlier. Conversely, *walk* seems be more strongly associated with *move* that with *exercise*. *Walk* seems like a less canonical form of exercise than *jog*, and thus exhibits a weaker association with its para-hypernym and a correspondingly stronger link to its strict superordinate.

The relative frequency of one reading as compared to another presumably influences expectation. Just as, say, hawks as pets may be more conventional in certain cultures than in others, there are probably cultures where jogging is not done for exercise purposes but, say, for pursuing game in a hunt.

Of course, the higher frequency of one reading as compared to the other makes the former more expected and thus stronger. It would therefore be desirable to firm up intuitions about the relative strength or weakness of the (para)troponymy relation with the aid of corpus data.

Almost any verb that is a hyponym of *move* could be made a para-troponym of *exercise*, just as any animal can be called a pet. If one wants to code para-relations in the database, it is important to avoid flooding it with links that reflect readings with very low expectancy. Here, too, corpus data would be useful to identify genuine from spurious para-links.

6. Para-tronomymy in the Lexicon

This paper has cited only a handful of examples of para-troponymy. At this point, we don't know how prevalent this relation is in the lexicon, or how many cases of concepts that exist merely by virtue of contingent subordinates are lexicalized in English. To find them, we need characteristic syntactic frames and a tool to search a corpus for appropriate occurrences of such patterns[5]. This section merely offers some thoughts and suggestions for future work.

We saw that para-troponyms pass the tests adapted from the one for para-hyponymy; in this respect, para-troponyms are indistinguishable from strict troponyms:

(17) X-ing and other manners/ways/methods of Y-ing.

(18) To X is to Y in some way/manner.

Using Google to search the Web for the string *and other manners/ways of*, we turned up quite a few examples of para-troponymy and para-hyponymy, as well as some cases of regular troponymy and noun hyponymy, in addition to cases of verbs co-occurring with nominalizations. Here are some cases of para-troponymy:

(19) Flirtation, courting and other manners of getting the attention of the opposite sex is certainly a form of manipulation ... www.mothersmagic.net/Goddess/maiden.html

(20) Befriending, listening and other ways of helping.... www.britishcouncil.org/sudan/science/ - 17k

(21) volunteering and other ways to help www.fcs-sf.org/page5.html

(22) Home Cooking and other ways to save Money. www.geocities.com/ dvsclothing/cooking.html

(23) Walking and other exercise use many muscles. www.lungusa.org/diseases/exercise.html

(24) activities that repeatedly flex the knee (ie, jumping, squatting, running and other exercise). orthoinfo.aaos.org/fact/thr_report.cfm? Thread_ID=252&topcategory=Knee

(25) Swimming, running, biking, walking and other exercise that are at a time length of over 20 minutes.. www.pmssolutions.com/Hiddentruth.html

To limit the search to para-troponyms, we searched for instances where the expected relation is negated, as in the pattern in (26):

(26) It's X-ing but not Y-ing
(e.g., it's swimming but not exercising)

We found:

(27) ...and then spraying the action with a little WD-40 is not cleaning. It is a slow methodical destruction of a considerable investment. Like everything ... www.doubleought.com/cleaning.html

(28) No, this is not "cleaning for the cleaning lady", it's picking up so that the cleaning lady can clean ... www.bitchypoo.com/2000/May/11.html - 7k

Similarly, on can search for cases where the para-hyponymy is asserted, possibly over a negative presupposition, as in the pattern in (29):

(29) This X-ing is Y-ing
(e.g., This swimming is exercising)

A web search turned up examples like these:

(30) Shotblasting is a way of cleaning or preparing surfaces for recoating, using an abrasive material forced through a jet nozzle... www.westshotblasting.co.uk/

(31) ... shake hands, using the right hand, and explain that this is a way of greeting one another. Pair up children and allow them to practice shaking hands. www.atozkidsstuff.com/math.html

(32) Tipping-leaving a gratuity-is a way of thanking people for their service. www.istudentcity.com/stages/3mannerstipping.asp

Another possibility is to examine co-occurrences of verbs in contexts for cases of (defeated) para-troponymy, without using any specific patterns. The following are actual examples:

(33) really get the job done. If the goal is to have clean sidewalks, they're going to have to be swept and bagged, not just blown. www.heartlight.org/two_minute/2min_971015.html

(34) will be swept by City crews. Residential streets are now swept once a month, while downtown streets are cleaned three times a week... www.ci.walnut-creek.ca.us/street

(35) These sociologists think that interrupting is a way of exercising power. They say, "Here we are dealing with a class of speakers ... www.glc.k12.ga.us/qstd-int/ancill/guidance/schoices/sc-f20.htm

We hope to develop more sophisticated and efficient ways for finding para-relations in the lexicon in the near future and to test their usefulness in applications.

[5]Resnik, Fellbaum, and Olsen are currently developing a tool to search the Web for specific syntactic patterns.

7. Summary and Conclusions

We have argued for retaining instances of paratroponymy in a lexical database like WordNet. Furthermore, we advocate collecting and adding naturally attested cases of this relation. Semantic relations that are not based on logical necessity but on expectations grounded in pragmatics or world knowledge are an interesting area for research in lexical semantics. Enriching a lexical database with para-relations can not only shed light on the organization of the lexicon, but may yield benefits for NLP applications relying on this database.

8. Acknowledgment

We thank Alessandro Oltramari for commenting on an earlier version of this paper. This work was supported by grant number IIS-ITR 0112429 from the National Science Foundation to the author.

9. References

Alan Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

Christiane Fellbaum. 1998. *WordNet*. MIT Press.

Christiane Fellbaum. 2002. On the Semantics of Troponymy. In R. Green, S. Myang and C. Bean, editors, *Relations*, Dordrecht. Kluwer.

Christiane Fellbaum and Judy Kegl. 1989. Taxonomic Structure and Object Deletion in the English Verbal System. In: K. deJong and Y. No, editors, *Proceedings of the Sixth Eastern States Conference on Linguistics*, Columbus, OH: Ohio State University.

Nicola Guarino. 1998. Some Ontological principles for Designing Upper Level Lexical Resources. *First International Conference of Language Resources and Evaluation*, Granada, Spain.

Nicola Guarino. 1999. The Role of Identity Conditions in Ontology Design. *Proceedings of the IJCAI Workshop on Ontologies and Problem-Solving Methods*, 1–7, Stockholm.

Nicola Guarino and Chris Welty. 2002. Identity and Subsumption. *LADSEP-CNR Internal Report*, Padua, Italy.

Nicola Guarino and Chris Welty. 2002. Evaluating Ontological Decisions with Ontoclean. *Communications of the ACM*, 45.2:61-65.

Aldo Gangemi, Nicola Guarino, and Alessandro Oltramari. 2001. Conceptual Analysis of Lexical Taxonomies: The Case of WordNet Top Level. *Proceedings of FOIS*, Ogunquit, Maine.

Aldo Gangemi, Nicola Guarino, Alessandro Oltomari, and Stefano Borgo. 2002. Cleaning up WordNet's Top Level. In U. N. Singh, editor, *Proceedings of the First Global WordNet Conference*, Mysore, India. CIIL.

On the Ontological Basis for Logical Metonomy

Telic Roles and WORDNET

Sandiway Fong

NEC Research Institute
4 Independence Way
Princeton NJ
sandiway@research.nj.nec.com

Abstract

The analysis of examples of Logical Metonomy, where an event-taking verb is combined a non-eventive object, intuitively involves the recovery or insertion of a missing verb generally known as a Telic Role. For example, for *Mary enjoyed the meal*, an appropriate might be *eat*, i.e. *Mary enjoyed eating the meal*. The question for lexical semantics is where do telic roles reside and how are they accessed? In this paper, we investigate the use of WORDNET, a widely used semantic network, both as an appropriate repository and also as an organization suitable for the recovery or assignment of telic roles.

1. Introduction

The interaction of aspectual verbs such as *begin* or *finish* with simple, non-eventive noun phrases (NPs) has been used to motivate an account of *logical metonymy* in which telic (purpose/function) and agentive (creation) roles are distinguished components of the lexicon, see (Pustejovsky, 1995). Others, e.g. (Lascarides and Copestake, 1995) and (Verspoor, 1997), have highlighted the role of context and convention. Consider (1).

(1) a. John began the novel (*reading*/*writing*)
b. The author began the unfinished novel back in 1962 (*writing*)

(1a) can mean *John began reading the novel*, accessing the functional sense or telic role of *novel*, or *John began writing the novel*, accessing the specific means of creation or agentive role of *novel*. The telic/agentive role ambiguity seen in (1a) can be made less apparent in context, either within the same sentence, as in (1b) above, or through discourse or semantic inference, as in (11) and (12), to be discussed below. Note that there are important constraints, e.g. with respect to boundedness and aspect, on the possible NPs that can appear with *begin*. See (Verspoor, 1997) and the references cited therein for discussion of the relevant factors.

Other verbs such as the subject-experiencer psych verb *enjoy*, or verbs such as *refuse*, exclude the agentive role.[1] For example, contrast (2a) with (1a).

(2) a. Mary enjoyed the novel (*reading*)
b. Timmy refused the meal (*to eat*)

In (2b), *refuse* can access the telic role for *meal*, namely *to eat*. However, there is room for ambiguity here; (2b) is also compatible with the interpretation *Timmy refused* to accept *the meal*, cf. (3) below.

(3) Timmy refused the present (*accept*)

In (3), arguably the telic role of *present*, meaning *gift*, is *accept*. However, the same account cannot be posited for *meal*; its basic function (if one exists) is to be consumed or eaten; thus creating a problem for enumeration in lexical representation. In other cases, such as (4), there is no (felicitous) telic role at all.

(4) a. !John enjoyed the rock
b. !!John enjoyed the door

A physical object like *rock* has no obvious function. Yet (4a) can be marginally interpreted in the context that some (physical) aspect of the object gave *John* pleasure, e.g. its appearance as in *John enjoyed looking at the rock*. Or we can appeal to other perceptual properties, e.g. the tactile sense as in *the blind man enjoyed* touching *the rock*. To take one more example, consider (5):

(5) Mary enjoyed the garden

The prototypical definition of a garden as a pleasing arrangement of plants and other natural (or non-natural) objects admits not only the (putative) telic role *to see* but also a range of other possibilities, illustrated in (6).

(6) a. Mary enjoyed *seeing* the garden
b. Mary enjoyed *inspecting* the garden
c. Mary enjoyed *visiting* the garden
d. Mary enjoyed *strolling* through the garden
e. Mary enjoyed *rollerblading* in the garden
f. Mary enjoyed *sitting* in the garden
g. Mary enjoyed *dozing* in the garden

The ease of defeasibility of telic roles and the productivity of plausible alternatives is striking. In general, the recovery of appropriate contextual function falls outside the domain of local or specific lexical knowledge. It belongs more appropriately to systems that carry out reasoning and inference about the real world.

In fact, the recovery of contextual function is more ideally suited to ontological networks, which encode general semantic relations between abstract and concrete concepts

[1] *Enjoy* can take *write* explicitly, as in *Mary enjoyed writing the novel*. But this is not an instance of what Pustejovsky terms "coercion".

in the real world. This paper explores the application of such a network, WORDNET, to this problem. In particular, we will make use of the *isa*, or hypernymy, relation on, assuming (as required) the existence of certain commonsense, or real-world, properties of higher-level concepts, to account for a range of data.[2]

2. Hypernymy

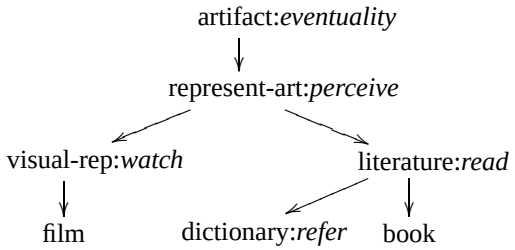
The idea that hypernymy may inform interpretation in logical metonymy has already been hinted at, or tacitly assumed, in several places in the literature. For example, this is apparent from the summary of logical metonymy in the BNC corpus, (Verspoor, 1997), excerpted in (7):

(7) eat FOOD/MEAL
drink LIQUID
tell STORY
play MUSIC
read/write WRITTEN_OBJECT
take MEDICINE/TREATMENT

(The capitalized terms in (7) denote semantically relevant concepts.)

(Lascarides and Copestake, 1995) assume the following telic roles for artifacts:

(8)



Finally, (Asher and Pustejovsky, forthcoming) assert the following complex types ($\otimes$ a type constructor):

(9) a. $p \otimes see$ and $p \otimes hear$ to encode the fact that objects with extension are typically visible, and objects involving sound are typically audible, respectively.
b. all artifacts inherit a general dependent type that gives their cause.
c. *wine*: liquid $\otimes_T$ drink ($\otimes_T$ introduces the telic role)
d. *class*: people $\otimes_T$ teach

In this paper, we explicitly test the hypothesis using the somewhat coarse-grained *isa*-relation available in WORDNET.[3] In conjunction with two principles, specificity and locality, defined with respect to hypernymy, we explain why telic/agentive roles are available for some cases but not for others. If this is the case, the locus of variation should be in ontological not lexical structure (as suggested by lexical entries such as the following):

(10) *novel*(y): telic:λx.*read*(x,y) agentive:λx.*write*(x,y)

In fact, in generative grammar, the lexicon is generally taken to be a repository of exceptions, see (Chomsky, 1965) citing Bloomfield. In this framework, non-idiosyncratic properties are factored out into grammar or further afield. Obviously, the evaluation of properties implicating mechanisms peculiar to language must stay within the domain of the language faculty. Non-language particular properties are perhaps best assimilated to general systems of reasoning and cognition.

Ontological relationships play a large role in lexical semantics and, more generally, semantic inference. Any account of language phenomenon involving the interaction of lexical entries with inheritance and (semantic) class-based behavior falls into this category. Computation involving defeasible reasoning and knowledge about the physical properties of objects in the real world should therefore fall outside the scope of the lexicon.

Furthermore, as (11), from (Lascarides and Copestake, 1995), illustrates, telic roles are easily overridden through discourse priming:

(11) a. He really enjoyed your book (*reading*)
b. My goat eats anything. He really enjoyed your book (*eating*)

Even in cases where arguably no felicitous telic role exists to be overridden, as in (12a), discourse may play a part in supplying the missing event, as in (12b):

(12) a. !He enjoyed your shoe[4]
b. My dog eats everything. He really enjoyed your shoe (*eating*)

3. The WORDNET Framework

3.1. The Hypernym Hierarchy

In WORDNET, nouns are grouped into synonym sets, known as "synsets", representing single concepts. For example, *cigarette*, *coffin nail*, *butt* and *fag* are generally substitutable, and thus belong to the same synset. Concepts are related through (possibly iterated) application of the hypernymy ("$\rightarrow$") or *isa*-relation, illustrated in (13).[5] Inheritance is strictly unidirectional in this model. For example, *tobacco* may be termed a *street drug*, but the reverse need not be true. Furthermore, multiple inheritance may obtain for some concepts. For example, *tobacco* is a *plant product* as well as a *street drug*.

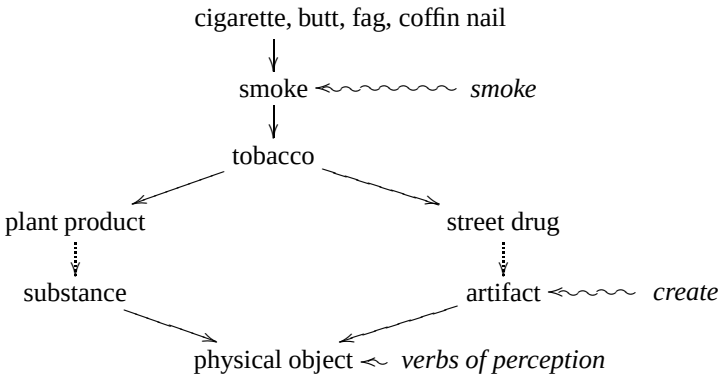
[2]The idea of using WORDNET on object NPs to pick out contexts in which those NPs represent events on a class-based model is not new. (Siegal, 1998) performed a (medical) corpus study in conjunction with WORDNET to distinguish eventive and stative *have*, e.g. *the patient had a fever* (stative)/*blood loss* (eventive).

[3]As (Gangemi et al., 2001), have noted, WORDNET's hypernymy relation is a heterogeneous one, merging functional and non-functional *isa*-relations alike, e.g. isa(tobacco,plant_product) and isa(tobacco,street_drug).

[4]In the framework described in this paper, *shoe* is a "foot covering". The telic role is cover(NP,FOOT), which is incompatible with prototype V(PRO,NP) defined in section 3.. The next higher concept is "footwear" with telic role *wear*, perhaps accessed in contexts like *He enjoyed the comfortable shoes you lent him*.

[5]For brevity, a dotted arrow ("$\dashrightarrow$") will sometimes be used to represent a hypernym sequence.

(13)

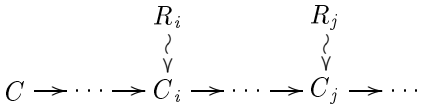


In this paper, we will assume annotation of concepts with characteristic verbs where relevant (to be indicated by "$\lesssim$"). For example, in (13) *artifact*, defined in the gloss as a "man-made object", is associated with the verb *create*. Similarly, the noun *smoke* is associated with the related verb *to smoke*.[6] Finally, the concept *physical object*, defined as "a tangible and visible entity", is characterized by verbs of perception such as *see*/*look at* and *touch*.

3.2. Contextual Function Search Rules

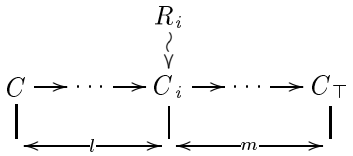
In this paper, we employ two simple principles of contextual function search over the hierarchy outlined above. In the following section, concepts will be denoted by (subscripted) C. R_i will denote a characteristic verb for a concept C_i. Given a noun $N \in C$, we have the rule of preference (14):

(14) **Principle of Specificity**: Prefer R_i to R_j in the sequence



In other words, prefer a closer role R_i over a more general one R_j in the concept chain. The (one-way) hypernymy relation relates a specific concept to a more general concept, so the closer a matching concept is in terms of the number of links, the more specific it will be. Next, given a noun $N \in C$ and $C_\top$ representing the top or most general concept relative to N, we have the rule of evaluation of the "goodness" of a characteristic verb R_i (15):[7]

(15) **Principle of Locality**: Plausibility of R_i scales with m and inversely with l in



Scalars l and m represent the length of sequences $< C, \ldots, C_i >$ and $< C_i, \ldots, C_\top >$, respectively. The closer C_i is to C (l small), the more plausible R_i will be. On the other hand, if C_i is close to $C_\top$, m will be small, encoding the intution that R_i (then) is a general characteristic that is not strongly associated with specific concept C. Rules (14) and (15) operate in tandem. Although the closest concept is always preferred, *ceteris paribus*, it will be deemed implausible or requiring of strong contextual support if it is many links from C or close to $C_\top$.

3.3. Grammatical Constraints

In what follows, we will consider the problem of determining the value of the verb V in the configuration (16b) given (16a), a restricted version of the telic role determination problem.

(16) a. EXP enjoy NP
b. EXP_i enjoy [PRO_i [V(ing) NP]]

In (16), EXP is the experiencer subject of *enjoy*, NP the object, PRO the controlled subject of V, and V a transitive verb V(PRO,NP). The twin requirements that the NP as must be the embedded object and that the subject be controlled limits the possibilities for telic roles to appear as V, as will be seen in the next section.

4. Worked Examples

Cigarette: Consider (17).

(17) Mary enjoyed the cigarette (*smoking*)

Given the hypernym hierarchy in (13), *smoke*(PRO,*cigarette*) is the strongly preferred interpretation since the concept *smoke* is highly specific (l small) and distant from general concepts *artifact* and *physical object* (m large).

Sonata: Consider the possibilities in (18).

(18) a. Mary enjoyed the sonata (*listening to*/*playing*)

[6] Concepts in WORDNET have associated glosses. A gloss will typically contain a brief definition and examples of use. In some cases, the characteristic verbs can be inferred from the gloss or from members of the synset. Further exploration of this idea is beyond the scope of this paper.

[7] In WORDNET's hypernym hierarchy there is no unique $C_\top$ concept. For example, *dirt* as *material* and as *gossip* have top concepts *entity* and *act*, respectively. See (34).

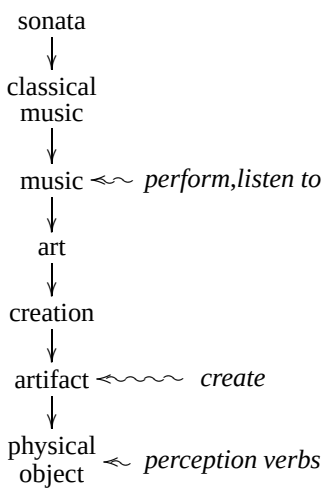
b. Mary began the sonata (*playing*/*composing*)

According to (Asher and Pustejovsky, forthcoming), the agentive and telic roles associated with *sonata* are *compose* and *play*, expressed in their type logic notation as (19).

(19) sonata: $(p \bullet i) \otimes_{A,T} (\text{compose}, \text{play})$

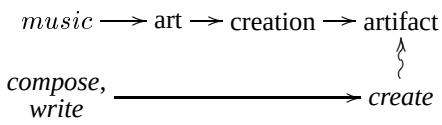
The hierarchy for *sonata* is given in (20).[8]

(20)



(20) predicts that *perform* and *listen to* are preferred in (18a). Verbs *begin* and *enjoy* differ in that *begin* allows an agentive role. This excludes subject-experiencer *listen to* but allows for *perform* and is also compatible with *create*. Note that *create* is associated with the general concept *artifact*. We can turn to WORDNET's verb hierarchy, shown superimposed in (21), to pick out the music-specific sense of *compose*.[9]

(21)

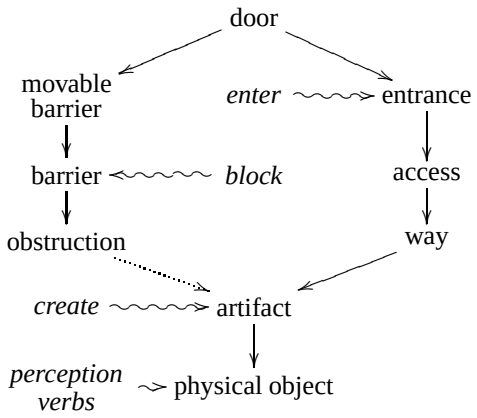


(18b) is explained since *compose* (or *write*) and *perform* are effectively equidistant from *sonata*.

Door: Consider (4b), repeated here as (22), with WORDNET hierarchy (23).

(22) !!John enjoyed the door

(23)

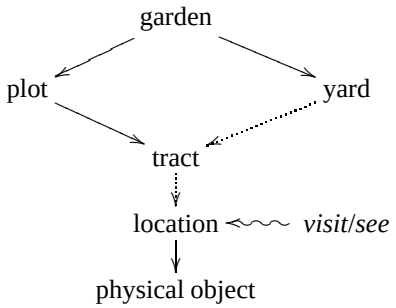


Specifically, a door can function both as an entrance (*enter*) and a barrier (*block*) to an enclosure. However, the telic verb *block* has form *block*(*door*,ENCLOSURE), which is incompatible with the prototype V(PRO,*door*), thus ruling out *block*. Similar reasoning applies to *enter*(PRO,ENCLOSURE). At the other end of the hierarchy, the canonical events associated with *physical object* are predicted to be implausible (l large, m small).

Garden: Consider (5), repeated here as (24), with WORDNET hierarchy (25).

(24) Mary enjoyed the garden (*seeing*/*visiting*)

(25)

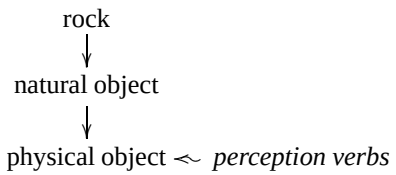


Assuming *visit* and visibility are characteristic of locations in general, (24) is accounted for. General mechanisms involved in reasoning about entailment may also play a large role in grounding *visit*. Note that the possibilities exemplified in (6) all entail *visit*.

Rock: Consider (4a), repeated here as (26).

(26) !John enjoyed the rock

(27)



Unlike *door* in (22), *rock* has no obvious function, as the simple hierarchy in (27) suggests. Hence, relatively speaking, we predict that (26), when picking out perceptual *looking at* or *touching*, is more acceptable than (22) (since l is smaller). However, the value of m is still small, indicating its acceptability can be improved significantly by contextual (discourse) support.

[8] Note, *physical object* $\rightarrow$ *entity* in WORDNET. $C_\top$ = *entity* has been omitted in (20) since *entity* has no possible characteristic functions.

[9] *Compose* and *write* belong to the same synset glossed as "*write music*". Thus the gloss locates this synset with the concept *music*.

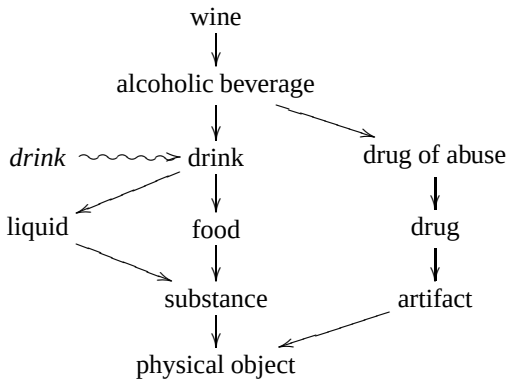
Note that WORDNET does not classify *rock* as a location, cf. *garden* in (5). Given the right context, the characteristic function *visit* may also be felicitous for *rock*, as in (28), where the rock in question is geographically significant.

(28) Mary enjoyed Ayer's Rock (*visiting*)

Wine: Consider (29) with hierarchy (30).

(29) Mary enjoyed the wine (*drinking*)

(30)

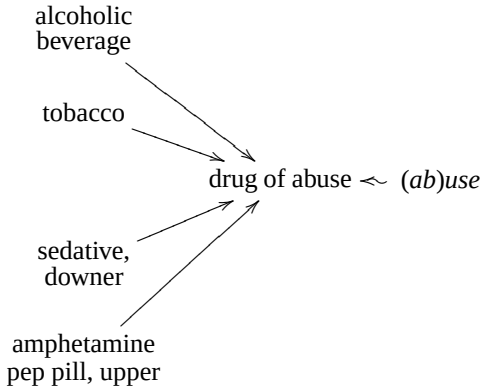


(30) strongly predicts (29) (l small, m large). However, this assumes the branch containing *drug of abuse* (with telic role (*ab*)*use*) is marginalized, i.e. *wine* as *drink* is preferred over *drug of abuse*. Contrast (29) with (31).

(31) Mary enjoyed the amphetamine/sedative (*using*)

(31) is also strongly predicted in our analysis as the elaborated WORDNET hierarchy fragment in (32) illustrates.

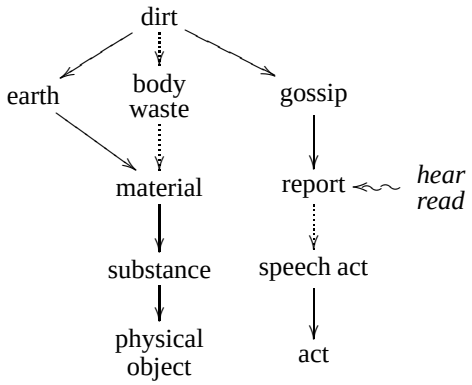
(32)



Dirt: Consider (33) with hierarchy (34).

(33) !John enjoyed the dirt

(34)



In (33), *dirt* as a natural substance has no plausible telic role. The corresponding WORDNET hierarchy is shown in (34). The relevant sense is given by the sequence <*dirt,earth,material,substance*>; the elements of which have no obvious purpose or function. Hence the status of (33).

According to WORDNET, *dirt* is also, perhaps little used, slang for fecal matter. Other (more common) words sharing the same synset are *crap*, *shit*, *poop* and *turd*. The telic role for *body waste*, perhaps *discharge*, is generally available for the synset, as can be seen by substitution of *dirt* in (33). So an appropriately annotated WORDNET makes essentially the right prediction for the synset as a whole. Finally, the right prediction is also made for *dirt* in the sense of malicious gossip, as in (35).

(35) John enjoyed the dirt on OJ Simpson (*hearing about*/*reading about*)

5. Conclusions

In this paper, we have argued for an ontological approach to the problem of logical metonymy using WORDNET's hypernymy relation for non-eventive nominals. That is, we interpret logical metonymy to be a phenomenon belonging to systems of semantic interpretation and general reasoning, governed by simple rules of specificity and locality with respect to concept hierarchy. We have shown, through worked examples, how such a mechanism accounts for data of the sort commonly cited in the literature.

Interesting questions remain for future work. For example, not all concepts in the WORDNET hierarchy have simple lexical realization satisfying the grammatical constraints, the question of what happens with lexical gaps remains. Since languages vary with respect to concept lexicalization, the question of whether the results obtained here generalize to other languages exhibiting logical metonomy remains open.

6. References

N. Asher and J. Pustejovsky. (forthcoming). The metaphysics of words in context. *Journal of Logic, Language and Information*.

N.A. Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

A. Gangemi, N. Guarino, and A. Oltramari. 2001. Conceptual analysis of lexical taxonomies: The case of wordnet top-level. In *Proceedings of FOIS 2001*.

A. Lascarides and A. Copestake. 1995. Pragmatics of word meaning. In *Semantics and Linguistic Theory (SALT5)*, Austin, Texas.

J. Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

E. V. Siegal. 1998. Disambiguating verbs with the wordnet category of the direct object. In *Workshop on Usage of WordNet in Natural Language Processing Systems*, University of Montreal, Montreal, Canada.

C. M. Verspoor. 1997. Conventionality-governed logical metonymy. In H. Bunt, L. Kievit, R. Muskens, and M. Verlinden, editors, *2nd International Workshop on Computational Semantics*, pages 302–312, Tilburg, Netherlands.

Differentiae Specificae in EuroWordNet and SIM*u*LLDA

Maarten Janssen

UiL-OTS, Utrecht University
Trans 10, 3512 ED Utrecht, The Netherlands
m.janssen@let.uu.nl

Abstract

(Euro)WordNet, like all other semantic network based formalisms, does not contain differentiae specificae. In this article, I will argue that this lack of differentiae specificae leads to a number of unsurmountable problems, not only from a monolingual point of view, but also in a multilingual setting. As an alternative, I will present the framework proposed in my thesis: SIM*u*LLDA. The SIM*u*LLDA set-up not just contains differentiae specificae (called definitional attributes), but differentiae specificae form the building blocks of the system: the relations between meanings are derived from the application of Formal Concept Analysis to the set of definitional attributes.

1. Introduction

Given the many shortcomings of systems based on semantic primitives, WordNet, like many other lexical databases and knowledge bases, is based on semantic networks (see for instance Miller (**?**)). In semantic networks, there is no need for anything like semantic markers or, as you would call them from a lexicographers point of view, differentiae specificae, since all information is formulated in terms of relations between (in the case of WordNet) synsets. In this article, I will argue that this lack of differentiae specificae leads to a number of insurmountable problems, not only from a monolingual point of view, but also in a multilingual setting.

As an alternative, I will present the framework proposed in my thesis (**?**): SIM*u*LLDA, a Structured Interlingua MultiLingual Lexical Database Application. The SIM*u*LLDA set-up not just contains differentiae specificae (which are called definitional attributes in the system), but differentiae specificae form the building blocks of the system: the relations between meanings are derived from the application of a logical formalism called Formal Concept Analysis (FCA) to the set of definitional attributes.

After the presentation of the framework, I will indicate why definitional attributes do not give these traditional problems by showing that the resulting framework should not be viewed as an ontological hierarchy, nor as a knowledge base, but as a modest lexical database.

In this article, the following notational conventions will be used: meaning-units, in the case of WordNet the synsets, will be typeset in SMALL-CAPS, word-forms are set in sans serif, differentiae specificae, as well as the relations in WordNet, in **bold-face**.

2. The Need for Differentiae Specificae

One of the main aspects of the WordNet system is its ontological hierarchy, provided by the **is_a** links. Although not de facto a separate system (the **is_a** link is just a link as any other), the hierarchy is often presented that way, and many applications of the WordNet database only make use of this ontology. So for the moment I will consider the (ontological) hierarchy of WordNet as a system on its own.

The **is_a** relation links a synset to its *genus proximum* (to use the lexicographer's term), hence strongly characterising the meaning of the synset by indicating what kind of meaning it is. But on its own, the **is_a** link does not fully characterise the meaning of the synset: it fails to distinguish the various hyponyms of the same synset. From the point of view of the hierarchy we also need *differentiae specificae* to keep the meanings/synsets within the same genus apart.

In the WordNet approach, this differentiation is done by means of the other links. As an example, one could define the synset ACTRESS by means of an **is_a** relation to ACTOR, and a **female** relation the other way around (or alternatively a **is** relation to FEMALE). But although the other links in WordNet do provide additional information about the synset, they are not designed to provide differentiae specificae. This shows in two ways: firstly, the other links give information independent of the **is_a** link, which means that they are independent of the information already provided by the **is_a** link. So they cannot structurally supplement the information lacking from the **is_a** link.

Secondly, not all differentiating information can be modelled by means of these other links. Consider for instance the word *millpond*, which **is_a** AREA OF WATER. But a millpond is not just any area of water, it is specifically one *used for driving the wheel of a watermill* (according to LDOCE). And there are no WordNet links for this type of differentiating information.

So differentiae specificae as such do not exist in WordNet, even though in some (or many) cases the differentiating information will be present or can be provided somehow. This absence of a structural modelling of differentiae specificae leads to serious problems. Let me illustrate this using three examples.

The first example is that, according to Vossen & Copestake (**?**), (Euro)WordNet has problems dealing with verb nominalisations: SMOKER is a hyponym of PERSON, but so are RUNNER, SLEEPER, JOGGER, etc. The point here is not so much that distinguishing these nominalisations is impossible in WordNet: in principle, these can be distinguished by means of the **involved_agent** relation. So we can express that the involved agent for SMOKE is SMOKER, and hence by means of backward search say that a smoker is a person *who smokes*. The point is that for synsets with large numbers of hyponyms, there is no structural way of telling them apart: WordNet in many cases depends on the ontological hierarchy, so the less layered it is, the less informative it is.

The second example makes a similar point: because of

the high dependence on hierarchy, WordNet is forced to accept as layered a structure as possible: to indicate the relation between ENEMY and MURDERER, WordNet has to introduce a synset for BAD PERSON, even though there are no words related to that synset. This introduction of 'empty synsets' is not really incorrect, but at least conceptually unattractive.

The lack of differentiae specificae is most disturbing when considered in a multilingual setting. As a third example, consider the Spanish word DEDO. It is a (translational) hyperonym of both the English FINGER, and the English TOE, since a finger is a *dedo del mano*, and a toe is a *dedo del pie*. The way this is modelled in EuroWordNet is as follows: the Spanish DEDO has an **eq_synonym** relation to an InterLingual Item (ILI) DEDO, and both the English FINGER and TOE are related to this same ILI with a relation **eq_has_hyperonym**[1]. In this way, the words finger and toe are correctly modelled as translational hyponyms of dedo.

But in this cross-linguistic linking, there is nothing keeping the two translational hyponyms *finger* and *toe* apart. That is to say, language internally, FINGER will have a **part_of** relation to HAND, and TOE to FOOT, but this information is not (directly) related to the cross-linguistic link to DEDO. Furthermore, if we would use these **part_of** relations to tell the translational hyponyms apart, they would be used as differentiae specificae. And there are other examples in which such differentiae specificae are not available. For instance, the French BIEF will be linked as a translational hyponym of CANAL, but the reason why bief is more specific (namely that it is a canal *bringing water from a stream to a hydraulic installation*) would not be modelled, because WordNet has no links to provide for it.

Such examples show that in a lexical database, there is a definite need for a structural modelling of differentiae specificae, especially in a multilingual setting. Although in this section, the criticism is specifically aimed at (Euro)WordNet, any hierarchy based system without a structural modelling of differentiae specificae will encounter the same problem, though they might show up in a different guise. Let me now turn to the system proposed in my thesis which does use differentiae specificae.

3. SIM*u*LLDA

In my thesis, I describe a multilingual lexical database set-up called SIM*u*LLDA, in which differentiae specificae play a crucial role. The differentiae specificae are modelled within the system by means of entities called *definitional attributes*. The SIM*u*LLDA system is designed to be a multilingual lexical database system from which bilingual definitions between arbitrary pairs of languages in the system can be derived.

The SIM*u*LLDA set-up consists of a number of steps: the data from monolingual dictionaries are reduced to sets of definitional attributes. These sets of definitional attributes are turned into a lattice structure by means of a logical formalism called Formal Concept Analysis (FCA). The result is a lattice structure, which can serve as a structured interlingua, connecting words from different languages. Let me show how this works using a simple example: the words for horses in English. This explanation is very brief; for a more complete explanation I refer to my thesis (**?**).

3.1. Creating Sets of Definitional Attributes

The hierarchical set-up of the SIM*u*LLDA system is best shown using a small and simple lexical field, such as the words for male, female, young, and adult horses in English. The SIM*u*LLDA system aims at modelling lexicographic data, so takes the definitions of these words as found in a monolingual dictionary as a starting point. The relevant definitions are given in table 1 (these are cleaned-up version of the definitions in the Longman Dictionary of Contemporary English, henceforth LDOCE).

colt	a young male horse
fil·ly	a young female horse
foal[1]	a young horse
mare	a fully-grown female horse
stal·lion	a fully-grown male horse

Table 1: Definitions of Words for Horses

The definitions in table 1 are analysed in the SIM*u*LLDA set-up as relating English words to defining aspects of the meanings expressed by these words. These defining attributes are called *definitional attributes*. As an example, the first definition in table 1 relates the word colt to the definitional attributes **male** and **young**. On top of these definitional attributes, colt is related to a sense of horse. But this meaning of horse is itself also related in the dictionary to definitional attributes and a further meaning of animal, etc. This will go on until the genus term is what you might call an *empty genus term*. The claim is that *thing* in a definition reading *a thing which ...* is just there because a lexical definition without a genus term is hard to formulate (in some cases). In this way, all lexical definition can be 'unravelled' into sets of definitional attributes. For simplicity, I will here ignore the relation of the words in table 1 to the word horse, and treat **horse** as if it were a definitional attribute. This leads to a situation in which the definitions in table 1 are analysed as in table 2.

	horse	**male**	**female**	**adult**	**young**
HORSE	×				
STALLION	×	×		×	
MARE	×		×	×	
FOAL	×				×
FILLY	×		×		×
COLT	×	×			×

Table 2: Definitional Attributes for Horses

So in the SIM*u*LLDA set-up, every word expresses a number of meanings, and these meanings are analysed in terms of sets of definitional attributes. And these defini-

[1]This situation is symmetrical in EuroWordNet: DEDO and FINGER are also related via the ILI FINGER. But that has no impact on the example.

tional attributes are nothing more than the accumulated differentiae specificae from their lexical definitions in monolingual dictionaries.

3.2. Formal Concept Analysis

The data in table 2 are organised within the SIM*u*LLDA set-up by means of a logical framework called Formal Concept Analysis (henceforth FCA). FCA was developed by Ganter and Wille in Darmstadt (**?**). It is an attempt to give a formal definition of the notion of a 'concept', within the boundaries of a model-theoretic framework. The idea behind FCA is the following: in a model, those objects that share a common set of attributes belong together; they form the extension of a concept, the intention of which is the set of attributes that they share.

The formal representation of FCA is follows. Take a set of objects *G*, a set of attributes *M*, and a relation *I* relating the objects to the attributes. We define the set of formal concepts $\mathfrak{B}$ over a context (G, M, I) in the following way:

$$B^{\downarrow} = \{g \in G \mid \forall b \in B \,.\, (g, b) \in I\} \quad (1)$$

$$A^{\uparrow} = \{m \in M \mid \forall a \in A \,.\, (a, m) \in I\} \quad (2)$$

$$\mathfrak{B}(G, M, I) = \{\langle A, B\rangle \mid A = B^{\downarrow} \wedge B = A^{\uparrow}\} \quad (3)$$

The way FCA is applied in SIM*u*LLDA is as follows: the meanings in table 2 are taken as formal objects (the elements of *G*), and the definitional attributes relation to them are taken as formal attributes (the elements of *M*). This lead to a set $\mathfrak{B}$ of formal concepts consisting of pairs of sets of meanings and sets of definitional attributes. There are ten such formal concepts in total, which are listed in table 3.

$\langle\{$HORSE, COLT, STALLION, MARE, FOAL, FILLY$\}, \{$**horse**$\}\rangle$
$\langle\{$MARE, FILLY$\}, \{$**horse, female**$\}\rangle$
$\langle\{$MARE$\}, \{$**horse, female, adult**$\}\rangle$
$\langle\{$STALLION, COLT$\}, \{$**horse, male**$\}\rangle$
$\langle\{$STALLION, MARE$\}, \{$**horse, adult**$\}\rangle$
$\langle\{$STALLION$\}, \{$**horse, male, adult**$\}\rangle$
$\langle\{$FOAL, COLT, FILLY$\}, \{$**horse, young**$\}\rangle$
$\langle\{$COLT$\}, \{$**horse, male, young**$\}\rangle$
$\langle\{$FILLY$\}, \{$**horse, female, young**$\}\rangle$
$\langle\emptyset, \{$**horse, female, young, male, adult**$\}\rangle$

Table 3: Formal Concepts for Horses

The formal concepts in $\mathfrak{B}$ have a natural order: formal concepts with more defining attributes are more specific those with less defining attributes. And also, all those objects that belong to a subconcept also belong to its superconcept. So we define an order relation $\leq$ over $\mathfrak{B}$ as follows:

$$\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_2 \subseteq B_1 \quad (4)$$

The relation $\leq$ orders the formal concepts in table 3 into a lattice structure, which can be displayed in a Hasse-diagram as in figure 1. The nodes in this lattice represent the formal concepts, where the related sets of meanings and attributes can be found as follows: all formal concept below the node above which the definitional attribute **young** is placed have **young** in their set of definitional attributes, and conversely, all nodes above COLT have COLT in their set of meanings (i.e. a definitional attributes **a** is put above $\langle \mathbf{a}^{\downarrow}, \mathbf{a}^{\downarrow\uparrow}\rangle$, and a meaning A is depicted under $\langle \mathrm{A}^{\uparrow\downarrow}, \mathrm{A}^{\uparrow}\rangle$).

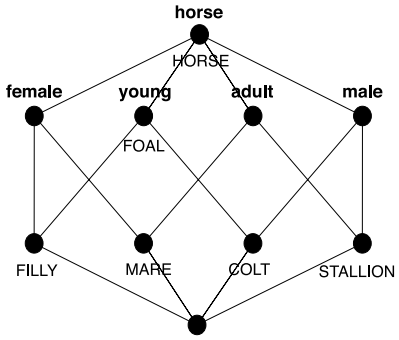


Figure 1: Concept Lattice for Horses

The construction of a concept lattice from a tabular representation of a context can be done automatically on-line by means of Java Applet written as part of my thesis. The Java-Applet is called JaLaBA (a Java Lattice Building Application). JaLaBA gives ask for a set of formal objects and a set of definitional attributes, and a relation between them, gives the related set of formal concepts, and then displays a 3D rotatable model of the corresponding Hasse diagram. JaLaBA can be found on the web-site of my thesis: `http://maarten.janssenweb.net/simullda`.

3.3. Interlingual Concept Lattice

The meanings in SIM*u*LLDA are abstracted from monolingual dictionaries. So the meanings STALLION in table 2 is derived from LDOCE. But the meaning STALLION as such is not an English meaning: the same meaning can be expressed by the French word étalon. Therefore the formal objects in SIM*u*LLDA are not taken to be language dependent meanings, but rather *interlingual meanings*, which can be expressed by words in various languages. It is clear that the definitional attributes defining these interlingual meanings cannot be language specific themselves. So also definitional attributes in SIM*u*LLDA are interlingual entities: **female** is a language independent definitional attribute, that can be lexicalised in English by the expression *female*, but also in French by the expression *femelle*, or in Dutch by the expression *mannelijk*.

Since the lattice in figure 1 thus contains only language independent entities, it can be taken as an interlingual structure, to which words of various languages can be related. This gives the situation as depicted in figure 2. Some notational conventions related to this figure: every interlingual meaning *y* has a (possibly empty) set of words lexicalising it in every language *X*, denoted by $\mathrm{wrd}_X(y)$, and every word *x* of every language has a set of interlingual meanings *Y* it expresses, denoted by $\mathrm{mng}(x)$.

In the set-up depicted in figure 2, it is possible to find translational synonyms: *x* is a translational synonym of *y*, iff $\mathrm{wrd}_Y(\mathrm{mng}(x)) \supseteq y$. To give an example:

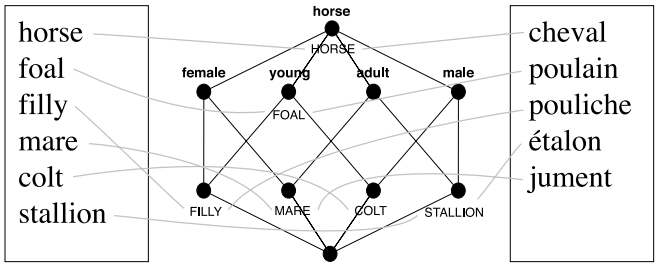


Figure 2: Concept Lattice with Words

mng(stallion) $\supseteq$ STALLION, and $\text{wrd}_{French}(\text{STALLION}) \supseteq$ étalon, so étalon is a translational synonym of stallion. In other words, just following the lines gives you translational synonyms.

More interesting is the situation when there is a lexical gap. In the SIM*u*LLDA set-up, there is a lexical gap iff $\text{wrd}(\text{mng}(x)) = \emptyset$. An example of a lexical gap in figure 2 is that there is no French translational synonym for colt. There only is the more general translational hyperonym poulain.

To find a translational hyperonym for a word *x*, first take mng*x*), and look up the lattice to find the first superconcept which has an interlingual meaning depicted under it for which there is a lexicalisation in the target language. So for colt, this interlingual meaning would be FOAL, and the fact hat poulain is a translational hyperonym of colt is modelled by the fact that COLT $\subseteq$ mng(colt), the related formal concept $\langle \text{COLT}^{\uparrow\downarrow}, \text{COLT}^{\uparrow} \rangle$ (I will use **COLT** as a name for this formal concept) is a subconcept of **FOAL**, and $\text{wrd}_{French}(\text{FOAL}) \supseteq$ poulain.

As claimed in the previous section, the things keeping colt and poulain apart should be the differentiae specificae. And differentiae specificae are implicitly present in the SIM*u*LLDA set-up: if we consider the formal concepts **COLT** and **FOAL**, then by the simple fact that **COLT** $\leq$ **FOAL**, we know that **COLT** has more definitional attributes than **FOAL**. If we define a function *ext* to give the set of definitional attributes of a formal concept ($\text{ext}(\langle A, B \rangle) = B$), then this *definitional surplus* will be $\text{ext}(\textbf{COLT}) \backslash \text{ext}(\textbf{FOAL}) =$ **male**. So **male** is the differentiam specificam distinguishing COLT from other hyponyms of FOAL such as FILLY.

The differentiae specificae, as well as the genus proximum, are hence modelled at the interlingual level. Within the interlingua, you could say that 'COLT = FOAL + **male**'. The language specific differentiae specificae are obtained by taking the lexicalisation in the desired language of this definitional surplus. We get the translation of our lexical gap by lexicalising both parts of the right-hand side of this equation in the target language. Since the French lexicalisation of **male** is *mâle*, we can conclude that colt in French is *poulain mâle*. This process of generating a translation for a lexical gap is called *lexical gap filling*. Notice that the lexical gap filling procedure renders what Zgusta (**?**) calls an *explanatory equivalent*, and not a *translational equivalent*.

We could also have opted to lexicalise all elements of the above equation within the same language, hence in English relating the word colt to the description *male foal*. In this way, also lexical definitions can be retrieved from the system. Notice that this lexical definition *male foal* is not the same definition as the one that formed the starting point of the analysis (see table 1): LDOCE does in fact not give the genus proximum, but a more remote genus term. But firstly, the rendered definition is nevertheless correct, and secondly, the LDOCE definition can also be rendered in the same way: we also have that **COLT** $\leq$ **HORSE**, with a larger definitional surplus: {**young, male**}. This leads to the original definition of colt as *young male horse*. The claim is that the generation of lexical definitions, as well as the lexical gap filling procedure, does not give a unique result, but does give only correct results.

Let me conclude this section by observing that not all definitional attributes are as 'simple' as the ones in this example. For instance, the Petit Robert definition of bief is *canal qui conduit les eaux d'un cours d'eau vers une machine hydraulique*[2]. There is no translational synonym in English for bief, but given an analysis of the data in SIM*u*LLDA, we would have that 'BIEF = CANAL + **qcled-cvumh**', where the lexicalisation in English of CANAL would be canal, and the English lexicalisation of **qcled-cvumh** would be *bringing water from a stream to a hydraulic installation*. So any differentiam specificam can be captured by a definitional attribute.

4. Definitional Attributes

As I have tried to show in the previous two sections, there is a definite need for differentiae specificae in a lexical database, especially in a multilingual one. That it is possible to set up a system using such differentiae specificae such as in the SIM*u*LLDA set-up. And that such a set-up leads to a correct modelling of lexical relations even in such problematic cases as lexical gaps. But of course the differentiae specificae introduced in a system, such as the definitional attributes in the case of SIM*u*LLDA, are at least reminiscent of the very thing WordNet reacted against: Katz & Fodor style semantics primitives (**?**). So naturally, from the perspective of semantic network theories, there is a reluctance to introduce differentiae specificae.

In the theory of Katz & Fodor, semantic markers are supposed to provide the foundation of knowledge, by their being innate building blocks to which all concepts can be reduced. But the presence of semantic primitives does not necessarily entail such a strongly reductionistic theory of meaning; there are more modest versions of semantic primitives, such as for instance in the French tradition of *sémantique interpretative*, as advocated by Rastier (**?**), Pottier (**?**) and others. The semantic primitives in this theory are called *sèmes*, which constitute meaning units calles *sémèmes*. Rastier explicitly discusses that sèmes do not have any of the strong properties semantic markers are supposed to have: they are not innate, not universal, not (interestingly) indivisible, they are not (necessarily) small in number, and they are not qualities of a referent or part of

[2]It actually is *canal de dérivation qui* ..., but I want to avoid here the for this point irrelevant question whether *canal de dérivation* should be taken as a complex genus term, or whether *de dérivation* counts as a differentiam specificam.

a concept. Especially in its description by Messelaar (**?**), sèmes have a striking resemblance to definitional attributes.

I do not want to give here an elaborate description of sèmes, their relation to semantic markers or a comparison to the SIM*u*LLDA set-up: definitional attributes are not sèmes either. But it is important to observe that the introduction of definitional attributes does not entail a strong theory of meaning. Definitional attributes are meant to be little more than what they are: theoretical entities that help to distinguish hyponyms of the same genus, and that make it possible to generate bilingual lexical definitions even for non-corresponding meanings. In my thesis, I give a lengthy discussion of the nature of the basic element of the SIM*u*LLDA set-up: words, word-forms, languages, interlingual meanings, and definitional attributes. For the moment, I will merely mention three properties definitional attributes are explicitly *not* supposed to have.

Firstly, definitional attributes do not form a special closed set of indivisible, innate semantic primitives. This should be clear from the example in section 2: the differentiam specificam *used for driving the wheel of a watermill* will constitute a definitional attribute, even though it has a clear internal structure. As a definitional attribute, it will count as an atomic entity, disregarding its internal structure[3]. So it is not an interestingly indivisible definitional attributes. And it would clearly be absurd to suppose that such a definitional attribute is in any way innate. New concepts arise every day, and new concepts can entail new definitional attributes, so there is not even a closed set of definitional attributes: new definitional attributes are introduced when need arises.

Secondly, sets of definitional attributes do not constitute a complete description of the concept related to the word that expresses the interlingual meaning in question. That is to say, interlingual meanings in the SIM*u*LLDA set-up are in a way defined in terms of sets of definitional attributes. But that does not result in saying that all information related to the word expressing that interlingual meaning is captured by the definitional attributes. For instance, stylistic information and other language-internal characteristics of the word are not modelled by the interlingual meaning, but handled at the level of the individual languages. Also, prototypes play an important role in the information/concept related to a word. But prototypes cannot be interlingual since, as shown by for instance Putnam (**?**), prototypes do not translate[4]. So the SIM*u*LLDA set-up is not supposed to provide a knowledge base: it is a lexical database, containing some aspects of word-meaning. In particular those aspects necessary for producing the kind of bilingual definition found in bilingual dictionaries.

Thirdly, definitional attributes are not denotational in nature. Definitional attributes are aspects of word meanings, not of (the) objects denoted by those words. And the interlingual meaning and/or the related set of definitional attributes are not supposed to fix the denotation of the word. Denotational semantics is very problematic, and it is even very dubious if every word(meaning) can be said to have a fixed denotation at any given moment. Furthermore, denotational semantics can never give a complete picture of word meaning. For instance, words can be metaphorically attributed to objects, where the meaning of the word is applied without the claim that the object to which it is attributed falls under the denotation of the word. So the fact that within the SIM*u*LLDA set-up, **COLT** is a subconcept of **FOAL** is not intended to express the ontological inclusion of the class of colts in the class of foals[5]: SIM*u*LLDA provides a lexical hierarchy, which should not be taken as an ontological hierarchy.

This last point is independent of the presence of differentiae specificae: also hierarchical systems without differentiae specificae, such as WordNet, should be taken as providing a lexical hierarchy, and not an ontological hierarchy. It is even dubious whether there really is an ontological ordering on the world. This is not to say that SIM*u*LLDA is not an ontology in the sense often used in computer science. For instance, the set-up is in many ways comparable to the *ontology clustering* set-up proposed by Visser & Tamma (**?**), which has a shared ontology and attributes over the concepts in it. Also in their set-up, a translation for a lexical gap is created after "*the attributes of the concept in the source ontology are compared with the attributes of the hypernym [found in the shared ontology] to select the distinguishing features.*" The point is that SIM*u*LLDA does not provide an ontological hierarchy in the philosophical sense.

Given the modest nature of definitional attributes, it will be clear that there are no strong claims concerning the meanings in the SIM*u*LLDA set-up. This is not surprising if you consider that SIM*u*LLDA aims at modelling lexicographic definitions, and lexicographic definitions do not really 'give' a description of the meanings of a word; they rely on knowledge of related words to 'hint at' the meaning of the word. A nice example of this is given by Hanks (**?**), who shows that a lexicographic definition of a chinaman (say *a left-hander's googly*) is only useful if you know about googlies, leg breaks, off-breaks and related cricket terms. Given the elusive nature of words, any theory that makes strong(er) claims is likely to runs into grave problems.

5. Conclusion

In this article, I hope to have shown the need for a structural modelling of differentiae specificae in a (multilingual) lexical database, and the advantages of the SIM*u*LLDA set-up which has such differentiae specificae by means of its definitional attributes. As already said, the criticism in this article was mainly directed at the EuroWordNet set-up, but applies equally to other hierarchical systems without differentiae specificiae. For instance, as far as I can tell, the

[3] In my thesis, I discuss some cases in which adopting a certain internal structure for definitional attributes proves beneficial, and also discuss order sets of definitional attributes, but in general, definitional attributes are atomic.

[4] Putnam goes on to claim that *perceptual prototypes may be psychologically important, but they just aren't* meanings – *not even "narrow" ones (op.cit. p.46).*. Although I am not unsympathetic with this point, it is not this strong claim I am aiming at here.

[5] This independently of the questions whether all colts are in fact foals.

SIMPLE framework, which in a way is a succesor of EuroWordNet, does not add structure to overcome the problems described in section 2.

Of course, the question whether SIM*u*LLDA could really provide a better alternative for a system like EuroWordNet is an (at least partly) empirical question: lexical databases and knowledge bases are designed for practical applicability. The SIM*u*LLDA approach is, however, a theoretical feasibility study, performed as a PhD-project, and the SIM*u*LLDA system has not (yet) been implemented or tested at large scale.

This is not to say that there is no empirical evidence for the applicability of the system: in my thesis, there is an empirical test whether the around 50 words for bodies of water from 6 different languages (English, French, Dutch, German, Italian, and Russian) can be correctly handled within the SIM*u*LLDA set-up. Describing the results of this test here would be too lengthy, and the test did bring forward some problems (or weaknesses) of the set-up. But the claim is that all the problems that have a solutions could be solved to satisfaction within the system. Although this does not provide a large-scale test, it does show that within an actual domain of lexical definitions, the systems works properly. The lexical field was not arbitrarily chosen, but was taken because it is a lexical field that is often quoted as problematic, both in terms of definability, as in terms of cross-linguistic differences, such as the often cited case of *river* and *fleuve*. So it is intended to provide some empirical evidence for the practical applicability of the system. But the only way to really test it is of course to build an application and fill it with data.

6. References

Bernhard Ganter and Rudolf Wille. 1996. *Formale Begriffsanalyse: mathematische grundlagen*. Springer Verlag, Berlin.

Patrick Hanks. 2000. Contributions of lexicography and corpus linguistics to a theory of language performance. In *Proceedings of the Ninth Euralex International Congress*, Stuttgart.

Maarten Janssen. 2002. *SIMuLLDA: a Multilingual Lexical Database Application using a Structured Interlingua*. Ph.D. thesis, Universiteit Utrecht, Utrecht.

Jerrold J. Katz and Jerry A. Fodor. 1963. The structure of a semantic theory. *Language*, vol. 39:170 – 210.

P.A. Messelaar. 1990. *La Confection du Dictionnaire Générale Bilingue*. Peeters, Leuven.

George A. Miller. 1998. Foreword. In Christiane Fellbaum, editor, *Wordnet: an Electronic Lexical Database*. MIT Press, Cambridge.

Bernard Pottier. 1980. Sémantique et noémique. *Annuario de Estudios filológicos*, vol. 3:169 – 177.

Hillary Putnam. 1988. Fodor and block on "narrow content". In *Representation and Reality*. MIT Press, Cambridge.

François Rastier. 1987. *Sémantique Interprétative*. Presses Universitaires de France, Paris.

Pepijn R.S. Visser and Valentina A.M. Tamma. 1999. An experience with ontology-based agent clustering. In Benjamins, Chandrasekaran, Gomez-Perez, Guarino, and Uschold, editors, *Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5)*, Stockholm.

Piek Vossen and Ann Copestake. 1993. Untangling definition structure into knowledge representation. In Ted Briscoe, Valeria de Paiva, and Ann Copestake, editors, *Inheritance, Defaults, and the Lexicon*. Cambridge University Press, Cambridge.

Ladislav Zgusta. 1971. *Manual of Lexicography*. Mouton, Den Haag.

Merging Global and Specialized Linguistic Ontologies

Bernardo Magnini and Manuela Speranza

ITC-irst, Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), Italy
{magnini, manspera}@itc.it

Abstract

There is an increasing interest in linguistic ontologies (e.g. WordNet) for a variety of content-based tasks, including conceptual indexing, word sense disambiguation and cross-language information retrieval. A relevant contribution in this direction is represented by linguistic ontologies with domain specific coverage, which are a crucial topic for the development of concrete application systems.
This paper tries to go a step further in the direction of the interoperability of specialized linguistic ontologies, by addressing the problem of their integration with global ontologies. This scenario poses some simplifications with respect to the general problem of merging ontologies, since it enables to define a strong precedence criterion so that terminological information overshadows generic information whenever conflicts arise. We assume the EuroWordNet model and propose a methodology to "plug" specialized linguistic ontologies into global ontologies. Experimental data related to an implemented algorithm, which has been tested on a global and a specialized linguistic ontology for the Italian language, are provided.

1. Introduction

Ontologies have become an important topic in research communities across several disciplines in relation to the key challenge of making the Internet and the Web a more friendly and productive place by filling more meaning to the vast and continuously growing amount of data on the net. The surging interest in the discovery and automatic or semi-automatic creation of complex, multi-relational knowledge structures, in fact, converges with recent proposals from various communities to build a Semantic Web relying on the use of ontologies as a means for the annotation of Web resources.

There is also an increasing interest in linguistic ontologies, such as WordNet, for a variety of content-based tasks, such as conceptual indexing and semantic query expansion to improve retrieval performance. More recently, the role of linguistic ontologies is also emerging in the context of distributed agents technologies, where the problem of meaning negotiation is crucial. A relevant perspective in this direction is represented by linguistic ontologies with domain specific coverage, whose role has been recognized as one of the major topics in many application areas.

This paper tries to go a step further in the direction of the interoperability of specialized linguistic ontologies, by addressing the problem of their integration with global linguistic ontologies. The possibility of merging information at different levels of specificity seems to be a crucial requirement at least in the case of large domains where terminologies include both very specific terms and a significant amount of common terms that may be shared with global ontologies.

The global-specialized scenario poses some simplifications with respect to the general problem of merging ontologies at the same degree of specificity (Hovy, 1998); in particular, in the case of conflicting information, it is possible to define a strong precedence criterion according to which terminological information overshadows generic information. We assume the EuroWordNet model and propose a methodology to "plug" specialized linguistic ontologies into global ontologies. The formal apparatus to realize this is based on plug relations that connect *basic concepts* of the specialized ontology to corresponding concepts in the generic ontology. We provide experimental data to support our approach, which has been tested on a global and a specialized linguistic ontology for the Italian language.

The paper is structured as follows. Section 2 presents the main features and uses of linguistic ontologies as opposed to formal ontologies. Section 3 introduces specialized linguistic ontologies, i.e. linguistic ontologies with domain specific coverage, as opposed to global linguistic ontologies containing generic knowledge. Section 4 deals with the problem of the interoperability of linguistic ontologies and describes the relations and the procedure enabling an integrated access of pairs of global and specialized linguistic ontologies.

2. Linguistic ontologies versus formal ontologies

In the recent years the increasing interest in ontologies for many natural language applications has led to the creation of ontologies for different purposes and with different features; therefore, it is worth pointing out the distinction between two main kinds of existing ontologies, i.e. formal and linguistic ontologies.

Linguistic ontologies are large scale lexical resources that cover most words of a language, while at the same time also providing an ontological structure where the main emphasis is on the relations between concepts; linguistic ontologies can therefore be seen both as a particular kind of lexical database and as particular kind of ontology.

Linguistic ontologies mainly differ from formal ontologies as far as their degree of formalization is concerned. Linguistic ontologies, in fact, do not reflect all the inherent aspects of formal ontologies. As Guarino et al. (1999) point out, for instance, WordNet's upper level structure shows no distinction between types and roles, whereas most of the original Pangloss (Knight and Luk, 1994) nodes in the Sensus ontology are actually types; to give a further example, WordNet's hierarchical structure lacks information about mutual disjointness between concepts.

Moreover, what distinguishes linguistic ontologies from formal ontologies, is their size: linguistic ontologies are very large (WordNet, for instance, has several dozen thousand synsets), while formal ontologies are generally much smaller.

The duality characterizing linguistic ontologies is reflected in their most prominent features. If we consider the linguistic level, they are strongly language-dependent, like electronic dictionaries, glossaries and all other linguistic resources, which focus on the words used in one specific language (in the case of monolingual resources) or in two or more specific language (in the case of bilingual or multilingual resources). On the other hand, if we consider the semantic level, we can observe that concepts denotated by different words in different languages can be shared, as it happens with the concepts in formal ontologies. In fact it is possible, at least for the core Indo-European languages, to identify a common ontological backbone behind the lexical surface of different languages (Guarino et al., 1999).

WordNet (Fellbaum, 1998), the best-known linguistic ontology, is an electronic lexical database where each sense of a lemma belongs to a different synset, i.e. a set of synonyms. Synsets are organized hierarchically by means of hypernymy and hyponymy relations. In WordNet other kinds of semantic relations among synsets are defined (e.g. role relation, part-of relation and cause relation), so as to build a more rich and complex semantic net. WordNet thus offers two distinct services: a lexicon, which describes the various word senses, and an ontology, which describes the semantic relationships among concepts.

As a linguistic ontology, WordNet is strongly language-dependent, but as an ontology it could also be adapted to a cross-language environment using the EuroWordNet multilingual database (Vossen, 1998) and mapping synsets into the EuroWordNet InterLingual Index, i.e. the index that links monolingual wordnets for all the languages covered by EuroWordNet. There are several examples of monolingual wordnets for many other languages, such as Dutch, Spanish, Italian, German and Basque.

A formal ontology based on linguistic motivation is the Generalized Upper Model (GUM) knowledge base (Bateman et al., 1995), an ontology primarily developed for Natural Language Processing applications. An upper model is an abstract linguistically motivated ontology meeting two requirements at the same time: i) a sufficient level of abstraction in the semantic types employed, as to escape the idiosyncrasies of surface realization and ease interfacing with domain knowledge, and ii) a sufficiently close relationship to surface regularities as to permit interfacing with natural language surface components.

2.1. Uses of formal ontologies

Recently ontologies have been used in the context of the Semantic Web. Ontologies can be employed to associate meaning with data and documents found on the Internet thus boosting diverse applications of information-retrieval systems. For the retrieval of information from the Web, Luke et al. (1996) propose a set of simple HTML Ontology Extensions to manually annotate Web pages with ontology-based knowledge, which performs high precision but is very expensive in terms of time.

OntoSeek (Guarino et al., 1999) is also based on content, but uses ontologies to find user's data in a large classical database of Web pages. Erdmann and Studer (1999) use an ontology to access sets of distributed XML documents on a conceptual level. Their approach defines the relationship between a given ontology and a document type definition (DTD) for classes of XML document. Thus, they are able to supplement syntactical access to XML documents by conceptual access.
However, as pointed out by Guarino et al. (1999), the practical adoption of ontologies in information-retrieval systems is limited by their insufficiently broad coverage and their need to be constantly updated; linguistic ontologies encompass both ontological and lexical information thus offering a way to partly overcome these limitations.

2.2. Uses of linguistic ontologies

Linguistic ontologies, and WordNet in particular, are proposed for content-based indexing, where semantic information is added to the classic word-based indexing. As an example, *Conceptual Indexing* (Woods, 1997) automatically organizes words and phrases of a body of material into a conceptual taxonomy that explicitly links each concept to its most specific generalizations. This taxonomic structure is used to organize links between semantically related concepts, and to make connections between terms of a request and related concepts in the index.

Mihalcea and Moldovan (2000) designed an IR system which performs a combined word-based and sense-based indexing exploiting WordNet. The inputs to IR systems consist of a question/query and a set of documents from which the information has to be retrieved. They add lexical and semantic information to both the query and the documents, during a preprocessing phase in which the input question and the texts are disambiguated. The disambiguation process relies on contextual information, and identifies the meaning of the words using WordNet.

The proble of sense disambiguation in the context of an IR task has been addressed, among the others, also by Gonzalo et al. (1998). In a preliminary experiment where disambiguation had been done manually, the vector space model for text retrieval gives better results if WordNet synsets are chosen as the indexing space, instead of word forms.

Desmontils and Jacquin (2001) present an approach where linguistic ontologies are used for information retrieval on the Internet. The indexing process is divided into four steps: i) for each page a flat index of terms is built; ii) WordNet is used to generate all candidate concepts which can be labeled with a term of the previous index; iii) each candidate concept of a page is studied to determine its representativeness of this page content; iv) all candidate concepts are filtered via an ontology, selecting the more representative for the content of the page.

More recently, the role of linguistic ontologies is also emerging in the context of distributed agents technologies, where the problem of meaning negotiation is crucial (Bouquet and Serafini, 2001).

3. Specialized linguistic ontologies

A particular kind of linguistic ontologies is represented by specialized linguistic ontologies, i.e. linguistic ontologies with domain specific coverage, as opposed to global linguistic ontologies, which contain generic knowledge. Focusing on one single domain, specialized linguistic ontologies often provide many sub-hierarchies of highly specialized concepts, whose lexicalizations tend to assume the shape of complex terms (i.e. multi-words); high level knowledge, on the other hand, tends to be simplified and domain oriented.

Many specialized linguistic ontologies have been developed, especially for practical applications, in domains such as art (see the Art and Architecture Getty Thesaurus), geography (see the Getty Thesaurus of Geographical Names), medicine (Gangemi et al., 1999), etc. and the importance of specialized linguistic ontologies is widely recognized in a number of works. The role of terminological resources for Natural Language Processing is addressed, for instance, by Maynard and Ananiadou (2000), who point out that high quality specialized resources such as dictionaries and ontologies are necessary for the development of hybrid approaches to automatic term recognition combining linguistic and contextual information with statistical information.

Buitelaar and Sacaleanu (2002) address the problem of tuning a general linguistic ontology such as WordNet or GermaNet to a specific domain (the medical domain, in the specific case). This involves both selecting the senses that are most appropriate for the domain and adding novel specific terms. Similarly, Turcato et al. (2000), describe a method for adapting a general purpose synonym database, like WordNet, to a specific domain (in this case, the aviation domain), adopting an eliminative approach based on the incremental pruning of the original database.

The use of domain terminologies also arises the problem of the (automatic) acquisition of thematic lexica and their mapping to a generic resource (Buitelaar and Sacaleanu, 2001; Vossen, 2001; Lavelli et al., 2002). As far as automatic term extraction is concerned, Basili et al. (2001) investigate whether syntactic context (i.e. structural information on local term context) can be used for determining "termhood" of given term candidates, with the aim of defining a weakly supervised "termhood" model suitably combining endogenous and exogenous syntactic information.

4. Merging global and specialized linguistic resources: the plug-in approach

One of the basic problems in the development of techniques for the Semantic Web is the integration of ontologies. Indeed the Web consists of a variety of information sources, and in order to extract information from such sources, their semantic integration is required.

Merging linguistic ontologies introduces issues concerning the amount of data to be managed (in the case of WordNet we have several dozen thousand synsets), which are typically neglected when upper levels are to be merged (Simov et al., 2001).

This paper tries to go a step further in the direction of the interoperability of linguistic ontologies, by addressing the problem of the integration of global and specialized linguistic ontologies. The possibility of merging information at different levels of specificity seems to be a crucial requirement at least in the case of domains, such as Economics or Law, that includes both very specific terms and a significant amount of common terms that may be shared by the two ontologies. We assume the EuroWordNet model and propose a methodology to "plug" specialized ontologies into global ontologies, i.e. to access them in conjunction through the construction of an integrated ontology.

4.1. Correspondences between global and specialized linguistic ontologies

A global linguistic ontology and a specialized one complement each other. The one contains generic knowledge without domain specific coverage, the other focuses on a specific domain, providing sub-hierarchies of highly specialized concepts. This scenario allows some significant simplifications when compared to the general problem of merging two ontologies. On the one hand, we have a specialized ontology, whose content is supposed to be more accurate and precise as far as specialized information is concerned; on the other hand, we can assume that the global ontology guarantees a more uniform coverage as far as high level concepts are concerned. These two assumptions provide us with a powerful precedence criterion for managing both information overlapping and inheritance in the integration procedure.

In spite of the differences existing between the two ontologies, in fact, it is often possible to find a certain degree of correspondence between them. In particular, we have information *overlapping* when the same concept belongs to the global and to the specialized ontology, and *over-differentiation* when a terminological concept has two or more corresponding concepts in the global ontology or the other way round. Finally, some specific concepts referring to technical notions may have no corresponding concept in the global ontology, which means there is a *conceptual gap*; in such cases a correspondence to the global ontology can be found through a more generic concept.

The sections highlighted in the global and the specialized ontology represented in Figure 1 reflect the correspondences we typically find between the two kinds on ontologies.

As for the global ontology (the bigger triangle), area *B1* is highlighted since it corresponds to the sub-hierarchies containing the concepts belonging to the same specific domain of the specialized ontology (the smaller triangle). The middle part of the specialized ontology, which we call *B* area, is also highlighted and it corresponds to concepts which are representative of the specific domain but are also present in the global ontology.

When the two ontologies undergo the integration procedure, an integrated ontology is constructed (Figure 2). Intuitively, we can think of it as if the specialized ontology somehow shifts over the global. In the integrated ontology, the information of the generic is maintained, with the exclusion of the sub-hierarchies containing the concepts belonging to the domain of the specialized ontology, which are covered by the corresponding area of the specialized. The

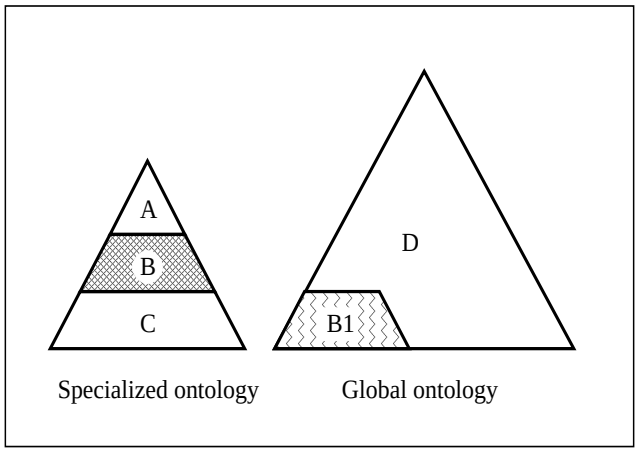


Figure 1: Separate specialized and global ontologies. Overlapping is represented in colored areas

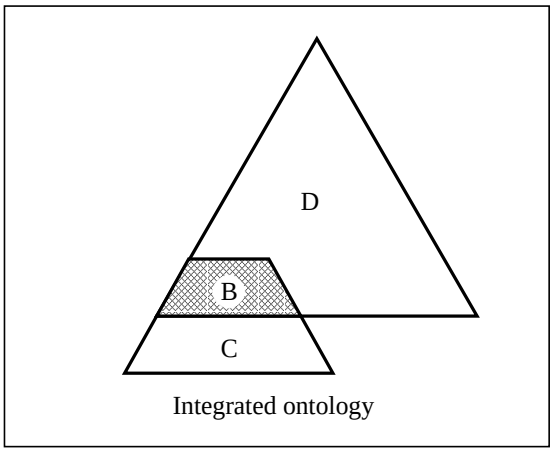


Figure 2: Integrated ontology. As to overlapping, precedence is given to the specialized ontology

integrated ontology also contains the most specific concepts of the specialized ontology (*C* area), which are not provided in the generic. What is excluded from the integrated ontology is the highest part of the hierarchy of the specialized ontology; it is represented by area *A* and contains generic concepts not belonging to a specialized domain, which are expected to be treated more precisely in the generic ontology.

4.2. Plug relations

The formal apparatus to realize an integrated ontology is based on the use of three different kinds of relations (plug-synonymy, plug-near-synonymy and plug-hyponymy) that connect basic concepts of the specialized ontology to the corresponding concepts in the global ontology, and on the use of eclipsing procedures that shadow certain concepts, either to avoid inconsistencies, or as a secondary effect of a plug relation.

A plug relation directly connects pairs of corresponding concepts, one belonging to the global ontology and the other to the specialized ontology. The main effect of a plug relation is the creation of one or more "plug concepts", which substitute the connected concepts, i.e. those directly involved in the relation. To describe the relations inherited by a plug concept, the following classification, adapted from Hirst and St-Onge (1998) is used: *up-links* of a concept are those whose target concept is more general (i.e. hypernymy and instance-of relations), *down-links* are those whose target is more specific (i.e. hyponymy and has-instance relations) and *horizontal-links* include all other relations (i.e. part-of relations, cause relations, derivation, etc.).

Plug-synonymy is used when overlapping concepts are found in the global ontology (hereafter *GO*) and in the specialized ontology (hereafter *SO*). The main effect of establishing a relation of plug-synonymy between concept *C* belonging to the global ontology (indicated as C^{GO}) and $C1^{SO}$ (i.e. concept *C1* belonging to the specialized ontology) is the creation of a plug concept $C1^{PLUG}$. The plug concept gets its linguistic forms (i.e. synonyms) from *SO*, up-links from GO, down-links from *SO* and horizontal-links from *SO* (see Table 1). As a secondary effect, the up relations of $C1^{SO}$ and the down relations of C^{GO} are eclipsed.

	$C1^{PLUG}$
Up links	*GO*
Down links	*SO*
Horizontal links	*GO* + *SO*

Table 1: Merging rules for plug-synonymy and plug-near-synonymy.

Plug-near-synonymy is used in two cases: (i) over-differentiation of the *GO*, i.e. when a concept in the *SO* has two or more corresponding concepts in the *GO*; this happens, for instance, when regular polysemy is represented in the GO but not in the SO; (ii) over-differentiation of the *SO*, i.e. when a concept in the *GO* corresponds to two or more concepts in the *SO*; this situation may happen as a consequence of subtle conceptual distinctions made by domain experts, which are not reported in the global ontology. Establishing a plug-near-synonymy relation has the same effect of creating a plug-synonymy (see Table 1).

Plug-hyponymy is used to connect concepts of the specialized ontology to more generic concepts in the case of conceptual gaps. The main effect of establishing a plug-hyponymy relation between C^{GO} (i.e. concept *C* of the global ontology) and $C1^{SO}$ (i.e. concept *C* of the specialized ontology) is the creation of the two plug concepts C^{PLUG} and $C1^{PLUG}$ (see Table 2). C^{PLUG} gets its linguistic forms from the *GO*, up-links from the *GO*, down-links are the hyponyms of C^{GO} plus the link to $C1^{PLUG}$ and horizontal-links from the *GO*. The other plug node, $C1^{PLUG}$, gets its linguistic form from the *SO*, C^{PLUG} as hypernym, down links from the *SO* and horizontal links from the *SO*. As a secondary effect, the hypernym of $C1^{SO}$ is eclipsed.

Eclipsing is a secondary effect of establishing a plug re-

	C^{PLUG}	$C1^{PLUG}$
Up links	*GO*	C^{PLUG}
Down links	*GO* + $C1^{PLUG}$	*SO*
Horizontal links	*GO*	*SO*

Table 2: Merging rules for plug-hyponymy

lation and is also an independent procedure used to avoid the case that pairs of overlapping concepts placed inconsistently in the taxonomies are included in the merged ontology; this could happen, for instance, when "whale" is placed under a "fish" sub-hierarchy in a common sense ontology, while also appearing in the mammal taxonomy of a scientific ontology.

4.3. Integration procedure

The plug-in approach described in the previous subsection has been realized by means of a semi-automatic procedure with the following four main steps.

(1) Basic concepts identification. The domain expert identifies a preliminary set of "basic concepts" in the specialized ontology. These concepts are highly representative of the domain and are also typically present in the global ontology. In addition, it is required that basic concepts are disjoint among each other and that they assure a complete coverage of the specialized ontology, i.e. it is required that all terminal nodes have at least one basic concept in their ancestor list.

(2) Alignment. This step consists in aligning each basic concept with the more similar concept of the global ontology, on the basis of the linguistic form of the concepts. Then, for each pair a plug-in configuration is selected among those described in Section 4.2.

(3) Merging. For each plug-in configuration an integration algorithm reconstructs the corresponding portion of the integrated ontology. If the integration algorithm detects no inconsistencies, the next plug-in configuration is considered, otherwise step 4 is called.

(4) Resolution of inconsistencies. An inconsistency occurs when the implementation of a plug-in configuration is in contrast with an already realized plug-in. In this case the domain expert has to decide which configuration has the priority and consequently modify the other configuration, which will be passed again to step 2 of the procedure.

5. Experiments

The integration procedure described in Section 4.3 has been tested within the SI-TAL project [1] to connect a global wordnet and a specialized wordnet that have been created independently. ItalWordNet (IWN) (Roventini et al., 2000), which was created as part of the EuroWordNet project (Vossen, 1998) and further developed through the introduction of adjectives and adverbs, is the lexical database involved in the plug-in as a generic resource and consists of about 45,000 lemmas. Economic-WordNet (ECOWN) is a specialized wordnet for the economic domain and consists of about 5,000 lemmas distributed in about 4,700 synsets. Table 3 summarizes the quantitative data of the two resources considered.

	Specialized	Generic
Synsets	4,687	49,108
Senses	5,313	64,251
Lemmas	5,130	45,006
Internal Relations	9,372	126,326
Variants/synsets	1.13	1.30
Senses/lemmas	1.03	1.42

Table 3: IWN and ECOWN quantitative data

As a first step, about 250 basic synsets (5.3% of the resource) of the specialized wordnet were manually identified by a domain expert, including, for instance "azione" ("share"), and excluding less informative synsets, such as "azione" ("action"). Alignment with respect to the generic wordnet (step 2 of the procedure) is carried out with an algorithm that considers the match of the variants. Candidates are then checked by the domain expert, who also chooses the proper plug relation. In the case of gaps, a synset with a more generic meaning was selected and a plug-hyponymy relation was chosen.

At this point the merging algorithm takes each plug relation and reconstructs a portion of the integrated wordnet. In total, 4,662 ECOWN synsets were connected to IWN: 577 synsets (corresponding to area *B* in Figure 2) substitute the synsets provided in the global ontology to represent the corresponding concepts (*B1* area in Figure 1); 4085 synsets, corresponding to the most specific concepts of the domain (*C* area in Figure 2) are properly added to the database. 25 high level ECOWN synsets (*A* area in Figure 1) were eclipsed as the effect of plug relations. The number of plug relations established is 269 (92 plug-synonymy, 36 plug-near-synonymy and 141 plug-hyponymy relations), while 449 IWN synsets with an economic meaning were eclipsed, either as a consequence of plug relations (when the two taxonomic structures are consistent) or through the independent procedure of eclipsing (when the taxonomies are inconsistent). Each relation connects on average 17,3 synsets.

6. Conclusions

After discussing the main features and uses of linguistic ontologies as opposed to formal ontologies, we have addressed the problem of the interoperability between linguistic ontologies. We have presented a methodology for the integration of a global and a specialized linguistic ontology. The global-specialized situation allows to define a strong precedence criterion to solve cases of conflicting information. The advantage of the approach is that a limited number of plug relations allows to connect a large amount of concepts (i.e. synsets) in the two ontologies.

[1] Si-TAL (Integrated System for the Automatic Treatment of Language) is a National Project devoted to the creation of large linguistic resources and software for Italian written and spoken language processing.

7. References

R. Basili, M.T. Pazienza, and F.M. Zanzotto. 2001. Modelling syntactic context in automatic term extraction. In *Proc. of Recent Advances in Natural Language Processing (RANLP '01)*, Tzigov Chark, Bulgaria, September.

J.A. Bateman, B. Magnini, and G. Fabris. 1995. The generalized upper model knowledge base: Organization and use. In *Proc. of International Conference on Bulding and Sharing of Very Large-Scale Knowledge Bases*, Twente, The Netherlands, April.

P. Bouquet and L. Serafini. 2001. Two formalizations of a context: a comparison. In *Proc. of Third International Conference on Modeling and Using Context*, Dundee, Scotland, July.

P. Buitelaar and B. Sacaleanu. 2001. Ranking and selecting synsets by domain relevance. In *Proc. of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, June. held in conjunction with NAACL2001.

P. Buitelaar and B. Sacaleanu. 2002. Extending synsets with medical terms. In *Proc. of the First Global WordNet Conference*, Mysore, India, January.

E. Desmontils and C. Jacquin. 2001. Indexing a web site with a terminology oriented ontology. In *Proc. of SWWS International Semantic Web Working Symposium*, Stanford University, USA, July, August.

M. Erdmann and R. Studer. 1999. Ontologies as conceptual models for XML documents. In *Proc. of the Twelfth Workshop on Knowledge Acquisition, Modeling and Management (KAW '99)*, Voyager Inn, Banff, Alberta, Canada, October.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, US.

A. Gangemi, D.M. Pisanelli, and G. Steve. 1999. Overview of the ONIONS project: Applying ontologies to the integration of medical terminologies. *Data and Knowledge Engineering*, 31.

J. Gonzalo, F. Verdejio, Chugur, and J. Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In S. Harabagiu, editor, *Proceeding of the Workshop "Usage of WordNet in Natural Language Processing Systems"*, Montreal, Quebec, Canada, August.

N. Guarino, C. Masolo, and G. Vetere. 1999. OntoSeek: Contet-based access to the web. *IEEE Intelligent Systems and Their Application*, 14(3):70–80.

G. Hirst and D. St-Onge. 1998. Lexical chains representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet. An Electronic Lexical Database*. The MIT Press.

E. Hovy. 1998. Combining and standardizing large-scale, pratical ontologies for machine translation and other uses. In *Proc. of the First International Conference on Language Resources and Evaluation*, Granada, Spain, August.

K. Knight and S. Luk. 1994. Building a large knowledge base for machine translation. In *Proceedings of the American Association of Artificial Intelligence Conference AAAI-94*, Seattle, WA.

A. Lavelli, B. Magnini, and F. Sebastiani. 2002. Building thematic lexical resources by bootstrapping and machine learning. In *Proc. of the Workshop "Linguistic Knowledge Acquisition and Representation: Bootstrapping Annotated Language Data", Workshop at LREC-2002*. to appear.

S. Luke, L. Spector, and D. Rager. 1996. Ontology-based knowledge discovery on the world-wide-web. In *Proc. of the AAAI1996 Workshop on Internet-based Information Systems*, Portland, Oregon, August.

D. Maynard and S. Ananiadou. 2000. Creating and using domain-specific ontologies for terminological applications. In *Proc. of Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, May, June.

R. Mihalcea and D. Moldovan. 2000. Semantic indexing using WordNet senses. In *Proc. of the ACL workshop on Recent Advances in Natural Language Processing and Information Retrieval*, Hong Kong, October.

A. Roventini, A. Alonge, F. Bertagna, B. Magnini, and N. Calzolari. 2000. ItalWordNet: a large semantic database for Italian. In *Proc. of the Second International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, May, June.

K. I. Simov, K. Kiryakov, and M. Dimitrov. 2001. OntoMap - the guide to the upper-level. In *Proc. of SWWS International Semantic Web Working Symposium*, Stanford University, USA, July, August.

D. Turcato, F. Popowich, J. Toole, D. Fass, D. Nicholson, and G. Tisher. 2000. Adapting a synonym database to specific domains. In *Proc. of Workshop on Information Retrieval and Natural Language Processing*, Hong-Kong, October. held in conjunction with ACL2000.

P. Vossen, editor. 1998. *EuroWordNet: a Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

P. Vossen. 2001. Extending, trimming and fusing WordNet for technical documents. In *Proc. of NAACL Workshop WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, June. held in conjunction with NAACL2001.

W.A. Woods. 1997. Conceptual indexing: A better way to organize knowledge. Technical report, SUN Technical Report TR-97-61.

Automatic Adaptation of WordNet to Domains

Roberto Navigli, Paola Velardi

Università di Roma "La Sapienza", Dipartimento di Scienze dell'Informazione, Via Salaria 113
00198 Roma, Italy, e-mail: velardi@dsi.uniroma1.it

Abstract

The objective of this paper is to present a method to automatically enrich WordNet with sub-trees of concepts in a given language domain. WordNet is then trimmed to reduce unnecessary ambiguity and singleton nodes. The process is based on the use of statistical method and linguistic processing to extract candidate domain *terms*. Multiword terms are semantically disambiguated and interpreted using ontological and contextual knowledge stored in WordNet on singleton words.

1. Introduction

As already pointed out by many researchers, WordNet is a very useful tool, but has some important drawbacks, namely, over-ambiguity and lack of domain terminology. Several published studies attempted to solve this problem in some automatic way, for example, (Vossen, 2001) (Harabagiu et al., 1999) (Milhalcea et al., 2001) and (Agirre et al. 1999). Other studies related to the work presented in this paper deal with the more general issue of automatic ontology construction. These contributions are collected in the web proceedings of two workshops dedicated to Ontology learning, (ECAI-OL, 2000) and (IJCAI-OL, 2001).

In many described approaches for ontology learning, domain terms are firstly extracted using a variety of statistical methods; then, taxonomic relations and other types of relations between terms are detected. In the literature, the notion of domain *term* and domain *concept* are used interchangeably, though no semantic interpretation of terms takes place. For example, in (Vossen, 2001) the "concept" *digital printing technology* is considered as a kind-of *printing technology* by virtue of simple string inclusion. However, *printing* has four senses in WordNet, and *technology* has two senses. There are hence 8 possible concept combinations for *printing technology*!

In this paper we propose a method for semantic interpretation of terms, using the information available in WordNet for the individual words that appear in a terminological string. Semantic interpretation allows us to detect non-trivial taxonomic relations between *concepts*, and other types of semantic relations.

The method described in this paper is implemented in a system called OntoLearn. OntoLearn is part of an Ontology Engineering architecture, described in (Missikoff et al., 2002), developed in the context of two European projects[1], aimed at improving interoperability in the Tourism sector.

Taxonomic information is extracted from the documents available in the considered domain in 5 steps: domain terminology is identified (section 2) and structured in syntactic trees (section 3), terms are mapped to concepts (section 4), that are arranged in a domain concept forest (section 5), and then used to create a domain-specific view of WordNet (section 6).

[1] ITS – 13015 (FETISH) and ITS- 29329 (HARMONISE).

2. Identification of Relevant Domain Terminology

The objective of this phase is to extract from the available documents a domain terminology. First, we use a linguistic processor, ARIOSTO[2], to extract from a corpus of documents a list of syntactically plausible terminological patterns, e.g. compounds (*credit card*), prepositional phrases (*board of directors*), adjective-noun relations (*manorial house*).

Then, two information theory based measures are used to filter out non-terminological (e.g. *last week*) and non-domain specific terms (e.g. *world wide web* in a Tourism domain). The first measure, called *Domain Relevance*, computes the probability of occurrence of a candidate term in the application domain (e.g. Tourism), as compared with other corpora that we use for a contrastive analysis (e.g. Medicine, Economy, Novels, etc.). The second measure, called *Domain Consensus*, computes the entropy of the probability of seeing a candidate term across the documents of the application domain. The underlying idea is that only terms that are *frequently* and *consistently* referred in the available domain documents reflect some *consensus* on the use of that term. These two measures have been formally defined and extensively evaluated in (Velardi et al, 2001).

3. Generation of Syntactic Trees

From the list of filtered terminology we generate *lexicalized trees*, on the basis of a simple inclusion relation. For example, given two strings x and wx (e.g. *telephone service* and *service*), we generate $wx \rightarrow^{@} x$, where '$\rightarrow^{@}$' stands for the hyperonymy relation. Figure 1 provides an example of a generated lexicalized tree $\mathfrak{I}$. It is clear that many taxonomic relations are not captured by this simple inclusion mechanism, like *bus service* $\rightarrow^{@}$ *public transport service*.

4. Semantic Disambiguation of Terms

The process of *semantic interpretation* is one that associates to each multiword term $t = w_n \cdot \ldots \cdot w_2 \cdot w_1$ (where w_i is an atomic word) the appropriate *concept name*.

[2] ARIOSTO is a joint effort of the Universities of Roma "La Sapienza" and "Tor Vergata".

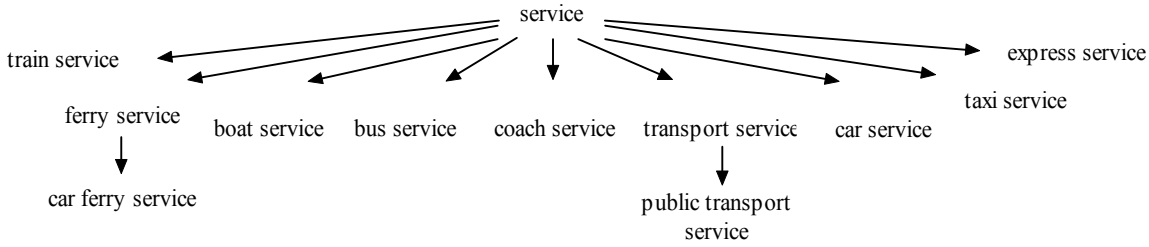


Figure 1. Example of a lexicalized tree.

Though complex terms are usually absent in WordNet, singleton words and occasionally word pairs included in a terminological string are mostly present. For example, *printing technology* as a unique term is not included, but *printing* and *technology* have an associated WordNet entry.

We derive the meaning of a complex terminological string *compositionally*, as explained hereafter.

Formally, a *semantic interpretation* is defined as follows: let $t = w_n \cdot \ldots \cdot w_2 \cdot w_1$ be a valid term belonging to a lexicalized tree $\mathfrak{I}$. The process of *semantic interpretation* is one that associates to each word w_k in t the appropriate WordNet synset S_i^k, the i-th synset ($i \in \{1,...,m\}$) associated to w_k in WordNet. The *sense* of t is hence defined as:

$$S(t) = \bigcup_k S^k,\ S^k \in Synsets(w_k) \text{ and } w_k \in t.$$

where $Synsets(w_k)$ is the set of synsets each representing a sense of the word w_k.

For instance: S("transport company") = { { transportation#4, shipping#1, transport#3 }, { company#1 } } corresponding to sense #1 of *company* ("*an institution created to conduct business*") and sense #3 of *transport* ("*the commercial enterprise of transporting goods and material*").

In order to disambiguate the words in a term $t = w_n \cdot \ldots \cdot w_2 \cdot w_1$ we proceed as follows:

a) If t is the first analyzed element of $\mathfrak{I}$, manually disambiguate the root node (w_1 if t is a compound) of $\mathfrak{I}$.

b) For any $w_k \in t$ and any synset S_i^k of w_k, create a *semantic net SN*. Semantic nets are automatically created using the following semantic relations: hyperonymy ($\rightarrow^{@}$), hyponymy ($\rightarrow^{\sim}$), meronymy ($\rightarrow^{\#}$), holonymy ($\rightarrow^{\%}$), pertonymy ($\rightarrow^{\backslash}$), attribute ($\rightarrow^{=}$), similarity ($\rightarrow^{\&}$), gloss ($\rightarrow^{gloss}$) and topic ($\rightarrow^{topic}$). The *gloss* and the *topic* relation are obtained parsing with ARIOSTO the WordNet concept definitions (*glosses*) and SemCor sentences (*topic*) including that sense. Every other relation is directly extracted from WordNet. To reduce the dimension of a SN, concepts at a distance of more than 3 relations from the SN centre, S_i^k, are removed. Figure 2a is an example of SN generated for sense #1 of *room*.

Let then $SN(S_i^k)$ be the semantic network for sense i of word w_k.

c) Starting from the "head" w_1 of t, and for any pair of words w_{k+1} and w_k (k=1,…,n-1) belonging to t, intersect alternative pairs of SNs. Let $I = SN(S_i^{k+1}) \cap SN(S_j^k)$ be one of such intersections for sense i of word w_{k+1} and sense j of word w_k. Note that, in each step k, the word w_k is already disambiguated, either manually (for k=1) or as a result of step k-1.

To identify common semantic patterns several heuristic rules are used, e.g.:

$$\exists G, M \in Synset_{wn} : S_1 \xrightarrow{gloss} G \xrightarrow{@^{\leq 3}} M \xleftarrow{@^{\leq 3}} S_2$$

The heuristic (named "gloss+parallelism") reads: "*given two central concepts S_1 and S_2, there exist two concepts G and M such that G appears in the gloss of S_1 and both G and S_2 reach the concept M in $SN(S_1) \cap SN(S_2)$ through a hyperonimy path.*

An example is the bold pattern in Figure 2b:

transport#3 $\xrightarrow{gloss}$ *enterprise*#2 $\xrightarrow{@^1}$ *organization*#1 $\xleftarrow{@^2}$ *company*#1.

5. Creating a Domain Concept Forest

Initially, all the terms in a tree $\mathfrak{I}$ are independently disambiguated. Subsequently, taxonomic information in WordNet is used to detect *is-a* relations between *concepts*, e.g. *ferry service* $\rightarrow^{@}$ *boat service*. In this phase, since all the elements in $\mathfrak{I}$ are jointly considered, some interpretation errors produced in the previous disambiguation step are corrected. In addition, certain concepts are *fused* in a unique concept name on the basis of pertonimy, similarity and synonymy relations (e.g. respectively: *manor house* and *manorial house*, *expert guide* and *skilled guide*, *bus service* and *coach service*).

Notice again that we detect semantic relations between *concepts*, not words. For example, *bus#1* and *coach#5* are synonyms, but this relation does not hold for other senses of these two words. Each lexicalized tree $\mathfrak{I}$ is finally transformed in a *domain concept* tree Υ.

Figure 3 shows the concept tree obtained from the lexicalized tree of Figure 1.

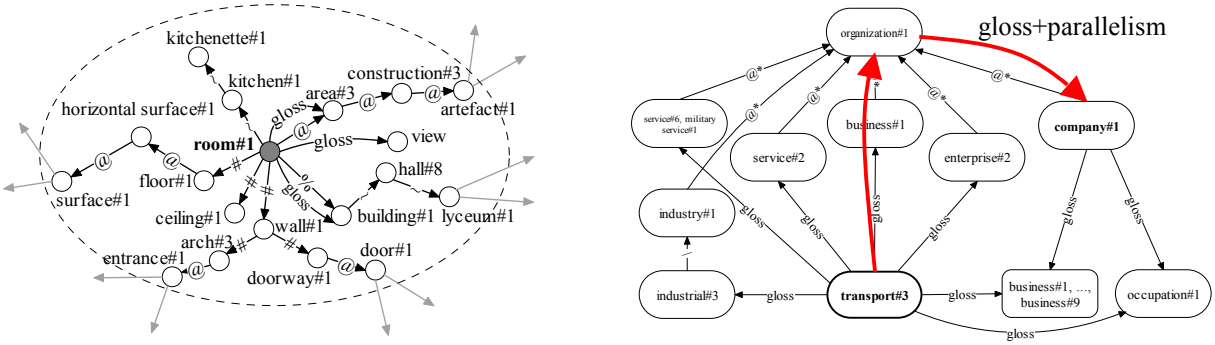


Figure 2. a) example of semantic net for *room#1*; b) example of intersecting semantic patterns for transport#3 and *company#1*.

For clarity, in Figure 3 concepts are labeled with the associated terms (rather than with synsets), and numbers are shown only when more than one semantic interpretation holds for a term, as for *coach service* and *bus service* (e.g. sense #3 of "bus" refers to "old cars").

6. Pruning and Trimming WordNet

The final phase consists in creating a domain-specialization of WordNet. In short, WordNet pruning and trimming is accomplished as follows:

1. The Domain Concept trees are attached under the appropriate nodes in WordNet.
2. An intermediate node in WordNet is pruned whenever the following conditions hold together

   i. it has no "brother" nodes;
   ii. it has only one direct hyponym;
   iii. it is not the root of a Domain Concept tree;
   iv. it is not at a distance ≤ 2 from a WordNet *unique beginner* (this is to preserve a "minimal" top ontology).

Figure 4 shows an example of pruning the nodes located over the Domain Concept tree with root *wine#1*. Appendix A shows an example of domain-adapted branch of WordNet in the tourism domain.

7. Evaluation

OntoLearn is a knowledge extraction system aimed at improving human productivity in the time-consuming task of building a domain ontology. Our experience in building a tourism ontology for the European project Harmonise reveals that, after one year of ontology engineering activities, the tourism experts were able to release the most general layer of the tourism ontology, comprising about 300 concepts.

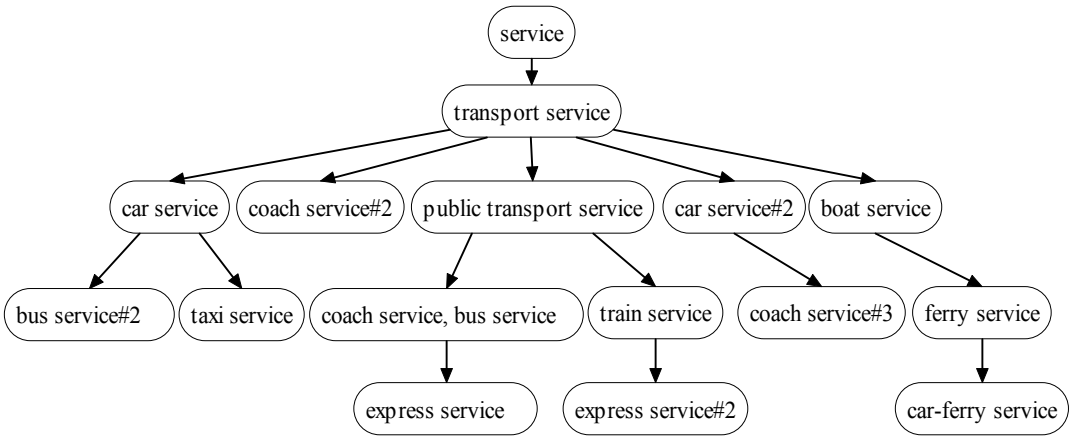


Figure 3. A Domain Concept Tree.

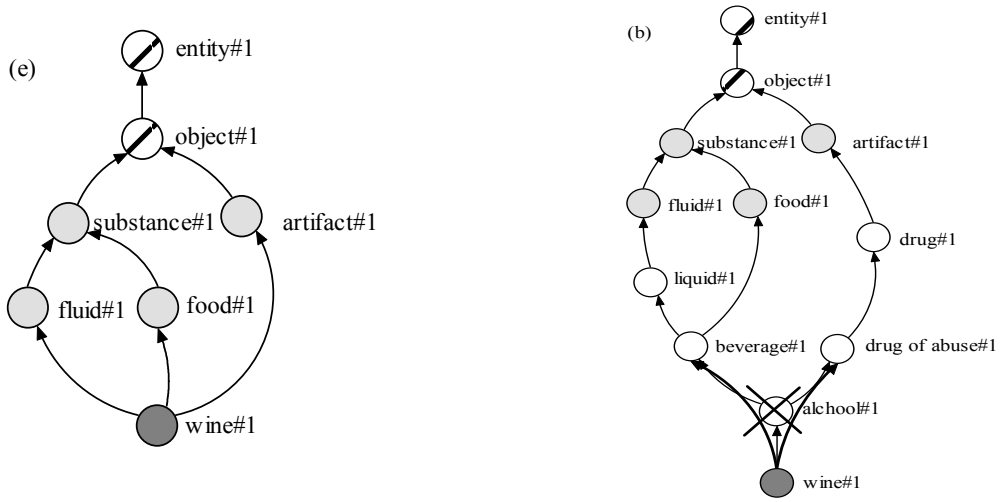


Figure 4. An intermediate step and the final pruning step over the Domain Concept Tree for "wine#1".

Then, we decided to speed up the process developing the *OntoLearn* system, aimed at supporting the ontology engineering tasks. This produced a significant acceleration in ontology building, since in the next 6 months[3] the tourism ontology reached about 3,000 concepts.

The OntoLearn system has been also evaluated independently from the ontology engineering process. We extracted from a 1 million-word corpus of travel descriptions (downloaded from Tourism web sites) a terminology of 3840 terms, manually evaluated[4] by domain experts participating in the Harmonise project. We obtained a precision ranging from 72.9% to about 80% and a recall of 52.74%. The precision shift is motivated by the well-known fact that the intuition of experts may significantly differ.

After this expert evaluation, we added few *ad hoc* heuristics that brought the precision to 97%. However, the use of heuristics limits the generality of the method.

The recall has been estimated by submitting a list of 6000 syntactic candidates to the experts, requiring them to mark truly terminological entries, and then comparing this list with that obtained by our statistical filtering method described in section 2.

We personally evaluated the semantic disambiguation algorithm using a test bed of about 650 extracted terms, which have been manually assigned to the appropriate WordNet concepts. These terms contributed to the creation of 90 syntactic trees. The entire process of semantic disambiguation and creation of domain trees has been evaluated, leading to an overall 84.5% precision. The precision grows to about 89% for highly structured sub-trees, as those in Figure 3. In fact, the phase described in section 5 significantly contributes at eliminating disambiguation errors (in the average, 5% improvement). We also analyzed the individual contribution of each of the heuristics mentioned in section 4 to the performance of the method, but a detailed performance report is omitted here for sake of space. The results of this performance analysis led to a refinement of the algorithm and the elimination of one heuristic.

[3] The time span includes also the effort needed to test and tune OntoLearn. Manual verification of automatically acquired domain concepts actually required few days.

[4] Here manual evaluation is simply deciding whether an extracted term is relevant, or not, for the tourism domain.

8. References

Agirre E., Ansa O., Hovy E. and Martinez D. *Enriching very large ontologies using the WWW*, in (ECAI-OL 2000).

Harabagiu S., Moldovan D. *Enriching the WordNet Taxonomy with Contextual Knowledge Acquired from Text*. AAAI/MIT Press, 1999.

Milhalcea R., Moldovan D. I. *eXtended WordNet: progress report.* NAACL 2001 Workshop, Pittsburg, June 2001.

Missikoff M., Velardi P. and Fabriani P. *Using Text Processing Techniques to Automatically enrich a Domain Ontology*. Proc. of ACM Conf. On Formal Ontologies and Information Systems, ACM_FOIS, Ogunquit, Maine, October 2002.

Velardi P., Missikoff M. and Basili R. *Identification of relevant terms to support the construction of Domain Ontologies*. ACL-EACL Workshop on Human Language Technologies, Toulouse, France, July 2001.

Vossen P. *Extending, Trimming and Fusing WordNet for technical Documents*, NAACL 2001 workshop on WordNet and Other Lexical Resources, Pittsbourgh, July 2001.

ECAI 2000, workshop on Ontology Learning http://ol2000.aifb.uni-karlsruhe.de/

IJCAI 2001, workshop on Ontology Learning http://ol2001.aifb.uni-karlsruhe.de/

Appendix A: A fragment of trimmed WordNet for the Tourism domain

{ activity%1 }
{ work%1 }
{ project:00508925%n }
{ tourism_project:00193473%n }
{ ambitious_project:00711113%a }
{ service:00379388%n }
{ travel_service:00191846%n }
{ air_service#2:00202658%n }
{ air_service#4:00194802%n }
{ transport_service:00716041%n }
{ ferry_service#2:00717167%n }
{ express_service#3:00716943%n }
{ exchange_service:02413424%n }
{ guide_service:04840928%n }
{ restaurant_service:03233732%n }
{ rail_service:03207559%n }
{ maid_service:07387889%n }
{ laundry_service:02911395%n }
{ customer_service:07197309%n }
{ guest_service:07304921%n }
{ regular_service#2:07525988%n }
{ outstanding_customer_service:02232741%a }
{ tourism_service:00193473%n }
{ waiter_service:07671545%n }
{ regular_service:02255650%a,scheduled_service:02255439%a }
{ personalized_service:01703424%a,personal_service:01702632%a }
{ secretarial_service:02601509%a }
{ religious_service:02721678%a }
{ church_service:00666912%n }
{ various_service:00462055%a }
{ helpful_service:02376874%a }
{ quality_service:03714294%n }
{ air_service#3:03716758%n }
{ room_service:03250788%n }
{ car_service#3:02384960%n }
{ car_service#4:02385109%n }
{ car_service#5:02364995%n }
{ hour_room_service:10938063%n }
{ transport_service#2:02495376%n }
{ car_service:02383458%n }
{ bus_service#2:02356871%n }
{ taxi_service:02361877%n }
{ coach_service#2:02459686%n }
{ public_transport_service:03184373%n }
{ bus_service:02356526%n,coach_service:02356526%n }
{ express_service#2:02653414%n }
{ local_bus_service:01056664%a }
{ train_service:03528724%n }
{ express_service:02653278%n }
{ car_service#2:02384604%n }
{ coach_service#3:03092927%n }
{ boat_service:02304226%n }
{ ferry_service:02671945%n }
{ car-ferry_service:02388365%n }
{ air_service:05270417%n }
{ support_service:05272723%n }

Extraction of Implicit Knowledge from WordNet

Wim Peters

NLP Group
Department of Computer Science
University of Sheffield
Regent Court
211 Portobello Street
Sheffield S1 4DP
U.K.
wim@dcs.shef.ac.uk

Abstract

Lexical knowledge databases such as WordNet contain much semantic information that is left implicit. In order to make maximal use of these resources it is important to make this implicit semantic information explicit. Metonymy and regular polysemy constitute a type of implicit ontological knowledge. This paper describes the semi-automatic extraction of systematically related word senses from WordNet by exploiting its hierarchical structure, and the identification of relations that link these on the basis of the glosses.

1. Introduction

WordNet (Fellbaum 1998) contains far more semantic information than its ontological organization shows. Word senses are related to senses of other words by means of a small number of basic semantic relations such as synonymy and hypernymy. Other types of encyclopaedic knowledge and semantic relations are implicitly present in the structure of WordNet in the form of taxonomic correspondences and glosses. This non-formalized semantic information in WordNet can be processed in order to distil more implicit knowledge (see e.g. Harabagiu 2000).

2. Relations between senses

Systematic relatedness between senses is one type of knowledge that is mostly left implicit in resources. This phenomenon is called metonymy, or, ore specifically, regular polysemy (Apresjan 1973).

Viewed traditionally, metonymy is a non-literal figure of speech in which the name of one thing is substituted for that of another related to it. It has been described as a cognitive process in which one conceptual entity, the vehicle, provides mental access to another conceptual entity (Radden 1999). In its basic form, it establishes a semantic relation between two concepts that are associated with word forms. The semantic shift expressed by the relation may or may not be accompanied by a shift in form. The semantic relation that is captured by metonymy is one of semantic contiguity, in the sense that in many cases there are systematic relations between metonymically related concepts that can be regarded as slots in conceptual frames (cf. Fillmore 1977).

Regular polysemy is a more specific instantiation of metonymy that covers the systematicity of the semantic relations involved. It can be defined as a subset of metonymically related senses of the same word displaying a conventional as opposed to novel type of semantic contiguity relation. Any systematic semantic relations between concepts are lexicalized, i.e. they are explicitly listed in dictionaries and independent of a pragmatic situation. For example, *The White House* is on the one hand an institution and on the other a building. The semantic relation between the two senses is 'is housed in'. It is a conventional pattern, not a nonce formation (a pragmatically defined novel metonymy), because it holds for related senses of two or more words (Apresjan, 1973) in the lexicon. It is this subtype of metonymy that we concentrate on in this paper.

3 Extraction from WordNet

A technique was developed (Peters 2000) for identifying sense combinations in WordNet where the senses involved potentially display a regular polysemic relation, i.e. where the senses involved are candidates for systematic relatedness.

In order to obtain these candidate patterns WordNet (WN) has been automatically analysed by exploiting its hierarchical structure for nouns. Wherever there are two or more nouns with senses in one part of the hierarchy, which also have senses in another part of the hierarchy, then we have a candidate pattern of regular polysemy. The patterns are candidates because there seems to be an observed regularity for two or more words. An example can be found in Figure 1 below.

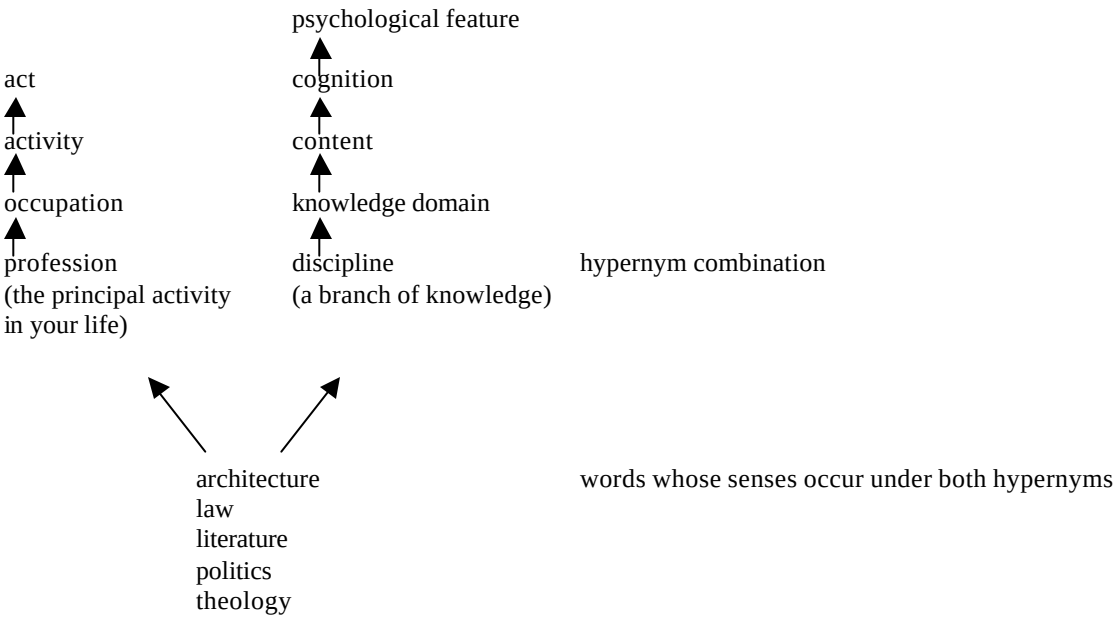


Figure 1: words in WordNet covered by the pattern profession/discipline

4 Relations

The results obtained from the manual analysis of reduced data sets according to (Peters 2001) and (Peters 2002) yields a set Regular Polysemic patterns. These patterns consist of combinations of the hypernymic nodes that subsume the words involved in the pattern. These combinations do not give any explicit information about the nature of the systematic relations that exist between them. This relationship can be determined by means of manual evaluation. The examination of the pair and the participating word senses will provide a human assessor with enough information to intuitively postulate a relationship. However, this is a costly and time consuming activity.

We have, up to a certain extent, automated this process of extracting explicit relations between the word senses involved in the regular polysemic pattern.

Our extraction process concentrates on the linguistic information available in the glosses associated with the word senses subsumed by the hypernymic pairs. The relations we have extracted take the form of verbs that link pairs of concepts. In each of these pairs one member is subsumed by member one of the hypernym pair and the other by number two. The glosses were first preprocessed. Part of speech tags were added and nominal and verbal content words were lemmatized.

For all nouns participating in the regular polysemic patterns listed above two bags of WordNet words were created, each associated with a sense captured by the regular polysemic pattern. The bag consisted of the noun under consideration, its synonyms and all the members of the hypernymic synsets. Then the words in the bag of the first word sense were matched against the processed gloss associated with the synset to which this sense belongs (henceforth synset 1).

If there was a match, the words from the bag of the second word sense (henceforth synset 2) were matched against the gloss.

If there was a match and the word from the synset 1 bag (word 1) preceded the word from the synset 2 bag (word 2) within the gloss, the text between the matches was extracted. If this span of text contained a verb, it was extracted, together with any associated prepositions. A distance of three positions between the matched nouns and the verb was applied in order to reduce spurious matches. Any extracted verb is considered to represent an instantiation of the relation(s) holding for the regular polysemic pattern.

The same matching process was repeated for the glosses associated with all hypernyms of synset 1. Then the whole process was repeated, looking for matches in the synset 2 gloss and all its hypernyms. Figure 2 below gives a graphical representation of the process.

The requirement that word 1 precedes word 2 is geared towards the extraction of transitive and prepositional verbs, both used in active form. The order constraint also determines the directionality of the relation, i.e. which hypernymic pair member is the subject and which is the object of an extracted verb.

We will illustrate this by means of an example.

The regular polysemic pattern **animal - food** is applicable to 172 words in WordNet. One of these words is '*herring*':

Sense 1: commercially important food fish of northern waters of both Atlantic and Pacific.

Sense 2: valuable flesh of fatty fish from shallow waters of northern Atlantic or Pacific; usually salted or pickled.

The bag of words associated with synset 1 contains 330 words (e.g. *fish, entity, life form, vertebrate, craniate*).

The synset 2 bag holds 518 words (*seafood, food, substance, food product, nutrient, object*). Only a subset of these words is related to *herring*, the rest are associated with the other words that are subsumed by the hypernymic pattern.
The concept 'food fish' is the hypernym of sense 1: "any fish used for food by human beings". Of the words in this gloss 'fish' is found in the synset1 bag and 'food' in the synset 2 bag. The intermediate text span is 'used for' which consists of a past participle and a preposition. The outcome is the relationship 'animal used for food'. This relation is found 37 times. The relation 'used for' is found 23 times.

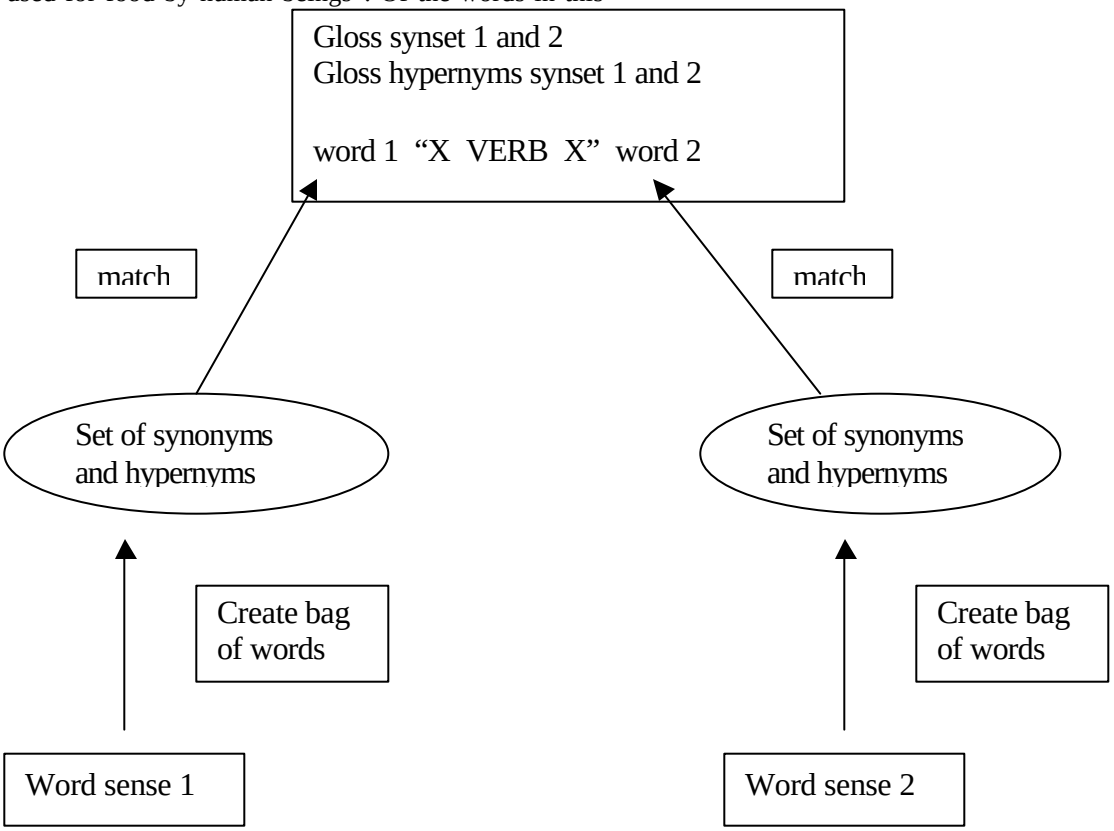


Figure 2: Mapping synset members onto glosses

The pattern **profession** and **discipline** (see figure 1) subsumes five words: *architecture, literature, politics, law* and *theology*.
Sense 6 of 'law' has the gloss 'the learned profession that is mastered by graduate study in a law school and that is responsible for the judicial system; "he studied law at Yale"'
Bag synset 1 contains 'profession', bag synset2 'study'. In between is the verb 'is mastered by' which yields the relation 'profession is mastered by discipline' for this regular polysemic pattern. This relation is found 2 times. One other relation was found: 'concerned with', which occurs only once.

Other relations are:

writing (reading matter; anything expressed in letters of the alphabet (especially when considered from the point of view of style and effect); "the writing in her novels is excellent")
message (what a communication that is about something is about)

This pattern covers 36 words. Examples are *account, conclusion, declaration, epitaph*. The relation 'express' occurs once, 'state' occurs 24 times.

fabric (something made by weaving or felting or knitting or crocheting natural or synthetic fibers)
covering (a natural object that covers or envelops)
This hypernymic combination subsumes 5 words: *fleece, hair, tapa, tappa, wool*.
'made from' occurs once.

person (a human being; "there was too much for one person to do")
language (a systematic means of communicating by the use of sounds or conventional symbols; "he taught foreign languages"; "the language introduced is standard throughout the text")
This pattern subsumes 257 words such as *Tatar, Assyrian, Hopi, Punjabi*.
The relation 'speak' occurs 132 times.

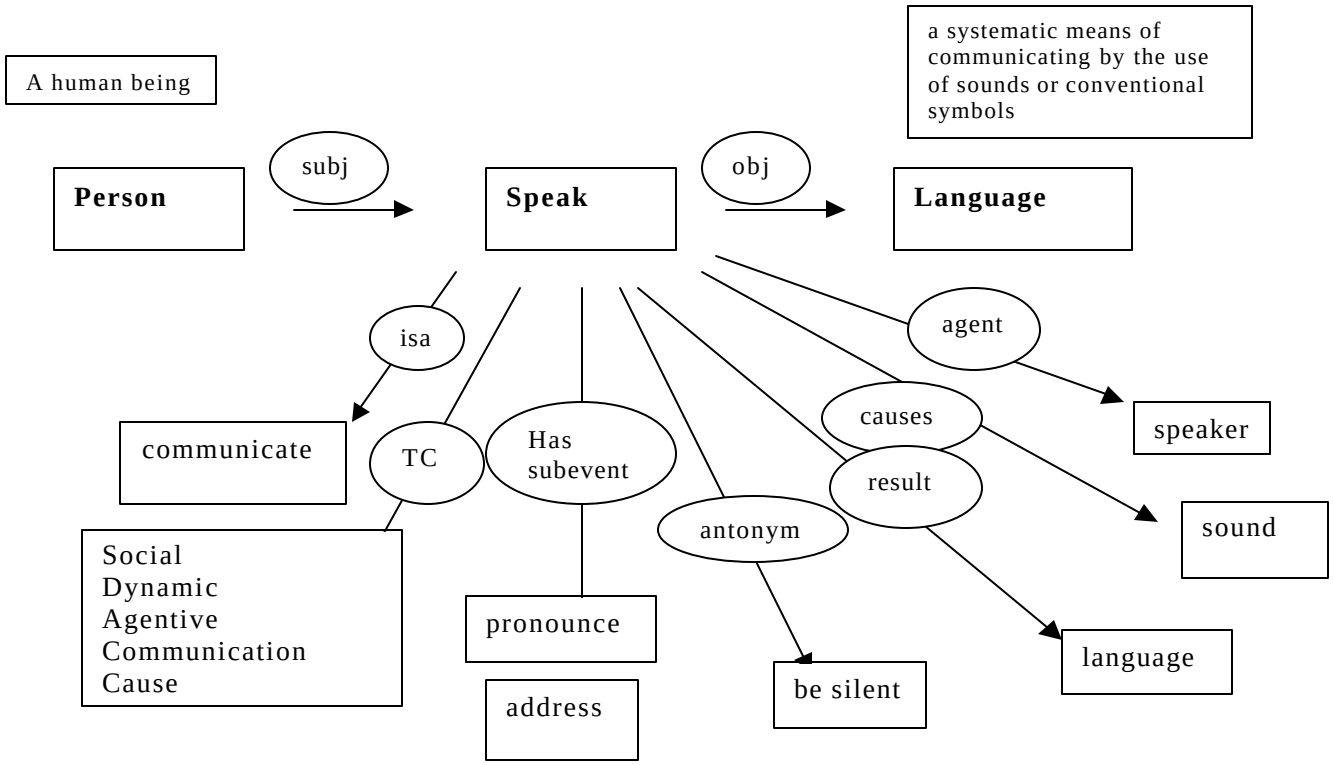


Figure 3: Expanded Ontological Fragment for the pattern person - speak - language

5 Expansion through EuroWordNet

Now we have obtained a number of patterns with specific relations it is possible to extend each ontological fragment consisting of concept triples (N-V-N) with explicit relations from EuroWordNet (Vossen, 1998). We have chosen this database over Wordnet because it contains more kinds of semantic relations than WordNet, such as thematic relations and links that hold between concepts lexicalized by different parts of speech.

First, the applicable verb senses was chosen manually. After that, relational chains in the database were extracted. Figure 3 and 4 exemplify this process for the verbs 'speak' associated with the pattern person - language and 'master' linking profession and discipline. The 'TC' relation indicates the EuroWordNet top concepts that are described in great detail in (Rodriguez et al., 1998).
The relations can all be considered additional slots in the partial knowledge frame that started as a regular polysemic pattern.
For instance, the additional knowledge fragments provided by EuroWordNet connect 'master' to 'knowledge', 'practice', 'learning' and 'teaching'. These can be used for inferencing purposes or knowledge extraction from texts.

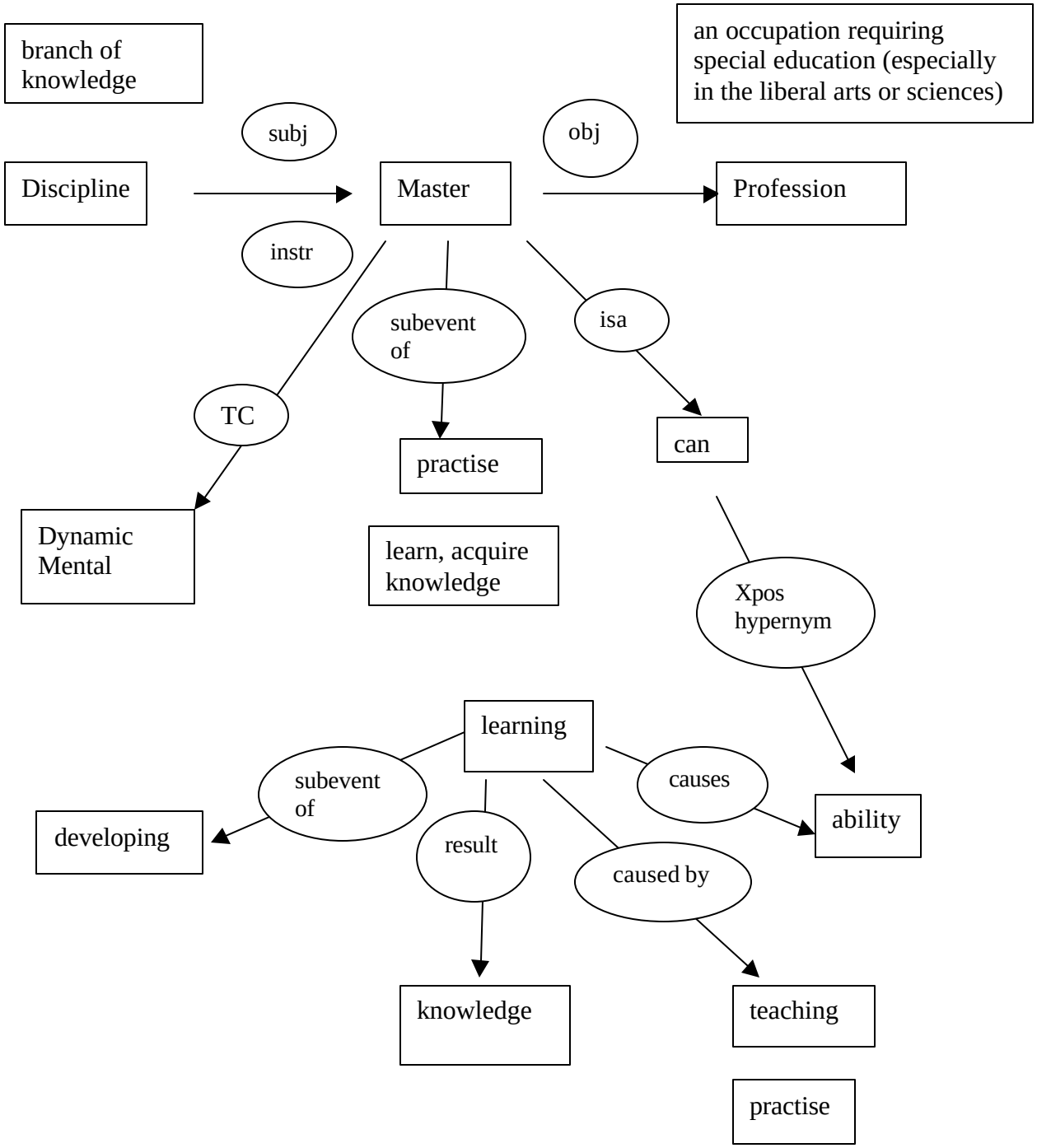


Figure 4: Expanded Ontological Fragment for the pattern **discipline - master - profession**

6 Discussion and conclusion

We have shown that the semi-automatic technique described above for extracting semantic relations between systematically related senses from WordNet glosses is feasible. There are cases, however, where no relations can be extracted, and where the extracted relations are wrong. Further experimentation with the syntactic properties of the glosses might improve results.

References

Apresjan, J. (1973), *Regular Polysemy*
In: Linguistics 142: 5-32

Fellbaum, Christiane (ed.) (1998)
WordNet: An Electronic Lexical Database.
Cambridge, Mass.: MIT Press.

Fillmore, C (1977)
Scenes and frames semantics.
In: Zampolli, A (ed.) Linguistic structures processing.
Amsterdam: Benjamins, 55-81.

Harabagiu, S. and Maiorano, S. (2000), *Acquisition of Linguistic Patterns for Knowledge-based Information Extraction*
LREC 2000, Athens

Peters, W. and Peters, I. (2000), *Lexicalised Systematic Polysemy in WordNet*
In *Proc. Secondt Intnl Conf on Language Resources and Evaluation*
Athens, Greece

Peters, W. and Wilks, Y. (2001), *Distribution-oriented Extension of WordNet's Ontological Framework,*
Proceedings RANLP2001, Tzigov Chark, Bulgaria

Peters, W., Guthrie, L. and Wilks, Y. (2002) , *Cross-linguistic Discovery of Semantic Regularity,*
Proceedings Global WordNet Association, Mysore, India

Radden, G. and Kövecses (1999), *Towards a Theory of metonymy*
In: Panther, K.U. and Radden, G. (eds.) Metonymy in language and Thought.
John Benjamins, Amsterdam

Rodriquez, H., Climent, S., Vossen, P. Bloksma, L., Roventini, A.,
Bertagna, F., Alonge, A., Peters, W. (1998), *The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology.*
In: Nancy Ide, Daniel Greenstein, Piek Vossen (eds), Special Issue on
EuroWordNet. Computers and the Humanities, Volume 32, Nos.
2-3 1998. 117-152.

Vossen, P. (1998), *Introduction to EuroWordNet.*
In: Nancy Ide, N., Greenstein, D. and Vossen, P. (eds), Special Issue on EuroWordNet. Computers and the Humanities, Volume 32, Nos. 2-3 1998. 73-89.

Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics

James Pustejovsky, Anna Rumshisky, José Castaño

Department of Computer Science, Brandeis University
Waltham, MA 02454
{jamesp, arum, jcastano}@cs.brandeis.edu

Abstract

In this paper, we discuss the utility and deficiencies of existing ontology resources for a number of language processing applications. We describe a technique for increasing the semantic type coverage of a specific ontology, the National Library of Medicine's UMLS, with the use of robust finite state methods used in conjunction with large-scale corpus analytics of the domain corpus. We call this technique "semantic rerendering" of the ontology. This research has been done in the context of Medstract, a joint Brandeis-Tufts effort aimed at developing tools for analyzing biomedical language (i.e., Medline), as well as creating targeted databases of bio-entities, biological relations, and pathway data for biological researchers (Pustejovsky et al., 2002). Motivating the current research is the need to have robust and reliable semantic typing of syntactic elements in the Medline corpus, in order to improve the overall performance of the information extraction applications mentioned above.

1. Introduction

Data mining and information extraction rely on a number of natural language tasks that require semantic typing; that is, the ability of an application to accurately determine the conceptual categories of syntactic constituents. Accurate semantic typing serves tasks such as relation extraction by improving anaphora resolution and entity identification. Domain-specific semantic typing also benefits statistical categorization and disambiguation techniques that require generalizations across semantic classes to make up for the sparsity of data. This applies, for example, to the problem of prepositional attachment, as well as identification of semantic relations between constituents within nominal compounds (see, for example, related discussion in Rosario & Hearst (2002)). Semantic typing has other direct applications, such as query reformulation, the filtering of results according to semantic type restrictions, and so on.

The set of categories used in semantic typing must be adequate enough to serve such tasks. In the biomedical domain, there are a number of efforts to develop specialized taxonomies and knowledge bases (UMLS, Gene Ontology, SWISS-PROT, OMIM, DIP). In this paper, we describe a method for adapting existing ontology resources for the natural language processing tasks and illustrate this technique on the National Library of Medicine's UMLS.

The UMLS, like many industry-standard taxonomies, contains a large number of word-concept pairings (over 1.5M typed terms), making it potentially attractive as a resource for semantic tagging information. However, these types are inadequate for NL tasks for two major reasons. First, the overall type structure is very shallow. For example, for the semantic tag "Amino Acid, Peptide, or Protein" (henceforth AAPP), there are 180,998 entries, for which there are dozens of functional subtypes that are routinely distinguished by biologists, but not in the UMLS.

One specific example of the type system deficiencies illustrates this point very clearly: the extraction of relations and their arguments from text is greatly improved with entity and anaphora resolution capabilities. However, entity and event anaphora resolution rely on (among other things) the semantic typing of the anaphor and its potential antecedents, particularly with sortal and event anaphora, as shown in (1) below.

(1) a. "For separation of nonpolar compounds, the prerun can be performed with *hexane*$_i$; ... The selection of *this solvent*$_i$ might be considered .."

b. [*p21*$_i$ inhibits the regulation of ...] ... [*This inhibitor*$_i$ binds to ...]

c. [A *phosphorylates*$_i$ B.] ... [*The phosphorylation*$_i$ of B ...]

Strict UMLS typing presents a problem for our anaphora resolution algorithm (Castaño et al., 2002). For example, for the case of anaphora in (1a), the UMLS Metathesaurus types *hexane* as either 'Organic Chemical' or 'Hazardous or Poisonous Substance'. However, *solvent* is typed as 'Indicator, Reagent, or Diagnostic Aid'. In the UMLS Semantic Network, these semantic types are not related. Therefore the resolution of the sortal anaphora would fail, due to the type mismatch. The fact is that hexane is a solvent, and this is simply not reflected in UMLS.

Functional subtyping is also missing, as (1b) illustrates. This example shows a known protein (p21) being subsequently referred to as an 'inhibitor' (a functional class of proteins). This type does not exist in UMLS and the noun 'inhibitor' is merely typed as 'Chemical Viewed Functionally', while p21 itself is typed as 'Gene or Genome', AAPP, or 'Biologically Active Substance'. It is therefore difficult to discriminate p21 from other proteins (as potential antecedents) for the sortal anaphor "this inhibitor".

A related difficulty is encountered with event anaphora cases such as (1c), where an event nominal anaphor binds to a tensed event as its antecedent, both of which are of different types in the UMLS. Hence, the existing UMLS system does not allow for recognition of type-subtype relations of the kinds that are needed in order to identify anaphoric bindings in Medline texts.

Given these motivations, we have developed a set of techniques for "rerendering" an existing semantic ontology to satisfy the requirements of specific features of a (set of) application(s). For the present case (i.e., the UMLS and bio-entity and relation extraction), we identify candidate subtypes for inclusion in the type system by two means: (a) corpus analysis on compound nominal phrases that express unique functional behavior of the compound head; (b) identification of functionally defined subtypes derived from bio-relation parsing and extraction from the corpus. The results of rerendering are evaluated for correctness against the original type system, and against additional taxonomies, should they exist, such as the GO ontology. In our preliminary experiments, we had domain experts partially verify it aganst the Gene Ontology. Full automatic verification will be done in the future.

2. Semantic Rerendering

Many NLP tasks in the service of information extraction can benefit from more accurate semantic typing of the syntactic constituents in the text. As mentioned above, the semantic taxonomy available from UMLS is lacking in several respects. With specific applications such as content summarization, anaphora resolution, and accurate relation identification in mind, we describe how an existing type system can be systematically adapted to better serve these needs, using a technique we call *semantic rerendering*. Semantic rerendering is a process that takes as input an existing type system (such as UMLS) and a text corpus, and proposes refinements to the taxonomy on the basis of two strategies:

- *Linguistic Rerendering*: Syntactic and semantic analysis of NP structures in the text;
- *Database Rerendering*: Analysis of "ad hoc abstractions" from a database of relations automatically derived from the corpus.

In the first strategy, we use the syntax of noun groups to identify candidate subtypes to an existing UMLS type. For example, categories that are of interest to biologists but which are not explicitly represented in the type system are functional categories such as *phosphorylators*, *receptors*, and *inhibitors*. These are each significant categories in their own right but also have a rich number of subtypes as well, as illustrated in (2) below.

(2)
```
CB(2) receptor
cannabinoid receptor
cell receptor
D1 dopamine receptor
epidermal growth factor receptor
functional GABAB receptor
gastrin receptororphan receptor
orphan nuclear receptor
major fibronectin receptor
mammalian skeletal muscle acetylcholine receptor
normal receptor
PTHrP receptor
protein-coupled receptor
ryanodine receptor
```

If individual proteins can be identified (i.e., semantically tagged) as belonging to a functionally defined class, such as *receptor*, then richer information extraction and textual binding is enabled.

There has been some recent research on extracting hyponym and other relations from corpora (Hearst, 1992; Pustejovsky et al., 1997; Campbell & Johnson, 1999; Mani, 2002). Our work extends the techniques described in (Pustejovsky et al., 1997) using more extensive corpus analytic techniques as developed in Pustejovsky & Hanks (2001).

2.1. Linguistic Rerendering

We first describe the linguistic rerendering procedure for inducing subtypes from corpus data, given an existing taxonomy such as the UMLS. We being by taking the strings classified as <*supertype*> in the current type system. On the basis of their behavior in the corpus, we identify candidate subtypes, derived from an analysis of the structure of nominal compounds and clusters. We use the RHHR (*righthand head rule*, cf. Pustejovsky et al. (1997)) for compound nominals (CN) and create subtype <*head-of-CN*> from the type of the head of CN. We then create a node for type N' and insert it into the existing UMLS hierarchy.

More explicitly, the procedure for identifying candidate subtypes from the structure of nominal compounds is given below.

(1) Acquire corpus C.

(2) Apply existing type system *UMLS*$_1$ over C:
TS-UMLS$_1(C) = C_{S-Tag}$.

(3) Select from the resulting semantically tagged corpus C_{S-Tag} all *NPs* with semantic tag A with $\theta > \delta$, where θ is a measure of how interesting semantic type is for rerendering:

(4) For a given noun *N* that is the headword of a phrase with semantic tag A, propose *N* as *name* of a subtype of *S-Tag A*, $N' \sqsubseteq A$, if:

- N appears as head in a certain number of NPs of length $l \geq 2$;
- N falls under the threshold set for the headwords above, but is an LCS (longest common subsequence) of a number of syntactic heads that achieve it when combined [1];
- there is sufficient variation in words comprising the remainder of phrase (so as to exclude using collocations as subtypes).

(We will refer to the nodes inserted into the ontology at this stage *first-level extension*)

(5) Nouns in the residue of *NP* with *N* as head α as modifier can be proposed as subtypes of $\alpha N' \sqsubseteq N'$ (*second-level extension*).

[1] E.g. For AAPP, *oxidase* might not achieve the threshold by itself. However, it does when all headwords containing it as a subsequence are combined (i.e.*myeloperoxidase, peroxidase, deepoxidase,* etc.)

Further subcategorization of induced types, based on the analysis of modifiers within the nominal phrases, uses a combination of template filtering of noun phrases and the LCS (longest common subsequence) algorithm (Cormen et al., 1990). Notice that one must use different thresholds for headwords and modifiers (in step (4) or step (5) of the algorithm). However, at step (4), a candidate subtype may replicate exactly the parent node ($receptor \sqsubseteq Receptor$). In that case, first-level extension types must derived from subphrase analysis, but using the threshold established for step (4).

Once the candidate subcategories are selected, the next step is to obtain the instances for the induced subtypes. These instances and their type bindings can be identified from the corpus using a number of standard methodologies developed in the field for the expansion of ontology coverage (Hearst, 1992; Campbell & Johnson, 1999; Mani, 2002). For the moment, in the experiments we conducted, we used syntactic pattern templates to identify the strings that map to the proposed extensions of UMLS types (see examples in Table 1 below).

This procedure will result in differential depth of UMLS extension for functionally defined vs. naming categories. For example no strings should map to $\{head, neck, arm, leg\} \sqsubseteq$ *<Body Location or Region>*, while string mappings are easily obtained for relational nouns such as $\{solvent, antibody, conjugate\} \sqsubseteq$ *<Indicator, Reagent, or Diagnostic Aid>*.

2.2. Database Rerendering

The second strategy uses a database of biological relations constructed through the application of robust natural language techniques as outlined in Pustejovsky et al. (2002) and Castaño et al. (2002). Over this database, "ad hoc" categories are created by projecting statistically thresholded arguments. More formally, for a particular relation, a typed projection is obtained:

$\pi X = \{X : T_1 | R(X, Y) \wedge T_1 \in UMLS_1\}$

R	X	Y
phosphorylate	"TNIK"	"Gelsolin"
phosphorylate	"GSK-3"	"NF-ATc4"
phosphorylate	"IKK-beta"	"IkappaB"
...	...	...
inhibit	"PD-ECGF"	"DNA synthesis"
inhibit	"BMP-7"	"terminal chondrocyte differentiation"
block	"DFMO"	"ODC activity"
abrogate	"Interleukin-4"	"hydrocortisone-induced apoptosis"
...	...	...

Table 2: A sample segment of relations database

The noun forms for such ad hoc categories are determined by checking each relation against the first-level extension subtypes derived through NP structure analysis as outlined above. Thus,

- For relation R and each subtype $N' \sqsubseteq T_1$, associate N' with πX if $Sim(N, \pi R) > \epsilon$.

e.g. $Sim($ "kinase", "phosphorylate" $)$,
$Sim($ "inhibitor", "inhibit" $)$, etc.

Note that the ad hoc category created through projection of the relation's argument can be matched with the types obtained at the second-level of NP-based ontology extension.

The similarity measure is constructed as a weighted combination of string similarity (e.g. *LCS*-based score), and an integrated composite measure derived from the training corpus and the outside knowledge sources. The latter might use standard IR similarity measures on contexts of occurrence of R and N in Medline abstracts, in definitions of R and N in domain-specific MRDs (such as the On-line Medical Dictionary), etc. Thus, we have:

$$Sim(N, \pi R) =$$
$$= z_0 * \text{ LCS-score } (N, \pi R) + \Sigma_{i=1}^{k} z_i * Sim_i(N, \pi R)$$

where $Sim_i(N, \pi R)$ is the similarity score derived from the source i, and z_i is the weight assigned to the source i.

3. Methodology

3.1. Seed Ontology

The Unified Medical Language System (UMLS) which was used as the seed ontology has three components: the UMLS Metathesaurus, the UMLS Semantic Network, and the SPECIALIST Lexicon (UMLS Knowledge Sources, 2001). The UMLS Metathesaurus maps single lexical items and complex nominal phrases into unique concept IDs (CUIs) which are then mapped to the semantic types from the UMLS Semantic Network. The latter type taxonomy is what was used in the experimental applications of rerendering procedure. It consists of 134 semantic types hierarchically arranged via the 'isa' relation and interlinked by a set of secondary non-hierarchical relations. UMLS Metathesaurus in the UMLS 2001 distribution contains over 1.5 million string mappings.

In the Metathesaurus, multiple semantic type bindings are specified for many of the concepts. Due to this ambiguity of UMLS concepts and to a lesser extent, the ambiguity of the strings themselves, the mappings obtained from the Metathesaurus give a number of semantic types for each lexical item or phrase. We intentionally avoid superimposing any disambiguation mechanism on this typing information while applying it in corpus analysis. Since corpus-based derivation of subtypes uses a frequency cutoff, this ambiguity essentially resolves itself. For example, if a given lexical item is typed as both T_1 and T_2 in the seed ontology, and occurs as a headword in $> 1\%$ of nominal phrases typed as T_1, but in $< 1\%$ of nominal phrases typed as T_2, it will only be proposed as a candidate subtype of T_1. Thus, under the 1% cutoff, *isozyme*, which the seed UMLS types as either $Enzyme$ or $AAPP$, will only be identified as a good candidate subtype for *Enzyme*.

3.2. Corpus preprocessing with UMLS types

The experimental application of the rerendering procedure was conducted on a relatively small corpus of Med-

Pattern Type	TEMPLATE	
apposition	"X, a Y inhibitor"	"X, the solvent
	"X, an inhibitor of Y"	"the solvent, X"
	"X, an inhibitor of Y"	"X, a common solvent for Y"
nominal compounds	"Y inhibitor"	"the solvent X"
definitional constructions	"X is an inhibitor of Y"	
aliasing constructions	"X (inhibitor of Y)"	"X (the solvent)"
	"an inhibitor of Y (X)"	"the solvent (X)"
enumeration	"Y inhibitors such as X, ..."	"solvents (e.g. X)"
		"solvents, e.g. X"
		"the following solvents: X, .."
relative clauses	"X which is an inhibitor of Y"	"the solvent used was X"
		"X proved to be a suitable solvent"
adjuncts		"in X and Y as solvents"
		"X as solvent"

Table 1: Sample syntactic patterns for string-to-semantic type mappings

line abstracts (around 40,000). Medline abstracts were tokenized, stemmed, and tagged. They were then shallow-parsed, with noun phrase coordination and limited prepositional attachment (only *of*-attachment) using finite-state techniques. The shallow parse was obtained using five separate automata each recognizing a distinct family of grammatical constructions, very much in the spirit of Hindle (1983), McDonald (1992) and Pustejovsky et al. (1997). The machinery used in preprocessing is described in more detail in Pustejovsky et al. (2002).

Semantic type assignment of the resulting nominal chunks is determined through lookup as follows. Each noun phrase is put through a cascade of hierarchically arranged type-assignment heuristics. Higher priority heuristics take absolute precedence; that is, if a semantic typing is possible, it is assigned. In this implementation, we use the full UMLS semantic type hierarchy, including the mappings to both leaves and intermediate nodes.

During direct lookup, a string is assigned a given semantic type if the UMLS Metathesaurus contains a mapping of that string to a concept so typed. If a semantic type for the whole phrase is not found in UMLS, we attempt to identify its syntactic head using a modification of RHHR (*righthand head rule*), and determine the semantic type of the headword. For chunks with *OF*-attachment, i.e. phrases of the form, *<NP-1> of <NP-2>*, the lookup is first attempted on *NP-1* as a whole.

If the lookup on a particular prospective head fails, it is tested for a match with morphological heuristics recognizing semantically vacant categories, such as 'NUMERIC', 'ABBREVIATION', 'SINGLE CAPITAL LETTER', 'SINGLE LOWER-CASE LETTER', etc. These, and phrases headed by common words occurring in a non-specialized dictionary are filtered out. The last layer of heuristics applied to a prospective syntactic head successively attempts to strip a groups of suffixes and prefixes and perform lookup on the remaining stem.

3.3. Inducing candidate subtypes

In these initial series of tests, we experimented primarily with the first part of the rerendering procedure as it is outlined in Section 2.1. In the first stage of identifying the subtypes based on the syntactic analysis of noun phrase structure, a headword was considered a candidate subtype of type T if it occured in more that 1% of all nominal chunks tagged as T. Note that the same chunk is frequently tagged with several UMLS types.

The candidate subtypes for the second (NP modifier-based) level of UMLS extension were identified using a combination of template and frequency-based filtering of noun phrases and the LCS (longest common subsequence) algorithm. Thus, for a given headword proposed as subtype at first level of extension (e.g., *kinase*) the LCS algorithm was run on all phrases with that headword that matched a certain template (e.g. <Indefinite Article> <Modifier>* N). The substrings that occurred in the corpus in more than a certain percentage of phrases with that headword were identified as candidate subtypes for insertion into the ontology at the next level. The cut-off threshold had to be kept very low for this series of experiments, as it was conducted over a relatively small corpus. In working with a larger corpus the thresholds are set separately for each template, so e.g. it is much higher for the unfiltered set of nominal compounds than for those occurring with an indefinite article. Frequency-based filtering involves discarding as potential candidates noun phrases with modifiers that occur frequently in separate non-specialized corpus, which allows to automatically discard phrases such as 'multiple receptors', 'specific kinase', etc. [2]

Identification of sample instances for the induced types was performed over shallow-parsed text using syntactic pattern templates. The definitional construction patterns were extracted using relation extraction machinery (see Pustejovsky et al. (2002) for details). It was applied to our test corpus and another sample set of Medline abstracts (approx. 60,000).

4. Results

Semantic typing over our sample set of Medline data produced type bindings for over 1 million noun phrases.

[2] Similar filtering was also applied to the first-level extensions

4.1. NP analysis-based subtypes

The choice of particular UMLS categories as supertypes for extension of the seed UMLS semantic type taxonomy is dictated by the particular natural language application. Semantic types given below are derived from nominal phrase analysis for some of the supertypes that have been used in anaphora resolution tasks (cf Castaño et al. (2002)). Each UMLS type is shown with the number of noun phrases of that type which occurred in our test corpus, followed by a list of derived candidate subtypes with their respective frequencies. The subtypes shown below were derived as described above in step 4 of the rerendering procedure specification in Section 2.1.

```
Enzyme 4724
      dehydrogenase 140
      protease 160
      reductase 73
      metalloproteinase 48
      isozyme 54
      oxidase 79
      phosphatase 111
      enzyme 1142
      kinase 741

Amino Acid, Peptide, or Protein 20830
      receptor 2444
      protein 4521
      peptide 947
      kinase 741
      cytokine 287
      isoform 412

Cell 16348
      macrophage 251
      clone 350
      neuron 1094
      lymphocyte 412
      fibroblast 257
      cell 11586

Cell Component 2508
      cytosol 84
      nucleus 469
      liposome 43
      organelle 40
      vacuole 35
      ribosome 28
      cytoskeleton 55
      dendrite 53
      cytoplasm 195
      soma 26
      granule 80
      chromatin 36
      microtubule 45
      chromosome 319
      axon 99
      microsome 132
```

Notice that the categories derived in this manner would include functionally defined types (e.g. *isoform*).

4.2. NP modifier-based extension (second-level)

As mentioned above, some of the UMLS extension candidates that are derived according to the procedure are replicas of the supertype category, e.g. $enzyme \sqsubseteq Enzyme$, or $receptor \sqsubseteq Receptor$. For example, among the lexical items tagged as *Receptor* in UMLS Metathesaurus, NPs headed by the word "receptor" comprise 87% of all NPs tagged as *Receptor* in our test corpus:

```
Receptor 2820
      integrin 91
      receptor 2444
```

The appropriate extensions to the comparable level within the type taxonomy in this case are derived from subphrase analysis. Thus, for the case of $enzyme$, the candidate subtypes so derived would be:

```
cytosolic enzyme
heterologous enzyme
male enzyme
metalloenzyme
multifunctional enzyme
proof-reading enzyme
proteolytic enzyme
rate-limiting enzyme
recombinant enzyme
rotary enzyme
tetrameric enzyme
```

These are identified at step 5 of rerendering procedure through a combination of template filtering of noun phrases and longest common substring identification. They are then added to the same level of the type taxonomy as all $N' \sqsubseteq Enzyme$ (see Figure 1).

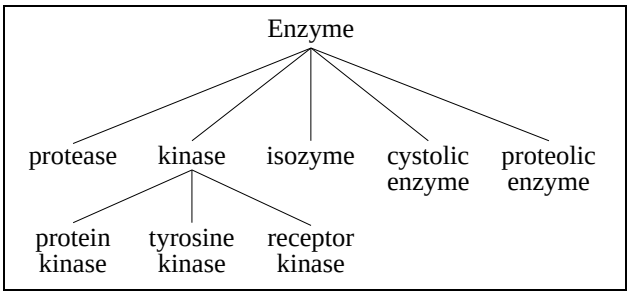


Figure 1: Extension subtree for $Enzyme$ (partial)

The results produced at this stage by the automated processing described above need further filtering before good subtype candidates can be identified. This can be achieved by fine-tuning the use of corpus frequencies, as well as type filtering of modifiers using the seed ontology type system. Table 3 below shows UMLS types for selected NP modifier-based subcategories of $receptor$.

4.3. Corpus-based identification of the instances of induced semantic categories

The rerendering procedure gives different results for different segments of the taxonomy, depending on the class of supertype category. Thus, for functionally defined semantic types, such as, "Chemical Viewed Functionally", or "Indicator, Reagent, or Diagnostic Aid", corpus-based derivation of instances for the induced subcategories is clearly much more feasible. Consider the first level extension types for the categories below:

```
Indicator, Reagent, or Diagnostic Aid 3424
      buffer 151
      conjugate 112
      stain 75
      agar 38
      antibody 1640
      indicator 373
      solvent 38
      tracer 53
      dye 95
      reagent 113
      nitroprusside 51
      hydrogen peroxide 58
```

Candidate Subtypes $\alpha N' \sqsubseteq N'$	**Seed UMLS Type for Modifier** α
cell surface receptor	'Cell Component'
membrane receptor	'Tissue'
adhesion receptor	'Acquired Abnormality', 'Disease or Syndrome'
	'Natural Phenomenon or Process'
activation receptor	no type binding
contraction receptor	'Functional Concept'
estrogen receptor	'Steroid', 'Pharmacologic Substance', 'Hormone'
dopamine receptor	'Organic Chemical', 'Pharmacologic Substance',
	'Neuroreactive Substance or Biogenic Amine'
adenosine receptors	'Nucleic Acid, Nucleoside, or Nucleotide',
	'Pharmacologic Substance', 'Biologically Active Substance'
insulin receptor	'Amino Acid, Peptide, or Protein',
	'Pharmacologic Substance', 'Hormone'
TSH receptor	'Amino Acid, Peptide, or Protein', 'Hormone'
	'Neuroreactive Substance or Biogenic Amine'
EGF receptor	'Amino Acid, Peptide, or Protein', 'Hormone',
	'Pharmacologic Substance',
transferrin receptor	'Amino Acid, Peptide, or Protein', 'Biologically Active Substance',
	'Indicator, Reagent, or Diagnostic Aid', 'Laboratory Procedure'
receptor	**'Amino Acid, Peptide, or Protein', 'Receptor'**

Table 3: UMLS Typing of modifiers α for some sample subtypes $\alpha N' \sqsubseteq N'$ for N' =*receptor*

```
Chemical Viewed Functionally 3494
      inhibitor 1668
      prodrug 62
      basis 1075
      vehicle 107
      radical 144
      base 265
      pigment 36
      surfactant 36

Pathologic Function 17752
      impairment 383
      stenosis 274
      other 450
      illness 209
      problem 1133
      dysfunction 493
      block 244
      carrier 219
      inflammation 243
      pathogenesis 497
      cavity 273
      hemorrhage 180
      occlusion 266
      lesion 1820
      infarction 449
      regression 237
      pathology 242
      infection 1782
      complication 1248
      separation 320
      degeneration 180
      stress 487
```

Table 4 shows the derivation of instances for the categories induced through noun phrase analysis (step 5), using the definitional construction template. The first column shows the actual strings that get the new type binding as $kinase$ (in blue) and their original UMLS types (in black). Notice that for many of the strings that can be so typed, the seed UMLS type is either generic $AAPP$ or the type binding is absent altogether.

If the candidate subtype is a valid semantic category, such corpus-based identification of instances should work equally well irrespective of the level at at which the induced type is inserted. For example, see Table 5 below for NP modifier extensions of *receptor*.

cell-surface receptors:
polycystin-1 is a *cell surface receptor*
Fas is a *cell surface* death $receptor$
CD40 is a *cell surface receptor*
CD44 is a *cell surface receptor*
The scavenger receptor BI is a *cell surface* lipoprotein $receptor$
membrane receptors:
Neuropilin-1 is a trans*membrane receptor*
APJ is a seven trans$membrane$ domain G-protein-coupled $receptor$
HER2 is a *membrane receptor*

Table 5: Sample semantic type instances derived with the definitional construction template for subtypes of *receptor*

5. Evaluation of Rerendering Procedure

The evaluation of the performance for rerendering essentially boils down to whatever improvement is produced in precision and recall for the client applications. However, in order to do an earnest evaluation of performance of the rerendering algorithm, we would need to run it on a much larger corpus. This would allow for better candidate choices for the portions of the procedure that have been plagued by sparsity (e.g., in NP modifier-based candidate subtype selection). But most importantly, it would increase the coverage in terms of instances for which the type bindings are produced in the new type system.

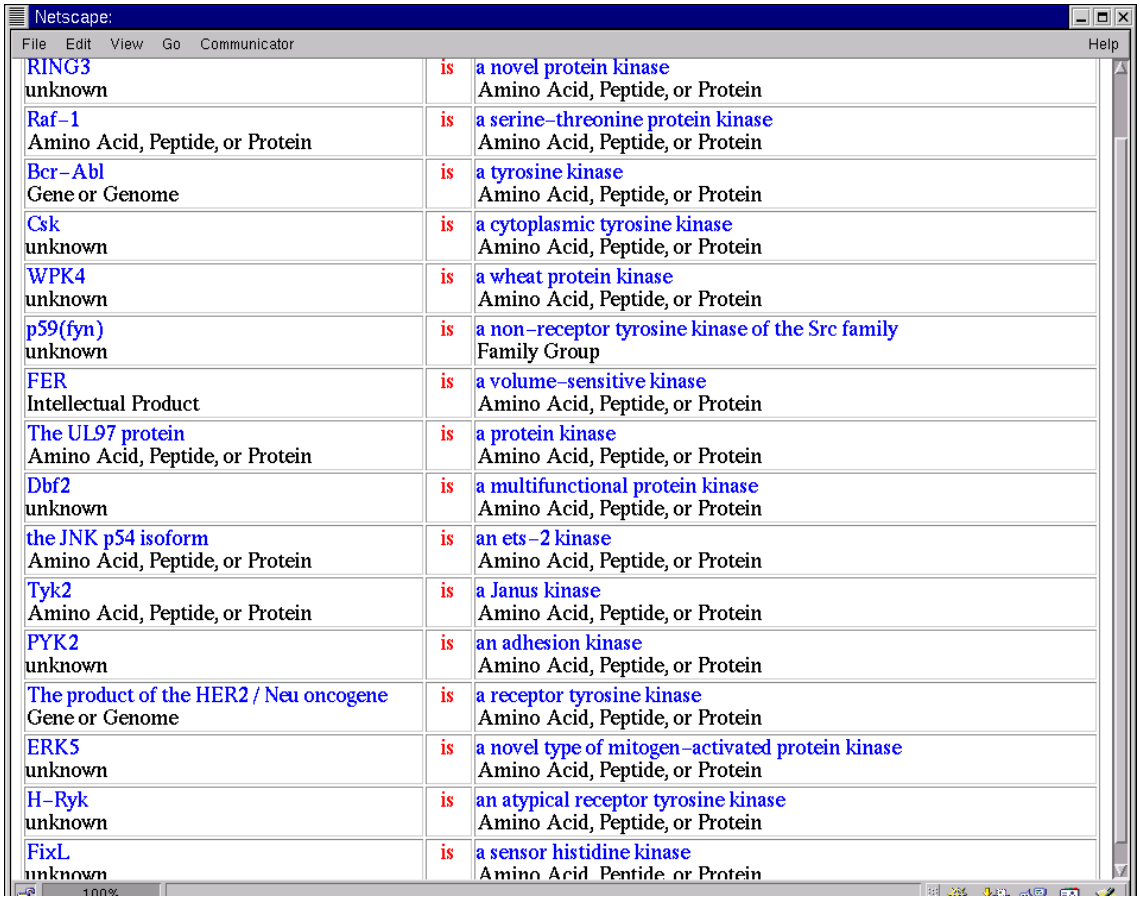


Table 4: Definitional construction template at work for the $N' = kinase$

5.1. Usability in natural language applications

One of the client applications for the experiments we report here is coreference resolution. The anaphora examples in (3) below illustrate the impact of using the derived types. Even the test corpus we used actually contained enough information to produce the type bindings for some of the strings used in (3).

(3) a. "Assays were conducted in *chloroform, toluene, amyl acetate, isopropyl ether, and butanol*. ... In *each solvent*,"

b. "The extracts were prepared separately in *methanol, ethanol, phosphate buffer saline (PBS), and distilled water* as part of our study to look at ... Our results have shown that *all four solvents* were ..."

c. "A 47-year-old man was found dead in a factory where *dichloromethane (DCM)* tanks were stocked. He was making an inventory of t he annual stock of DCM contained in several tanks (5- to 8000-L capacity) by transferring *the solvent* into an additional tank with the help of compressed air."

(emphasis added)

The seed ontology induces a type mismatch between the anaphor and the antecedent. For example, in (3c), the original type bindings are:

- *TS-UMLS*$_1$(solvent)= 'Indicator, Reagent, or Diagnostic Aid';
- *TS-UMLS*$_1$(dichloromethane)= {'Organic Chemical', 'Pharmacologic Substance', 'Injury or Poisoning'}

The rerendered ontology allows the induced semantic type *solvent* $\sqsubseteq$*<Indicator, Reagent, or Diagnostic Aid>* to be included in the type bindings for "dicloromethane".

5.2. Evaluation against existing ontologies

We performed some test evaluations of the second-level extension subtypes against the Gene Ontology. Despite the very modest side of our test corpus, we observed significant overlap in some categories. Thus, for example, the 388 second-level extension subtype candidates for *receptor*, 12% were identified as concept names in the Gene Ontology.

In general, the preliminary results of applying the first step of the rerendering procedure algorithm to the UMLS semantic type taxonomy appear quite encouraging. In the future, better automated methods for the evaluation of rerendering results against the existing ontologies must be developed. And most importantly, the utility and usefulness of the rerendering algorithm must be evaluated vis-a-vis achieving improvement in precision and recall for client NLP applications.

6. References

D. A. Campbell and S. B. Johnson. 1998. A Technique for Semantic Classification of Unknown Words Using UMLS Resources. In *Proceedings of AMIA Fall Symposium.*

S. A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*.

J. Castaño, J. Zhang, and J. Pustejovsky. 2002. Anaphora Resolution in Biomedical Literature. *submitted to International Symposium on Reference Resolution 2002, Alicante Spain*.

T.H. Cormen, C.E. Leiserson, and R.L. Rivest. 1990. Introduction to Algorithms. *MIT press, Cambridge, MA, 1990*.

U. Hahn and S. Schulz 2002. Massive Bio-Ontology Engineering in NLP In *Proceedings of Human Language Technology Conference.*

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of 14th International Conference on Computational Linguistics.*

D. Hindle. 1983. Deterministic Parsing of Syntactic nonfluencies. *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics.*

B. L. Humphreys, D. A. B. Lindberg, H. M. Schoolman, and G. O. Barnett. 1998. The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association*.

I. Mani. 2002. Automatically Inducing Ontologies from Corpora. *submitted to ?*.

A. McCray, A. Burgun, and O. Bodenreider. 2001. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. In *Proceedings of Medinfo, London*.

D. D. McDonald. 1992. Robust Partial Parsing through incremental multi-algorithm processing. In *Text-based Intelligent Systems*, P. Jacobs, ed. 1992.

On-line Medical Dictionary. 1998-2002. http://cancerweb.ncl.ac.uk/omd/ *Academic Medical Publishing & CancerWEB*

D. M. Pisanelli, A. Gangemi, and G. Steve. 1998. An Ontological Analysis of the UMLS Metathesaurus Journal of American Medical Informatics Association, vol. 5 S4, pp. 810-814.

J. Pustejovsky, B. Boguraev, M. Verhagen, P. Buitelaar, and M. Johnston. 1997. Semantic Indexing and Typed Hyperlinking. *AAAI Symposium on Language and the Web, Stanford, CA*

J. Pustejovsky and P. Hanks. 2001. Very Large Lexical Databases: A Tutorial Primer. Association for Computational Linguistics, Toulouse, July, 2001

J. Pustejovsky, J. Castaño, J. Zhang, B. Cochran, and M. Kotecki. 2002. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In *Pacific Symposium on Biocomputing*.

B. Rosario and M. Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. ACL.*

UMLS Knowledge Sources. Documentation. 2001. 12th edition. U.S. National Library of Medicine.

Exploiting Sublanguage and Domain Characteristics in a Bootstrapping Approach to Lexicon and Ontology Creation

Dietmar Rösner, Manuela Kunze

Otto-von-Guericke-University Magdeburg,
Institute of Knowledge and Language Processing,
P.O.box 4120,
D-39016 Magdeburg, Germany
{roesner,makunze}@iws.cs.uni-magdeburg.de

Abstract

It is very costly to build up lexical resources and domain ontologies. Especially when confronted with a new application domain lexical gaps and a poor coverage of domain concepts are a problem for the successful exploitation of natural language document analysis systems that need and exploit such knowledge sources. In this paper we report about ongoing experiments with 'bootstrapping techniques' for lexicon and ontology creation.

1. Introduction

It is very costly to build up lexical resources and domain ontologies. Especially when confronted with a new application domain lexical gaps and a poor coverage of domain concepts are a problem for the successful exploitation of natural language document analysis systems that need and exploit such knowledge sources.

We are confronted with such a situation very often in our work with the XDOC document suite, a collection of tools created to support intelligent processing of corpora of interesting textual documents taken from domains like engineering and medicine. The XDOC document workbench is currently employed in a number of applications. These include:

- knowledge acquisition from technical documentation about casting technology,
- extraction of company profiles from WWW pages,
- analysis of autopsy protocols.

The latter application is part of a joint project with the institute for forensic medicine of our university. The paper is organised as follows: We start with background information about XDOC. Then we sketch characteristics of the sublanguage of autopsy protocols and describe the core idea of our experiments. This is followed by a description of syntactic structures that are currently processed. Then clustering of co-occurrence data and its exploitation is described. A discussion of results and problems and an outlook on future work completes the paper.

2. Background: the XDOC document suite

We have designed and implemented the XDOC document suite as a workbench for the flexible processing of electronically available documents in German. The tools in the XDOC document suite(Kunze and Rösner, 2001a), (Kunze and Rösner, 2001b) can be grouped according to their function:

- preprocessing
- structure detection
- POS tagging
- syntactic parsing
- semantic analysis
- tools for the specific application: e.g. information extraction

In the semantic analysis, similar to the POS tagging, the tokens are annotated with their meaning and a classification in semantic categories like e.g. concepts and relations. For the semantic tagging we apply a semantic lexicon. This lexicon contains the semantic interpretation of a word and its case frame combined with the syntactic valence requirements. When we are confronted with a new application domain, the lexical resources must be completed with the domain specific terms. Even semantic resources with broad coverage like the semantic lexicon GermaNet for German (GermaNet-Project-Site, 2002) and Wordnet(Wordnet-Project-Site, 2002) for English, can not cover all terms of all different domains.

2.1. Design principles

The work in the XDOC project is guided by the following design principles:

- The tools shall be usable for 'realistic' documents.
  One aspect of 'realistic' documents is that they typically contain domain-specific tokens that are not directly covered by classical lexical categories (like noun, verb, ...). Those tokens are nevertheless often essential for the user of the document (e.g. an enzyme descriptor like EC 4.1.1.17 for a biochemist).
- The tools shall be as robust as possible.
  In general it can not be expected that lexicon information is available for all tokens in a document. This is not only the case for most tokens from 'nonlexical' types – like telephone numbers, enzyme names, material codes, ... –, even for lexical types there will always

be 'lexical gaps'. This may either be caused by neologisms or simply by starting to process documents from a new application domain with a new sublanguage. In the latter case lexical items will typically be missing in the lexicon ('lexical gap') and phrasal structures may not or not adequately be covered by the grammar.

- The tools shall be usable independently but shall allow for flexible combination and interoperability.
- The tools shall not only be usable by developers but as well by domain experts without linguistic training.

2.2. XML as unifying framework

We have decided to exploit XML (Bray et al., 1998) and its accompanying formalisms (e.g. XSLT (Site, 2002)) and tools (e.g. xt (Clark, 2002)) as a unifying framework. All modules in the XDOC system expect XML documents as input and deliver their results in XML format.

This decision has positive consequences for many aspects in XDOC. Take e.g. the desideratum that the tools of XDOC shall not only be usable by developers but as well by domain experts without linguistic training. Here XML and XSLT play a major role: XSL stylesheets can be exploited to allow different presentations of internal data and results for different target groups; for end users the internals are in many cases not helpful, whereas developers will need them for debugging.

2.3. Bridging lexical gaps

We do not expect extensive lexicon coding at the beginning of an XDOC application. XDOC's POS tagger and syntactic parser have therefore been augmented with a number of techniques for dealing with such 'lexical gaps'.

For POS tagging we exploit the morphology component MORPHIX (Finkler and Neumann, 1988): If a token in a German text can be morphologically analysed with MORPHIX the resulting word class categorisation is used as POS information. Note that this classification need not be unique. Since the tokens are analysed in isolation multiple analyses are often the case. Some examples: the token 'der' may either be a determiner (with a number of different combinations for the features case, number and gender) or a relative pronoun, the token 'liebe' may be either a verb or an adjective (again with different feature combinations not relevant for POS tagging).

MORPHIX's coverage can be characterised as follows: most closed class lexical items of German as well as all irregular verbs are covered. The coverage of open class lexical items is dependent on the amount of coding. The paradigms for e.g. verb conjugation and noun declination are fully covered but to be able to analyze and generate word forms their roots need to be included in the MORPHIX lexicon.

Due to lexical gaps some tokens will not get a MORPHIX analysis, at least at the beginning of an XDOC application. We then employ two techniques: We first try to make use of heuristics that are based on aspects of the tokens that can easily be detected with simple string analysis (e.g. upper-/lowercase, endings, ...) and/or exploitation of the token position relative to sentence boundaries (detected in the structure detection module). If a heuristic yields a classification the resulting POS class is added together with the name of the employed heuristic (marked as feature SRC, cf. example 1). If no heuristics are applicable we classify the token as member of the class unknown (tagged with XXX).

To keep the POS tagger fast and simple the disambiguation between multiple POS classes for a token and the derivation of a possible POS class from context for an unknown token are postponed to syntactic processing (cf. below).

3. Bootstrapping in a new domain

XDOCs most recent application is part of a joint project with the institute for forensic medicine of our university. The medical doctors there are interested in tools that help them to exploit their huge collection of several thousand autopsy protocols for their research interests. The confrontation with this corpus from a new domain has stimulated experiments with 'bootstrapping techniques' for lexicon and ontology creation.

3.1. The core idea

The core idea is the following:

When you are confronted with a new corpus from a new domain, try to find linguistic structures in the text that are easy to detect automatically and that allow to classify unknown terms in a robust manner both syntactically as well as on the knowledge level. Take the results from a run of these simple but robust heuristics as an initial version of a domain dependent lexicon and ontology. Exploit these initial resources to extend the processing to more complicated linguistic structures in order to detect and classify more terms of interest automatically.

An example: In the sublanguage of autopsy protocols (in German) a very telegrammatic style is dominant. Condensed and compact structures like the following are very frequent:

- Harnblase leer. (Urinary bladder empty.)
- Harnleiter frei. (Ureter free.)
- Nierenoberflaeche glatt. (Surface of kidney smooth.)
- Vorsteherdruese altersentsprechend. (Prostate corresponding to age.)
- ...

These structures can be abstracted syntactically as <Noun><Adjective><Fullstop> and as semantically <Anatomic-entity><Attribute-value><Fullstop>. Furthermore they are easily detectable.

In our experiments we have exploited this characteristic of the corpus extensively to automatically deduce an initial lexicon (with nouns and adjectives) and an initial ontology (with concepts for anatomic regions or organs and their respective features and values).

3.2. A sublanguage analysis of autopsy protocols

The telegrammatic style of autopsy protocolls results in a preference for 'verbless' structures. It is e.g. much more likely that a finding like 'the mouth was open' is expressed as 'Mund geoeffnet.' (mouth open) although a more verbose paraphrase like 'Der Mund ist geoeffnet.' may occur sometimes.

Another consequence of the style is a preference for noun compounds in contrast to semantically equivalent noun phrases.

When referring to a concept like 'weight of the liver' the noun compound 'Lebergewicht' is more likely than the noun phrase 'Gewicht der Leber'. This generalizes for the weight of other organs: 'Organgewicht' is more likely than the noun phrase 'Gewicht des/der X'.

The need for contextual interpretation of terms may be seen as another consequence of the style. In local context with an organ as topic generic terms like 'Gewicht' (weight) or 'Durchmesser' (diameter) have to be interpreted as referring to the object in focus, i.e. the organ.

3.3. Refinements of the initial approach

In our corpus it is very likely that a syntactic structure of the type <Noun><Adjective><Fullstop> can semantically be interpreted as <Anatomic-entity><Attribute-value><Fullstop>, but there are exceptions. An example: 'Flachschnitt unauffaellig.' Here the noun does not denote an anatomic entity, but is referring to a diagnostic procedure in autopsy. On the other hand the adjective co-occurs with anatomic entities as well.

So the initial approach needs refinement: as long as the number of exceptions of a simple pattern (here: <Noun> <Adjective> <Fullstop>) in a heuristic remains small the exceptions (here: noun 'Flachschnitt') are simply checked first before the heuristic is applied for all cases in which the exceptions are not present.

3.4. Exploitation of syntactic constraints

Pattern based analysis is a first step only. For full syntactic parsing we apply a chart parser based on context free grammar rules augmented with feature structures. The output of a robust POS tagger is used as input to parsing. The POS tagger works on token in isolation. Its output may contain:

- multiple POS classes,
- unknown classes of open world tokens and
- tokens with POS class, but without or only partial feature information.

Example 1 *unknown token classified as noun with heuristics*

```
<NP TYPE="COMPLEX" RULE="NPC3" GEN="FEM"
        NUM="PL" CAS="_">
  <NP TYPE="FULL" RULE="NP1" CAS="_"
           NUM="PL" GEN="FEM">
       <N SRC="UNG">Blutanhaftungen</N>
  </NP>
  <PP CAS="DAT">
    <PRP CAS="DAT">an</PRP>
    <NP TYPE="FULL" RULE="NP2" CAS="DAT"
           NUM="SG" GEN="FEM">
      <DETD>der</DETD>
      <N SRC="UC1">Gekroesewurzel</N>
    </NP>
  </PP>
</NP>
```

The latter case results from some heuristics in POS tagging that allow to assume e.g. the class noun for a token but do not suffice to detect its full paradigm from the token (note that there are approximately two dozen different morphosyntactic paradigms for noun declination in German).

For a given input the parser attempts to find all complete analyses that cover the input. If no such complete analysis is achievable it is attempted to combine maximal partial results into structures covering the whole input (Rösner, 2000).

A successful analysis may be based on an assumption about the word class of an initially unclassified token (tagged XXX). This is indicated in the parsing result (feature AS) and can be exploited for learning such classifications from contextual constraints. In a similar way the successful combination from known feature values from closed class items (e.g. determiners, prepositions) with underspecified features in agreement constraints allows the determination of paradigm information from successfully processed occurrences. In example 2 features of the unknown word "Mundhoehle" (mouth) could be derived from the features of the determiner within the PP (e.g. gender feminine).

Example 2 *unknown token classified as adjective and features derived through contextual constraints*

```
<NP TYPE="COMPLEX" RULE="NPC3" GEN="MAS" NUM="SG"
    CAS="NOM">
  <NP TYPE="FULL" RULE="NP3" CAS="NOM" NUM="SG"
        GEN="MAS">
    <DETI>kein</DETI>
    <XXX AS="ADJ">ungehoeriger</XXX>
    <N>Inhalt</N>
  </NP>
  <PP CAS="DAT">
    <PRP CAS="DAT">in</PRP>
    <NP TYPE="FULL" RULE="NP2" CAS="DAT" NUM="SG"
        GEN="FEM">
      <DETD>der</DETD>
      <N SRC="UC1">Mundhoehle</N>
    </NP>
  </PP>
</NP>"
```

The grammar used in syntactic parsing is organised in a modular way that allows to add or remove groups of rules. This is exploited when the sublanguage of a domain contains linguistic structures that are unusual or even ungrammatical in standard German.

Example 3 *Excerpt from syntactic analysis*

```
<PP CAS="AKK">
  <PRP CAS="AKK">auf</PRP>
  <NP TYPE="COMPLEX" RULE="NPC3" GEN="MAS" NUM="SG"
   CAS="AKK">
    <NP TYPE="FULL" RULE="NP1" CAS="AKK" NUM="SG"
    GEN="MAS">
      <N>Flachschnitt</N>
    </NP>
    <PP CAS="AKK">
      <PRP CAS="AKK">in</PRP>
      <NP TYPE="FULL" RULE="NP2" CAS="AKK" NUM="SG"
      GEN="NTR">
        <DETD>das</DETD>
        <N>Gewebe</N>
      </NP>
    </PP>
  </NP>
</PP>
```

3.5. Beyond simple patterns

At the time we work with a 'light' grammar of 40 rules. This grammar contains basic rules (for the analysis of noun phrases and preposition phrases) and specific rules, based on the patterns of the sublanguage.

We have just started to extract binary relations from completely parsed sentences. Following patterns of the sublanguage are analysed in this manner: Simple structures like: <NP> <Adjective> <Fullstop> will be analysed as <Anatomic-entity> <Attribute-value> <Fullstop>.

Example 4 *For example: 'Gehirngaenge frei.'. The Analysis returns:*

```
<RATT-V>
   <ENTITY>Gehoergaenge</ENTITY>
   <VALUE CNT="1">frei</VALUE>
</RATT-V>
```

All results of this analysis are also marked as XML structure. The attribute 'CNT' contains the number of occurences of the attribute value in context with the anatomic entity. A similar pattern is the structure <NP> 'ist|sind'[1] <Adjective>|<Verb> <Fullstop>.

Example 5 *For example: 'Gangsysteme sind frei.' or 'Augen sind geschlossen'. The Analysis returns:*

```
<RATT-V>
   <ENTITY>Gangsysteme</ENTITY>
   <VALUE CNT="1">frei</VALUE>
</RATT-V>
<RATT-V>
   <ENTITY>Augen</ENTITY>
   <VALUE CNT="1">geschlossen</VALUE>
</RATT-V>
```

Further on we analyse structures which contain more than attribute and domain entity. We extended our analyses to structures, which e.g. contain a modifier like 'sehr' (very) or a negator like 'nicht' (not) and other adjectives.

Example 6 *Result of the example: 'Brustkorb nicht sehr breit.*'

```
<RATT-V>
   <ENTITY>Brustkorb</ENTITY>
   <VALUE CNT="1">nicht-sehr-breit</VALUE>
</RATT-V>
```

Here the attribute is compounded of a series of words from different wordclasses, because at the time we work with binary relations only. In ongoing work we will further detail this semantic interpretation. In addition we analyse complex structures like coordinated structures. There exist various pattern, e.g. <NP> <Adjective>|<Verb> 'und' <Adjective>|<Verb><Fullstop>. These structures are interpreted as <Anatomic-entity> <Attribute-value1> 'and' <Anatomic-entity> <Attribute-value2><Fullstop>.

Example 7 *For example: 'Beckengeruest festgefuegt und unversehrt.'. The result is:*

```
<RATT-V>
   <ENTITY>Beckengeruest</ENTITY>
   <VALUE CNT="1">festgefuegt</VALUE>
   <VALUE CNT="1">unversehrt</VALUE>
</RATT-V>
```

[1] is|are, | expresses alternatives in pattern

The inverse structure (the coordination at the beginning of the pattern) e.g. <Adjective> 'und' <Adjective> <NP> <Fullstop> can also be analysed.

Example 8 *For example: 'Akute und chronische Erweiterung des Herzens.*'

```
<RATT-V>
   <ENTITY>Erweiterung des Herzens</ENTITY>
   <VALUE CNT="1">akute</VALUE>
   <VALUE CNT="1">chronische</VALUE>
</RATT-V>
```

Another coordinated pattern is <NP> 'und' <NP> <Adjective>|<Verb> <Fullstop>. The semantic interpretation is similar to the analysis of the simple structures: <Anatomic-entity1> <Attribute-value> 'and' <Anatomic-entity2><Attribute-value><Fullstop>.

Example 9 *For example: 'Rippen und Wirbelsaeule intakt.' The result is:*

```
<RATT-V>
   <ENTITY>Rippen</ENTITY>
   <VALUE CNT="1">intakt</VALUE>
</RATT-V>
<RATT-V>
   <ENTITY>Wirbelsaeule</ENTITY>
   <VALUE CNT="1">intakt</VALUE>
</RATT-V>
```

The pattern, like the example 'Leber und Niere ohne Besonderheiten.'('Liver and kidney without findings.'), differs from the last described structures in the kind of the attribute. In this structure the attribute is described by a preposition phrase. The analysis returns

Example 10 *Result of 'Leber und Niere ohne Besonderheiten.':*

```
<RATT-V>
   <ENTITY>Leber</ENTITY>
   <VALUE CNT="1">ohne Besonderheiten</VALUE>
</RATT-V>
<RATT-V>
   <ENTITY>Niere</ENTITY>
   <VALUE CNT="1">ohne Besonderheiten</VALUE>
</RATT-V>
```

4. Ontology creation

4.1. Analysis of co-occurrence data

Co-occurrence data are used for clustering: We start e.g. with an adjective token that is related to a single noun type only in the analysed data.

If - again within the corpus given - this noun co-occurs only with this very adjective then the relation between the noun's concept and the property denoted by the adjective is very strong. It may even be the case that the adjective-noun-combination is a name like fixed phrase.

If the noun co-occurs with other adjectives as well it is interesting to uncover the relation between the adjectives and denoted properties respectively.

There are a number of possibilities:

- Two adjectives may be used as 'quasi-synonyms',
- Adjectives may be in an antinomy relation,
- Adjectives may refer to discrete values of a property that are linearly ordered on a scale,

- Adjectives refer to values of different properties.

We can proceed in a zig-zag-manner:

We have started with a single adjective and checked for its co-occurring noun. We then asked for other adjectives co-occurring with this noun. In the next step we extend the set of nouns with those nouns that co-occur with at least one of the adjectives in the adjective set.

Then we can extend the adjective set accordingly. The process will definitely stop if in a step the set to be expanded (either the noun or the adjective set) is no longer growing and has thus reached a fixed point.

As soon as the zig-zag-procedure adds an adjective to the adjective set thats co-occurs with many nouns of different type then in the next step, when the co-occurring adjectives of all these nouns are added, we may produce (nearly) a full covering of all adjectives and of all nouns respectively.

4.2. Exploiting co-occurrence information

4.2.1. Concept detection

A noun phrase of the type <Adj> <Noun> may be like the name of a concept but this does not always hold and depends on usage.

An example: 'fluessige Galle' as in 'In der Gallenblase fluessige Galle' is a property value, not a name. On the other hand 'harte Hirnhaut' is to be treated as nameing a concept. This can be inferred from the usage of the NP 'harte Hirnhaut' in structures of the type

<NP><Adj> like 'harte Hirnhaut perlmuttergrau'.

4.2.2. Concept classification

Currently linguistic structures are mapped into binary relations. An example:

Harte Hirnhaut grauweiss.

is an application of the grammar rule with

<NP> <Adjective> <Fullstop>

as right hand side. This establishes a <Property> <Concept> pair.

If we invert this relation (i.e. give a listing of all property values that co-occur – with number of occurrences above a threshold – with the concept) this yields:

Harte Hirnhaut: glaenzend, grauweiss, perlmuttergrau, weisslich-gelblich-verfaerbt, intakt, grauroetlich, blaeulich-durchscheinend

If we analyze these adjectives (and compounded adjective groups) we find the following:

- there is one very generic property 'intakt' (engl. 'intact') that is usable with almost any anatomic-entity
- the adjective 'glaenzend' is characterising the visual appearance of the brain skin as shiny
- all other adjectives denote a variety colors

Thus the brain skin can be classified as an anatomic-entity whose color values are relevant in autopsy reports.

4.2.3. Concept grouping

Clustering of co-occurrence data allows to detect candidates for semantic groups as well as synonyms and/or paraphrases.

- 'spiegelnd': 'Herzueberzug', 'Lungenueberzug'
- 'unversehrt': 'Haut des Rueckens', 'Stirnhaut'
- 'frei': 'Gehoergange', 'Ausfuehrungsgang', 'Kehlkopfeingang'

All concepts co-occurring with 'frei' are of the type tube.

4.3. Ontological relations

What ontological relations can be inferred?

- Is-a: Leber Is-a Organ
- Part-of: Schleimhaut Part-of Magen (generalized Schleimhaut Part-of Organ)
- other n-nary relations: e.g. 'nicht widernatuerlich beweglich'

Further on we can find a classification of relations resp. the domain range of an relation. For example the relation 'geoeffnet' (opened) can be changed by modifier in the attribute-value

- 'geoeffnet'
- 'spaltweit-geoeffnet'
- 'spaltfoermig-geoeffnet'
- 'geschlossen' (as opposite to 'geoeffnet')

5. Discussion

Our current work is of an investigative nature. The size of the corpus is still small. It is planned to apply the techniques developed with the initial corpus to the collection of several thousand protocols. The number of occurrences is still small and statistical methods are therefore not yet adequate. Even if quantitative measures are not applicable on the basis of this corpus occurrence data can be interpreted qualitatively.

Since we have just recently started with the domain of autopsy protocols there are e.g. still gaps in grammar coverage and in the tagging process (not every unknown word can be classified by heuristics). In the corpus currently approx 37 % of the sentences and telegrammatic structures can be fully processed (i.e. get at least one reading covering the structure as a whole; multiple readings are possible.) Experiments with the full corpus will allow to evaluate how reliable the results are.

The telegrammatic style results in shorter and – on the first sight – 'simpler' linguistic structures. As a trade-off these structures are less constrained and this e.g. complicates the derivation of morphosyntactic features from context or makes inferred results less reliable.

An example: If 'Nieren' is an unknown token the full sentence 'Die Nieren sind unversehrt' allows to infer that the token is a plural form, the same inference is not possible from the telegrammatic version 'Nieren unversehrt'.

5.1. XDOC as a workbench

We are aiming at a workbench with a rich functionality but we do not expect a fully automatic and autonomous solution. The user shall be supported as good as possible but s/he will still be involved in the process.

Our approach is interactive. The user has to confirm suggestions from the system. He is accepting or refusing, but can delegate searching, comparing, counting etc. to the system.

5.2. Acquisition of domain knowledge

Some findings in autopsy protocols are results of measurements: values of weights, sizes, diameters etc. are reported.

This allows to collect 'typical values' and to gain distributions for ranges of values.

For weights a typical pattern is:

<organ>gewicht <number> g.

'Lebergewicht ... g'

From the texts we derived the range of the weight-relation for example for the organ kidney as 135 g to 270 g (in a medical manual the weight of the kidney is defined in the range of 120 g to 300 g).

Sometimes contextual interpretation is necessary:

<organ><property-value>. Gewicht <number> g.

Here the generic term 'Gewicht' (weight) has to be interpreted as referring to the organ in focus.

Similar constructions are employed for other indicators like diameters.

5.3. Future work:

For the quality of inferences the detection of synonyms and paraphrases plays a major role, e.g. 'Blase' and 'Harnblase' do refer to the same organ, 'Stirnhaut' and 'Haut der Stirn' denote the same region: the skin of the forehead.

A general solution for coordinated structures will be necessary.

A subtype of coordinated structures includes truncation of compounds. An example: 'Wangen- und Kinnpartie unauffaellig.' The reconstruction of the untruncated term is not always as simple as in the example. For this task we need an approach similar to the one described in (Buitelaar and Scaleanu, 2002). It must not only be analysed which is the semantic meaning of the word, but rather which is the word, which was truncated. One criterion is, that the words must have the same semantic category.

A general component for the semantic treatment of noun compounds is needed. This will have to interact with contextual interpretation. In an example like

24. Hirngewicht 1490 g. Windungen abgeflacht, Furchen verstrichen. ...

it has to be detected that with the reference to the weight of the brain ('Hirngewicht') the brain is established as topic and that the terms 'Windungen' and 'Furchen' are referring to findings about the brain's visible appearance.

Autopsy protocols are written in a way such that the course of the autopsy is directly reflected in discourse structure. The autopsy on the other hand follows anatomic structures and their neighbourhood relations. In local contexts we both find part-of relations between anatomic structures as well as neighbourhood relations.

The analysis of noun phrases needs to be more fine grained. Structures like 'Haut des Rueckens' or 'Haut ueber der Nase' should e.g. be interpreted as localisation information that is specifying regions of the skin (here: 'skin of the back' and 'skin of the nose').

6. References

Tim Bray, Jean Paoli, and C.M. Sperberg-McQueen. 1998. Extensible Markup Language (XML) 1.0. *http://www.w3.org/TR/1998/REC-xml-19980210*.

P. Buitelaar and B. Scaleanu. 2002. Extending Synsets with Medical Terms.

J. Clark. 2002. http://www.jclark.com.

W. Finkler and G. Neumann. 1988. MORPHIX: a fast Realization of a classification-based Approach to Morphology. In H. Trost, editor, *Proc. der 4. Österreichischen Artificial-Intelligence Tagung, Wiener Workshop Wissensbasierte Sprachverarbeitung*, pages 11–19. Springer Verlag.

GermaNet-Project-Site. 2002. http://www.sfs.nphil.uni-tuebingen.de/lsd/.

M. Kunze and D. Rösner. 2001a. XDOC - Extraktion, Rep"asentation und Auswertung von Informationen. In *GLDV-Workshop: Werkzeuge zur automatischen Analyse und Verarbeitung*.

M. Kunze and D. Rösner. 2001b. An XML-based Approach for the Presentation and Exploitation of Extracted Information. In *International Workshop on Web Document Analysis*.

D. Rösner. 2000. Combining robust parsing and lexical acquisition in the XDOC system. In *KONVENS 2000 Sprachkommunikation*, ITG-Fachbericht 161, ISBN 3-8007-2564-9, pages 75–80. VDE Verlag, Berlin, Offenbach.

XSL Site. 2002. http://www.w3.org/style/xsl.

Wordnet-Project-Site. 2002. http://www.cogsci.princeton.edu/ wn/.