Modeling Monolingual and Bilingual Collocation Dictionaries in Description Logics

Dennis Spohr and Ulrich Heid
Institute for Natural Language Processing
University of Stuttgart
Azenbergstr. 12, D-70174 Stuttgart, Germany
{spohrds,heid}@ims.uni-stuttgart.de

Abstract

This paper discusses an approach to modeling monolingual and bilingual dictionaries in the description logic species of the OWL Web Ontology Language (OWL DL). The central idea is that the model of a bilingual dictionary is a combination of the models of two monolingual dictionaries, in addition to an abstract translation model. The paper addresses the advantages of using OWL DL for the design of monolingual and bilingual dictionaries and proposes a generalized architecture for that purpose. Moreover, mechanisms for querying and checking the consistency of such models are presented, and it is concluded that DL provides means which are capable of adequately meeting the requirements on the design of multilingual dictionaries.

1 Introduction

We discuss the modeling of linguistic knowledge about collocations, for monolingual and bilingual electronic dictionaries, for multiple uses in NLP and for humans.

Our notion of *collocation* is a lexicographic one, inspired by (Bartsch, 2004); we start from her working definition: "collocations are lexically and/or pragmatically constrained recurrent cooccurrences of at least two lexical items which are in a direct relation with each other." The fact of being lexically and/or pragmatically constrained leads to translation problems, as such constraints are language specific. With Hausmann (2004), we assume that collocations have a *base* and a *collocate*, where the base is autosemantic and thus translatable without reference to the collocation, whereas the collocate is synsemantic, i.e. its reading is selected within a given collocation. Examples of collocations according to this definition include adjective+noun-combinations (*heavy smoker*, *strong tea*, etc.), verb+subject- (*question arises*, *question comes up*) and verb+complement-groups (*give+talk*, *take+walk*) etc. The definition excludes however named entities (*Rio de Janeiro*) and frequent compositional groups (e.g. *the police said...*). Our data have been semi-automatically extracted from 200 million words of German newspaper text of the 1990s (cf. Ritz (2005)).

We claim that a detailed monolingual description of the linguistic properties of collocations provides a solid basis for bilingual collocation dictionaries. The types of linguistic information needed for NLP and those required for human use, e.g. in text production or translation into a foreign language, overlap to a large extent. Thus it is reasonable to define comprehensive monolingual data models and to relate these with a view to translation.

In section 2, we briefly list the most important phenomena to be captured (see also Heid and Gouws (2006)); section 3 introduces OWL DL, motivates its choice as a representation format and describes our monolingual modeling. In section 4, we discuss and illustrate the bilingual dictionary architecture.

2 Collocation Data

Properties of collocations. A mere list of word pairs or sequences (*give a talk*, *lose one's patience*) is not a collocation dictionary. For use in NLP, linguistic properties of the collocations and of their components must be provided: these include the category of the components ($give_V$ + $talk_N$), the

distribution of base (*talk*) and collocate (*give*), as well as morphosyntactic preferences, e.g. with respect to the number of an element (e.g. *have high hopes*), the use of a determiner (*lose one's$_{poss}$|{} patience*, cf. Evert et al. (2004)).

For collocations to be identifiable in the context of a sentence (e.g. to avoid attachment ambiguity in parsing) and, conversely, in generation, to be correctly inserted into a sentence, the syntagmatic behavior of collocations must be described. This includes their function within a sentence (e.g in the case of adverbial NPs) and the subcategorization of their components, e.g. with support verb constructions (*make the proposal to + INF*). As subcategorization is not fully predictable from the subcategorization of the noun (how to explain the preposition choice in *Unterstützung finden bei jmdm*, 'find support in so.', be supported?), we prefer to encode the respective data in the monolingual dictionary. To support translation mapping at the complement level, the representation of each complement contains its grammatical category (NP, AP, etc.), its grammatical function (subject, object, etc.) and a semantic role inspired by FrameNet[1]. This allows us to cater for divergence cases: *jmd$_{Subj/SPEAKER}$ bringt jmdm$_{Ind.Obj/ADDRESSEE}$ etw.$_{Obj/TOPIC}$ in Erinnerung* vs. *someone$_{Subj/SPEAKER}$ reminds someone$_{Obj/ADDRESSEE}$ of sth.$_{Prep.Obj/TOPIC}$*.

Relations involving collocations. For language generation, paraphrasing or for summarization, paradigmatic relations of collocations must also be modeled. These include synonymy, antonymy and taxonomic relations, but also morphological ones (word formation) and combinations of collocations. Synonymy and antonymy should relate collocations with other collocations, but also with single words and with idioms: all three types should have the same status. Next to strict synonymy, there may be 'quasi-synonymy'.

Transparent noun compounds tend to share collocates with their heads (*Pause einlegen*, *Rauchpause einlegen*, *Kaffeepause einlegen*): if the relation between compound and head (*Kaffeepause – Pause*) and between the respective collocations is made explicit, this knowledge can be exploited in translation, when a compositional equivalent is chosen (*have a (smoking/coffee) break*). Paraphrasing and its applications also profit from an explicit representation of morphological relations between collocates: *submit + proposal*, *submission of + proposal* and *submitter of + proposal* all refer to the same collocational pattern.

A formal model for a collocation dictionary, monolingual and/or bilingual, has to keep track of the above mentioned properties and relations of collocations; both should be queriable, alone and in arbitrary combinations.

Other collocation dictionaries and dictionary architectures. Most of the above mentioned properties and relations have been discussed in the descriptive literature, but to our knowledge, they have never been modeled all in an electronic dictionary. The Danish *STO* dictionary (Braasch and Olsen, 2000) and Krenn's (2000) database of German support verb+PP-constructions both emphasize morphosyntactic preferences, but do not include relations. The electronic learners' dictionaries DAFLES and DICE[2] focus on semantic explanations of collocations, but do not contain details about most of the properties and relations mentioned above. The implementation of Mel'čuk's Meaning⇔Text-Theory in the DiCo/LAF model[3] comes closest to our requirements, insofar as it is highly relational and includes some though not all of the morphological relations we described above.

The Papillon project (Sérasset and Mangeot-Lerebours, 2001) proposes a general architecture for the interlingual linking of monolingual dictionaries; as it is inspired by the DiCo formalizarion, it foresees links between readings, e.g. to account for morphological relations. This mechanism could in principle be extended to syntagmatic phenomena; we are, however, not aware of a Papillon-based collocation dictionary.

3 Modeling in OWL DL

In this section, we present the main features of OWL DL and their relevance to the modeling of lexical data. Section 3.2 addresses the design of a monolingual collocation dictionary using OWL DL (Spohr, 2005).

3.1 Main Features of OWL

OWL DL is the description logic sublanguage of the OWL Web Ontology Language (Bech-

[1] Cf. `http://framenet.icsi.berkeley.edu/`

[2] Cf. `http://www.kuleuven.ac.be/dafles/` and DICE: `http://www.dicesp.com/`

[3] Cf. `http://olst.ling.umontreal.ca/dicouebe/`

hofer et al., 2004), combining the expressivity of OWL with the computational completeness and decidability of Description Logics (Baader et al., 2003)[4]. Properties of OWL DL relevant for lexical modeling are listed and discussed in the following.

Classes. An OWL DL data model consists of a subsumption hierarchy of classes, i.e. a class X subsumes all its subclasses X_1 to X_n. While classes represent concepts, their instances (called *OWL individuals*) represent concrete manifestations in the model. Classes and their instances can be constrained by stating assertions in the model definition, e.g. a class can be defined as being disjoint with other classes, which means that instances of a certain class cannot at the same time be instances of the disjoints of this particular class.

Properties. Classes are described by properties. These can be used either to specify XML Schema Datatypes (*datatype properties*) or to relate instances of one class to instances of (probably) other classes (*object properties*). These classes are then defined as the domain and range of a property, i.e. a particular property may only relate instances of classes in its domain to instances of classes in its range. In addition to this, a property may be assigned several distinct formal attributes, such as symmetric, transitive or functional, and can be defined as the inverse of another property. Similar to classes, properties can be structured hierarchically as well, which, among others, facilitates the use of underspecified information in queries (see section 3.2).

Inferences. The possibility to infer explicit knowledge from implicit statements is a core feature of OWL DL and can be performed by using DL reasoners (such as FaCT[5], Pellet[6] or RacerPro[7]). The most basic inference is achieved via the subsumption relation among classes or properties in the respective hierarchy (see above), but also more sophisticated inferences are possible. Among others, these may involve the formal attributes of properties just mentioned. For example, stating that instance A is linked to B via a symmetric property P leads a reasoner to infer that B is also linked to A via P. In conjunction with transitivity, a relatively small set of explicit statements may suffice to interrelate several instances implicitly (i.e. all instances in a particular equivalence class created by P).

Consistency. In addition to inferences, DL reasoners can further be used to check the consistency of an OWL DL model. One of the primary objectives is to check whether the assertions made about classes and their instances (see above) are logically consistent or whether there are contradictions. This consistency checking is based on the *open-world assumption*, which states that "what cannot be proven to be true is not believed to be false" (Haarslev and Möller, 2005). Since lexical data occasionally demand a *closed world*, other checking formalisms are required, which are mentioned in section 3.2 below.

3.2 Monolingual Collocation Dictionary

A data model for a monolingual collocation dictionary based on OWL DL has been presented in (Spohr, 2005). It was designed using the Protégé OWL Plugin (Knublauch et al., 2004) and makes use of the advantages of OWL DL mentioned above.

Lexical vs. descriptive entities. On the class level, the model distinguishes between lexical entities (e.g. single-word and multi-word entities, such as collocations or idioms) and descriptive entities (e.g. gender, part-of-speech, or subcategorisation frames), with lexical entities being linked to descriptive entities via properties. More than 40 of these *descriptive properties* have been modeled. In order to reflect the distinction between metalanguage vocabulary and object language vocabulary, the two types of entities can be separated such that they are part of different models. In other words, the classes and instances of descriptive entities constitute a *model of descriptions*, which is imported by a *lexicon model* containing classes and instances of lexical entities (see also section 4.1 below).

Lexical relations. In addition to descriptive properties, the data model also contains a number of *lexical relations* linking lexical entities, such as morphological or semantic relations. These relations have been structured

[4] As the emphasis in our work is on morphology, syntax and lexical combinatorics, we profit from the formal properties of DL without feeling the need for non-monotonicity as implemented, for example, in DATR (Evans and Gazdar, 1996).

[5] `http://www.cs.man.ac.uk/~horrocks/FaCT/`

[6] `http://www.mindswap.org/2003/pellet/`

[7] `http://www.racer-systems.com`

hierarchically and contain several subproperties, such as `hasCompound` or `isSynonymOf`, which use the formal attributes mentioned in section 3.1. For instance, `isSynonymOf` has been defined as a symmetric and transitive property (as opposed to the non-transitive `isQuasiSynonymOf`; see section 2), while `hasCompound` has been defined as the inverse of a property `isCompoundOf`. A small sample of descriptive and lexical relations of the collocation *Kritik üben* is illustrated in Figure 1 below.

Property	Value
hasLemma	"Kritik üben"
hasCompound	Selbstkritik_ueben
isSynonymOf	kritisieren_VV_1
hasCollocationType	V-Nobj_acc
hasComplementation	SubcatFrame_12
hasExampleSentence	Example_84
isInCorpus	HGC-STZ

Figure 1: Sample of the properties of *Kritik üben*

Semantic relations link lexical entities on the conceptual (i.e. word sense) level. Therefore, the synonym of *Kritik üben* is not some general single-word entity `kritisieren_VV`, but a particular word sense of *kritisieren*, `kritisieren_VV_1` in this case (see Spohr (2005) for more detail).

Queries. The data model can be queried very efficiently using the Sesame framework (Broekstra et al., 2002; Broekstra, 2005) and its associated query language SeRQL. An example query retrieving all collocations and their types is given below, along with a sample of the results[8].

```
SELECT *
FROM  {A} rdf:type {lex:Collocation},
      {A} lex:hasCollocationType {B}
```

A	B
in_Frage_kommen	V-PPpobj
Kritik_ueben	V-Nobj_acc
Lob_aussprechen	V-Nobj_acc
zu_Last_legen	V-PPpobj

Figure 2: Query for retrieving collocations and their types, along with results

Due to the fact that the relations in the data model have been structured hierarchically, it is possible to state underspecified queries. Figure 3 illustrates an underspecified query for semantically related entities, regardless of the precise nature of this relation. Hence, the first two rows in the result table below contain synonym pairs, while the last two rows contain antonym pairs.

```
SELECT *
FROM  {A} lex:hasSemanticRelationTo {B}
```

A	B
Kritik_ueben	kritisieren_VV_1
kritisieren_VV_1	Kritik_ueben
Kritik_ueben	Lob_aussprechen
Lob_aussprechen	Kritik_ueben

Figure 3: Underspecified query for semantically related entities, along with results

As is indicated in Figure 3, the results appear twice, i.e. they contain every combination of those entities between which the relation holds. This is due to the fact that the respective semantic relations have been defined as symmetric properties (see above).

Consistency and data integrity. Section 3.1 mentioned the distinction between the open-world assumption and the closed-world assumption. While the consistency checking performed by DL reasoners is generally based on an open world, it is vital especially for lexical data to simulate a closed world in order to check data integrity. Consider, for instance, the assertion that every collocation has to have a base and a collocate. Due to the open-world assumption, a DL reasoner would never render a collocation violating this constraint inconsistent, simply because it cannot prove that this collocation has either no base or no collocate. In order for this to happen, the simulation of a closed world is needed. In our approach, this is achieved by stating consistency constraints in SeRQL. Figure 4 below illustrates a constraint for the purpose just mentioned.

This query retrieves all collocations and subtracts those who have a path to both a base and a collocate. The result set then contains exactly those instances which have either no base or no collocate.

[8] In these examples, `lex:` is the namespace prefix for resources defined in the data model.

```
SELECT Coll
FROM  {Coll} rdf:type {lex:Collocation}

MINUS

SELECT Coll
FROM  {Coll} lex:hasBase {};
             lex:hasCollocate {}
```

Figure 4: Constraint checking: does every collocation have a base and a collocate?

4 Bilingual Model Architecture

Based on the definition of a monolingual collocation dictionary described above, the architecture of a bilingual dictionary model can be designed such that it is made up of several components (i.e. OWL models). These are introduced in the following.

4.1 Components of a Bilingual Dictionary

The components of a bilingual dictionary are illustrated in Figure 5.

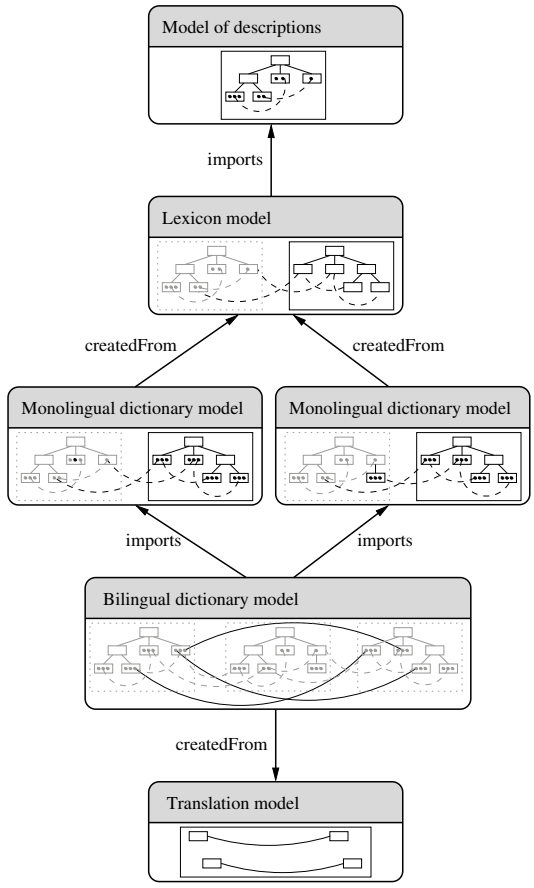

Figure 5: Architecture of a bilingual dictionary model

Model of descriptions. The most basic component of a bilingual dictionary model is a *model of descriptions*, which contains language-independent classes and instances of descriptive entities, as well as the relations among them (see section 3.2).

Lexicon model. The model of descriptions is imported by an abstract *lexicon model* via the `owl:imports` statement (see (Bechhofer et al., 2004)). The effect of using the import statement is that the lexicon model can access the classes, instances and properties defined in the description model without being able to alter the data therein. In addition to the thus available classes, the lexicon model further provides classes of lexical entities and relations among them, as well as relations linking lexical and descriptive entities.

Monolingual dictionary model. The lexicon model serves as input for the creation of a *monolingual dictionary model*, i.e. the lexicon model is not imported by the dictionary model, rather the dictionary model is an *instantiation* of it. There are practical reasons for doing so, the most important one being that the class of lexical entities (defined in the lexicon model) and its instances (defined in the monolingual dictionary) thus have the same namespace prefix, which would not be the case if the lexicon model was imported by the monolingual dictionary. The advantages are most obvious in the context of the mapping between monolingual dictionary models (see section 4.2). Finally, a monolingual dictionary may further introduce its own instances (or even classes) of descriptive entities, i.e. descriptions which are language-specific and which are hence not part of the language-independent model of descriptions (see above).

Translation model. The *translation model* is an abstract model containing only relations between monolingual dictionary models, i.e. it does not contain class definitions. Since the model is required to be generic, these relations do not have a specified domain and range, as otherwise the translation model would be restricted to a single language pair. The specification of the domain and range of the relations is performed in the final model of the bilingual dictionary.

Bilingual dictionary model. The *bilingual dictionary model* is an instantiation of the translation model. It further imports two monolingual dictio-

nary models and specifies the domain and range of the abstract relations in the translation model (see section 4.2 below).

4.2 Mapping between Models

By importing the monolingual dictionaries, each of these models is assigned a unique namespace prefix, e.g. `english:` or `german:`. Thus, in an English-German dictionary, for instance, a relation called `hasTranslation` may be defined as a symmetric property linking lexical entities of the English monolingual dictionary model (i.e. its *domain* is defined for instances with the `english:` prefix) to lexical entities of the German model (i.e. instances with `german:`). This translation mapping is illustrated in Figure 6 for the collocation *Kritik üben*.

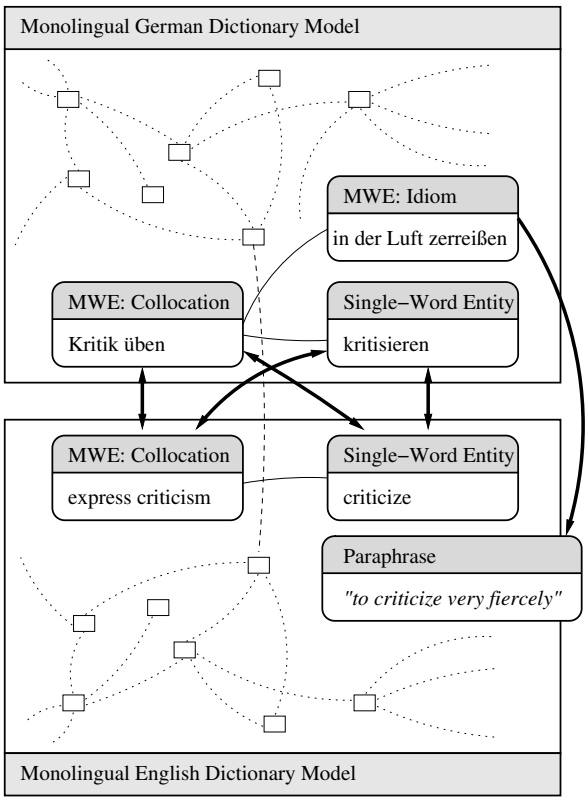

Figure 6: Translation mapping between monolingual dictionaries

As is indicated there, multi-word entities can be translated as single-word entities and vice versa. Moreover, since `hasTranslation` has been defined as a symmetric property, the translation mapping is bidirectional. However, since some instance in one language model might not have an equivalent instance in the other model, a further property can be defined which links the respective entity to a new instance created in the bilingual model (see *Paraphrase* in the figure above). As this instance is only required for the modeling of this particular bilingual dictionary, it is not part of the "original" monolingual models, and hence the relation between the respective entities is not bidirectional.

In addition to the translation mapping of lexical entities, it may further be necessary to map instances of descriptive entities of one model onto instances in the other model. As was mentioned in section 4.1, the model of descriptions contains language-independent descriptive entities. Since both monolingual dictionaries import the model of descriptions (via the lexicon model), the two "versions" of it are unified in the bilingual model. However, it is certainly conceivable to have two languages which both avail themselves of a descriptive entity that is not language-independent, but which is the same for the two languages in question. For example, not all languages have the gender `neuter`. English and German, however, do have it, and therefore an English-German bilingual dictionary has to express that `english:neuter` is the same as `german:neuter`. In OWL, this can be achieved by using the `owl:sameAs` statement, which expresses exactly the circumstances just mentioned.

4.3 Example Query

A query retrieving the situation depicted in Figure 6 is given below. It extracts the (quasi-)synonyms of *Kritik üben* (which *Kritik üben* itself is a part of) and their respective translations and/or paraphrases. The latter is achieved by restricting the properties that `Rel2` may stand for to those having the prefix `bdm:`, i.e. the prefix defined for the bilingual dictionary model. In other words, the query leaves the exact relation between `B` and `C` underspecified and simply restricts it to being defined in the bilingual dictionary, which only contains relations linking instances belonging to different monolingual dictionaries. The results are shown in the table below.

5 Conclusion

We have described a model for monolingual and bilingual collocation dictionaries in OWL DL. This formalism is well suited for the intended modularization of linguistic resources, be they language- or language-pair- specific (our dictio-

```
SELECT DISTINCT B, C
FROM {} german:hasLemma {A};
       Rel1 {B} Rel2 {C}
WHERE A LIKE "Kritik üben"
AND (Rel1 = german:isSynonymOf
 OR Rel1 = german:isQuasiSynonymOf)
AND namespace(Rel2) = bdm:
```

B	C
kritisieren_VV_1	express_criticism
kritisieren_VV_1	criticize_VV_1
Kritik_ueben	express_criticism
Kritik_ueben	criticize_VV_1
in_Luft_zerreissen	"to criticize very fiercely"

Figure 7: Query for retrieving the (quasi-)synonyms of *Kritik üben* and their translations and paraphrases, along with results

nary models), generalized over one or more languages (our lexicon model), or more abstract, in the sense of a meta-model or an inventory of the descriptive devices shared by the linguistic descriptions of several languages (our model of descriptions, see figure 5 above). This model of descriptions will be larger for related languages (e.g. the indo-european ones), and smaller for typologically very diverse languages; it is however by no means meant to have any interlingual, let alone universal function, but is rather understood in the sense of PARGRAM's shared inventory of descriptive devices[9].

We have modelled so far about 1000 collocations, their components, preferences and relations (also with single words); we intend to considerably enlarge the collocation dictionary, using the possibilities to combine OWL DL models with databases, offered by the Sesame framework. The formalism also supports experiments with credulous inferencing at the level of translation equivalence, e.g. by following not only explicit equivalence relations, but also synonymy relations: in line with the query discussed in section 4.3 above (cf. Figure 7), one could also start from the English *express criticism* and retrieve the equivalent collocation *Kritik üben* as well as its (quasi-)synonyms *kritisieren* (single word) and *in der Luft zerreißen* (idiom), which may thus be proposed as equivalent candidates for *express criticism*.

More such investigations into the data collection are planned; they may require non-standard access to the dictionary, i.e. access via paths involving other properties and relations than just lemmas and equivalents. The relational nature of the dictionary supports this kind of exploration; we intend to specify and implement a 'linguist-friendly' query overlay to SeRQL and a Graphical User Interface to make such explorations more easy.

[9] Cf. `http://www2.parc.com/istl/groups/nltt/pargram/gram.html`

References

Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. 2003. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, Cambridge, UK.

Sabine Bartsch. 2004. *Structural and Functional Properties of Collocations in English. A Corpus Study of Lexical and Pragmatic Constraints on Lexical Cooccurrence*. Narr, Tübingen, Germany.

Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. 2004. OWL Web Ontology Language Reference. Technical report.

Anna Braasch and Sussi Olsen. 2000. Formalised representation of collocations in a Danish computational lexicon. In *Proceedings of the EURALEX International Congress 2000*, Stuttgart, Germany.

Jeen Broekstra, Arjohn Kampman, and Frank van Hermelen. 2002. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In *Proceedings of the First International Semantic Web Conference (ISWC 2002)*, pages 54–68, Sardinia, Italy.

Jeen Broekstra. 2005. *Storage, Querying and Inferencing for Semantic Web Languages*. Ph.D. thesis, Vrije Universiteit Amsterdam, The Netherlands.

Roger Evans and Gerald Gazdar. 1996. DATR: A language for lexical knowledge representation. *Computational Linguistics*, 22(2):167–216.

Stefan Evert, Ulrich Heid, and Kristina Spranger. 2004. Identifying Morphosyntactic Preferences in Collocations. In *Proceedings of LREC-2004*, Lisbon, Portugal.

Volker Haarslev and Ralf Möller, 2005. *RacerPro User's Guide and Reference Manual, Version 1.8.1*.

Franz Josef Hausmann. 2004. Was sind eigentlich Kollokationen? In Karin Steyer, editor, *Wortverbindungen - mehr oder weniger fest*, pages 309–334. Institut für Deutsche Sprache: Jahrbuch 2003.

Ulrich Heid and Rufus H. Gouws. 2006. A model for a multifunctional electronic dictionary of collocations. Draft of a paper submitted to *EURALEX 2006*.

Holger Knublauch, Mark A. Musen, and Alan L. Rector. 2004. Editing description logic ontologies with the Protégé OWL plugin. In *Proceedings of the International Workshop in Description Logics - DL2004*, Whistler, BC, Canada.

Brigitte Krenn. 2000. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. Ph.D. thesis, DFKI Universität des Saarlandes, Saarbrücken, Germany.

Julia Ritz. 2005. Entwicklung eines Systems zur Extraktion von Kollokationen mittels morphosyntaktischer Features. Diploma thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.

Gilles Sérasset and Mathieu Mangeot-Lerebours. 2001. Papillon lexical database project: Monolingual dictionaries & interlingual links. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium: NLPRS-2001*, pages 119–125, Tokyo, Japan.

Dennis Spohr. 2005. A Description Logic Approach to Modelling Collocations. Diploma thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.