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.

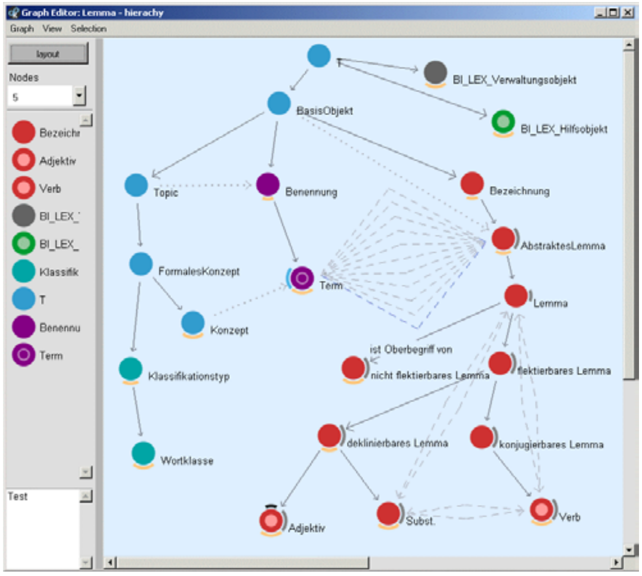

Figure 1: Top level of the Duden Ontology

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

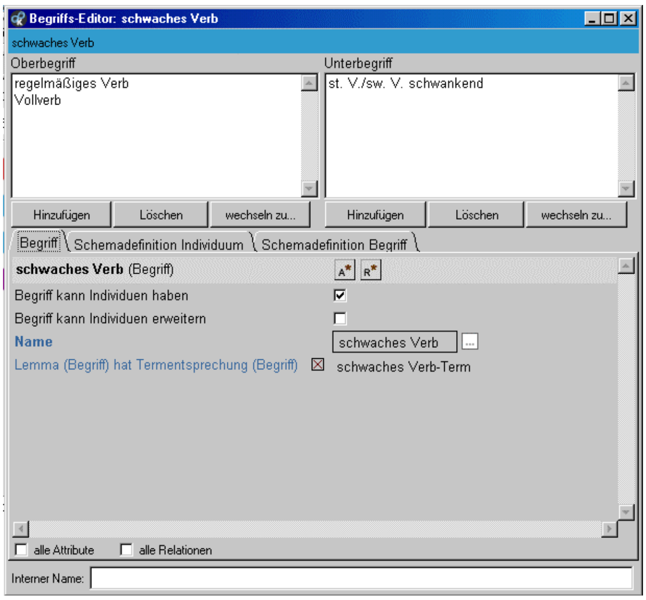

Figure 2: Concept Editor

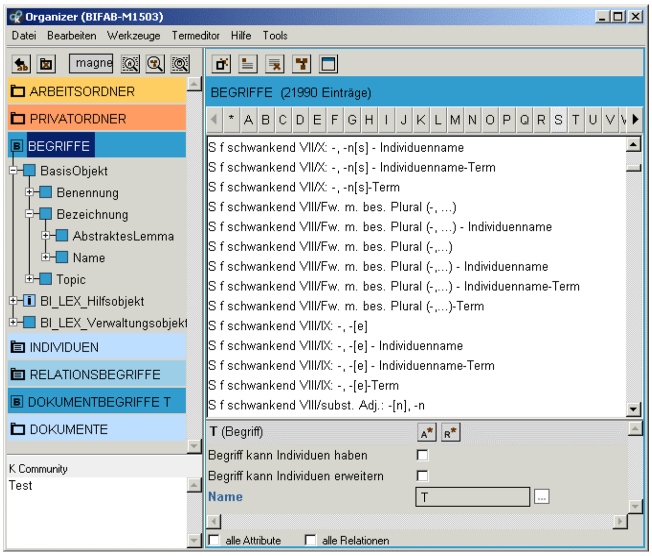

Figure 3: K-Organizer

extensions or roles, the terms, which we use to represent the different senses of a lemma.

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.

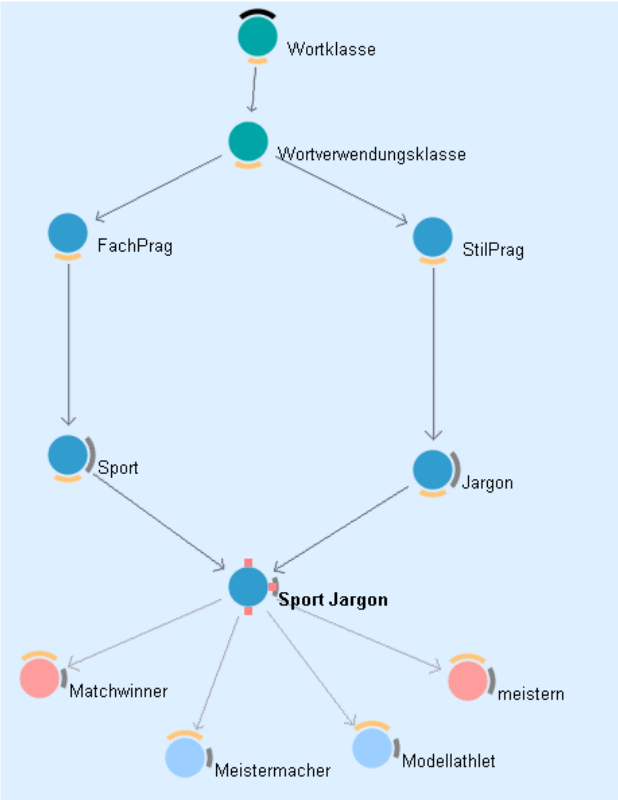

Figure 4: Example of pragmatic classes

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 polyhierarchy as a means for flexible navigation and access.

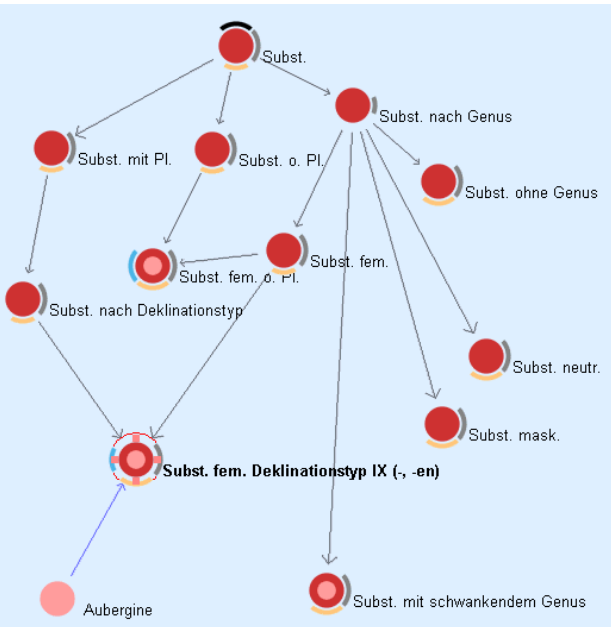

Figure 5: Example of noun classes

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".

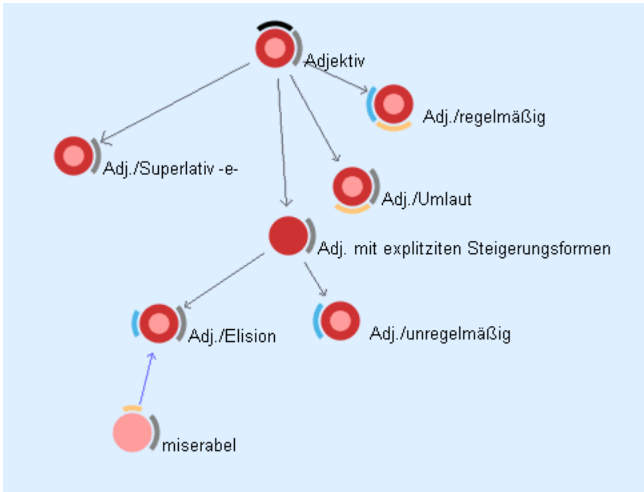

Figure 6: Example of adjective classes

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: Natürlichsprachige 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.