Inducing a Computational Lexicon from a Corpus with Syntactic and Semantic Annotation

Dennis Spohr[1,2], Aljoscha Burchardt[2], Sebastian Padó[2], Anette Frank[2] and Ulrich Heid[1]

[1]Inst. for Natural Language Processing
University of Stuttgart
Stuttgart, Germany
spohrds,heid@ims.uni-stuttgart.de

[2]Dept. of Computational Linguistics
Saarland University
Saarbrücken, Germany
albu,pado,frank@coli.uni-sb.de

Abstract

To date, linguistically annotated corpora are mainly exploited for feature-based training of automatic labelling systems. In this paper, we present a general approach for the Description Logics-based modelling of multi-layered annotated corpora which offers (i) *flexible and enhanced querying functionality* that goes beyond current XML-based query languages, (ii) a basis for *consistency checking*, and (iii) a general method for *defining abstractions over corpus annotations*.

We apply this method to the syntactically and semantically annotated SALSA/TIGER corpus. By defining abstractions over the corpus data, we generalise from a large set of individual corpus annotations to a corresponding *lexicon model*. We discuss issues arising from modelling multi-layered corpus annotations in Description Logics and illustrate the benefits of our approach at concrete examples.

1 Introduction

One of the most exciting developments in computational linguistics over the recent years is the increasing availability of large corpora with multiple layers of linguistic annotation. For example, the WSJ portion of the Penn Treebank is now not only annotated syntactically, but also with semantic roles and discourse connectives. Such corpora offer the possibility to investigate empirically the interactions between different layers of linguistic analysis, and much recent work has focussed on the acquisition of statistical models for automatic linguistic annotation at different linguistic levels. While the

resulting models fill an important need, they do not lend themselves easily to human interpretation or integration with other knowledge sources. We suggest that these needs can be addressed by inducing a *multi-level lexicon* by *generalising* over corpus instances. This lexicon makes the information on the individual linguistic levels accessible and explicitly represents their interaction, which gives rise to three major benefits:

Querying for linguistic data analysis. Corpus data are usually represented in XML; however, current XML-based query tools support queries that involve multiple linguistic levels only in very restricted ways. In a recent survey, Lai and Bird (2004) found that almost all query tools cannot deal with intersecting hierarchies, i.e., tree-shaped analyses on multiple linguistic levels, which are ubiquitous in corpora with multi-layer annotation. A powerful lexicon representation can overcome this limitation, allowing for integrated querying of multiple levels.

Consistency checking. The complexity of annotation schemes tends to increase for 'deeper' linguistic analysis, and so does the effort of ascertaining that given annotation instances are *consistent* with the annotation scheme. For example, the annotation of semantic roles requires a large number of categories. These are usually lexically specific, and not universally applicable; in addition, the observance of inter-category relations such as obligatoriness or mutual exclusion are usually not enforced by annotation tools. Checking for consistency of such complex constraints on the corpus level arguably requires a large effort. In contrast, a declarative formalisation of the lexicon model that integrates the annotation scheme can use general KR techniques.

Abstractions and application interfaces. The granularity chosen for a given corpus annotation may diverge considerably from the optimal granularity for a specific analysis. While it is possible to obtain more abstract representations procedurally in a corpus by collapsing categories, a declarative lexicon model allows for much more flexible and finer-grained control. Abstraction can then be driven empirically, as generalisations over a large body of corpus annotations, but also theory-driven. The latter point is especially important for the integration of corpus-derived data in large, symbolic processing architecture (see Frank (2004), who lists a number of problems arising from deriving syntax-semantics mapping information directly from a corpus, and Babko-Malaya et al. (2006), who encounter similar problems in synchronising Penn Treebank with PropBank annotations).

This paper demonstrates the benefits of a declarative lexicon model by reporting on the construction of a Description Logics-based lexicon for Ger-

man that represents information about morphological, syntactic and frame semantic levels of analysis. We highlight the benefits of lexicon modelling in Description Logics (DL) and show that complex annotation schemes require careful lexicon design. In Section 2, we motivate our decision to use DL to define our lexicon model. Section 3 describes our input data and the design of the lexicon model, and Section 4 shows concrete instances of the usage of this model and provides some statistics. Section 5 gives concluding remarks.

2 Description Logic and Lexicon Modelling

The formalisation of our lexicon model is based on OWL DL, a strongly typed framework which combines the expressivity of OWL[1] with the favourable computational properties of DL, most notably decidability and monotonicity (see Baader et al., 2003). Besides the availability of reasoning and consistency checking services, one of the major benefits of using OWL is the possibility to conceive of the lexicon as a graph, i.e. a complex entity with a high degree of interaction between various levels of linguistic description (cf. Spohr and Heid, 2006).

FrameNet and DL. Two earlier studies have used DL to model FrameNet, but have limited themselves to modelling the *definitional* part of the resource (Narayanan et al., 2002, Baumgartner and Burchardt, 2004). Our formalisation additionally comprises the modelling of *annotation instances* in a manner comparable to Scheffczyk et al. (2006). However, our conception of classes, properties and axioms is geared towards detecting inconsistencies in the corpus annotation with a theorem prover, and expressing various generalisations over annotated corpus instances. Moreover, our lexicon model is interfaced with a storage and querying architecture.

Other modelling frameworks. Description Logic is certainly not the only option for designing a lexicon model.[2] Recent alternatives include Lexical Systems (Polguère, 2006) and the Lexical Markup Framework (Francopoulo et al., 2006). Our main reason for preferring OWL DL is its strong logical foundation and its well-defined model-theoretic semantics. In addition, these models lack properties we consider essential, such as inheritance or hierarchical classification in general (Lexical Systems). While LMF is in

[1] `http://www.w3.org/2004/OWL/`

[2] Efforts in corpus-based extraction of "lexica" for LTAG grammars (cf. Xia et al., 2000) are loosely related to the present work, but are tied to a specific syntactic theory, and are lacking a general logical framework for generalisation and lexicon modelling.

principle powerful enough to define a model of FrameNet's frames and roles and of their interrelations[3], it is an open question whether LMF – in its current state – can represent these data in an equally principled way as OWL DL. For example, it appears difficult in LMF to capture information such as the roleset of a frame, or relations between roles, as restrictions on their syntactic and semantic properties. Consequently, checking the consistency of annotation instances wrt. their definitions is likely to require complex mechanisms.

3 A Lexicon Model For Syntax and Semantics

3.1 Role-semantic annotation in the SALSA corpus

The lexicon model we have designed is used to store the data annotated in the SALSA project (Burchardt et al., 2006). SALSA builds on the TIGER corpus, a German newspaper corpus with manual syntactic annotation (Brants et al., 2002), and adds a *role-semantic layer* following the FrameNet paradigm (Baker et al., 1998). FrameNet is a semantic dictionary which associates English words and expressions (*targets*) with semantic classes called *frames* and lists *semantic roles* for each frame. Since frames have been found to exhibit a high degree of language independence, SALSA re-uses English frames for German (see Burchardt et al. (2006) for details). An English example annotation is given in Figure 1: The verb "*criticise*" is annotated with the frame JUDGMENT_COMMUNICATION, the semantic roles EVALUEE and COMMUNICATOR pointing to "*Robert Tuttle*" and "*Washington*", respectively.

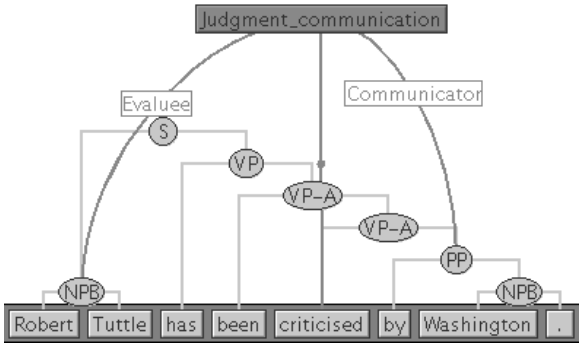

Figure 1: Frame-semantic annotation

[3]Cf. working paper "*Extended examples of lexicons using LMF*", G. Francopoulo (2005).

3.2 Core requirements for multi-level lexicons

In the process of creating our lexicon, we have identified three core requirements which we consider as central for any computational lexicon that models multi-level data, and which strongly influenced the design of our model.

Intersecting hierarchies. Making the information from different levels of linguistic analysis (cf. Figure 1) accessible requires a mapping of the various annotation layers and their intersections onto a common representation. In our case, this issue becomes particularly manifest in the complex interaction between syntax and semantics. Any syntactic unit in a corpus sentence – be it a constituent, part of a constituent or even part of a word – may potentially evoke a frame or represent a role. At the same time, it is assigned a (morpho-)syntactic category and may hold a grammatical function in the sentence. Moreover, a sentence may also contain multiple frame annotations. Consequently, a syntactic constituent can be assigned more than one role, within different frame annotations. We deal with this issue by means of *multiple instantiation* and tight linking between semantic annotation instances and their syntactic realisations (see Section 4).

Lexicographic relevance and generalisations. One of the main aims of the SALSA project has been to identify instances of lexicographically relevant phenomena in the corpus, such as metaphors or frame-evoking multiword expressions (Atkins et al., 2003). A requirement for the lexicon model is thus not only to capture this *lexical knowledge* and make it accessible in a straightforward yet flexible way, but also to support abstraction over specific annotation instances in order to derive further generalisations about these phenomena, such as valence patterns. In order to achieve this, we proceed in a bottom-up fashion: during lexicon generation, annotation instances are collected and grouped. In doing so, we *generalise* over particular annotation instances and make the respective types of phenomena explicit in the lexicon.

Quantitative tendencies. Closely connected to the previous issue is the demand to be able to derive quantitative tendencies from corpus annotations.

There are basically two approaches to this task: (i) frequencies of a fixed number of phenomena are calculated during the conversion process and then hard-coded in the lexicon (*static*), or (ii) the structure of the lexicon model and the query engine are designed such that it is possible to derive quantitative tendencies at query time (*dynamic*). Solution (i) is prescriptive wrt. the frequency information that can be accessed by the user, thus limiting the usability of the lexicon as a tool of active lexicographic and lexical-

semantic research. Solution (ii) compensates for the lack of the first option, but imposes higher demands on lexicon architecture and user skills.

The combination of OWL DL and the Sesame framework (Broekstra et al., 2002) supports both solutions: The graph-based view on the lexicon enables representing a high amount of interrelation between the different levels of linguistic analysis, where frequency information could be encoded within separate sub-graphs. As the flexibility of Sesame's query language SeRQL makes explicit information easily explorable and also enables the detection of *implicit* correlations, we opted for the more flexible solution (ii).

3.3 Filling TBox and ABox

When using DL, any information to be represented has to belong either to the *TBox* ('terminological box' – concepts) or the *ABox* ('assertional box' – facts). A straightforward division for our data would be to represent the annotation data in the ABox and the definition of the underlying syntactic and semantic categories in the TBox. However, as our annotation scheme itself is highly structured, this division is not so clear-cut. Moreover, we had to take into account the final size of the model, which contains more than 6.000 conceptual classes instantiated by about some 20.000 corpus sentences.

One fundamental question that arose was whether (i) to put individual frames such as SELF_MOTION or LEADERSHIP into the ABox as instances of a general class `Frame` or whether (ii) to represent them in the TBox as classes. The main reasons in favour of (i) are first, that our lexicon is built directly from a corpus containing fine-grained multi-level annotations which entails a certain degree of heterogeneity. Second, (i) is more explicit in that it supports querying information about the frames, such as their inherent semantic roles and relations. On the other hand, if individual frames are modelled as classes in the TBox (ii), they can impose well-formedness constraints on their annotated instances. As one of our main goals is consistency checking, we model frames as classes. We also extended the TBox with a small hierarchy that models the different annotation types.

Figure 2 illustrates part of the T-Box class hierarchy. The left-hand side illustrates the linguistic model, in which frames and roles are defined according to FrameNet's inheritance relation. Since both frames and roles may inherit from more than one frame/role, these are *multiple inheritance* hierarchies. The figure also shows (functional) edge labels and part-of-speech tags provided by TIGER and a corresponding set of (largely theory-independent) grammatical functions and categories we have defined to support the extraction of generalised valence information from the lexicon.

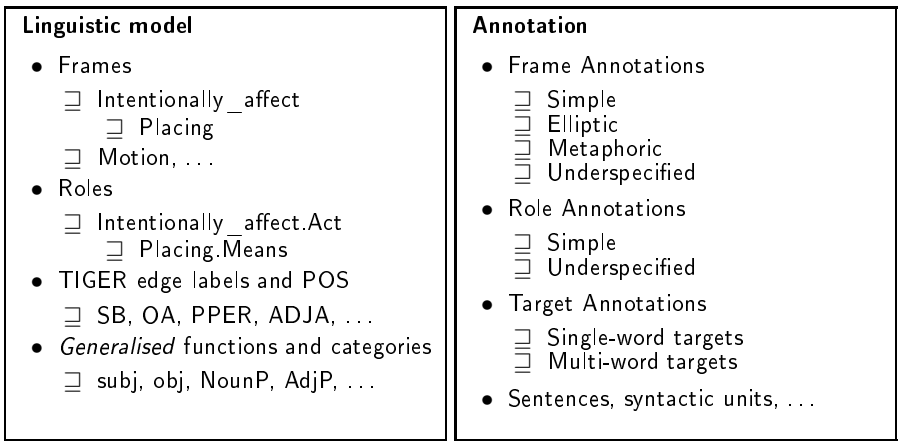

Figure 2: Schema of the TBox class hierarchy

The right-hand side of Figure 2 shows the hierarchy of bottom-up generalisations over the annotation scheme mentioned in Section 3.2. For example, a frame target is marked as a *multi-word* target if it is assigned to at least one non-terminal node, or two or more (terminal or non-terminal) nodes – excluding particle verbs and "zu" infinitives. Thus, annotation classes serve as a kind of filter for separating different types of annotation phenomena, for which different properties have been defined.

Annotation instances from the corpus instantiate (multiple) classes in both hierarchies (see Figure 2): On the annotation side according to their types of phenomena; on the definition side according to their frames or roles, and their syntactic functions and categories (both TIGER and generalised). Again, this interaction makes it possible to impose various well-formedness constraints on the annotation instances, e.g., axioms defining the admissible relations between a particular frame and its roles. This is illustrated in the DL statement below, which expresses that an instance of the PLACING frame may *at most* have the roles GOAL, PATH, etc.

$$\text{Placing} \sqsubseteq \exists.\text{hasRole}\ (\text{Placing.Goal} \sqcup \text{Placing.Path} \sqcup \ldots)$$
$$\text{Placing} \sqsubseteq \forall.\text{hasRole}\ (\text{Placing.Goal} \sqcup \text{Placing.Path} \sqcup \ldots)$$

Relations between roles can be formalised in a similar way. A prominent example is the *excludes* relation in FrameNet, which prohibits the co-occurrence of roles like CAUSE and AGENT of the PLACING frame. This can be expressed by the following statement.

$$\text{Placing} \sqsubseteq \neg((\exists.\text{hasRole Placing.Cause}) \sqcap (\exists.\text{hasRole Placing.Agent}))$$

The restrictions are used in checking the consistency of the semantic annotation; instances violating these constraints are identified by the theorem prover as incoherent.

4 Concrete Examples and Figures

An annotated corpus sentence. In order to sum up and substantiate the discussions in the previous sections, we present the partial lexicon representation of an annotated corpus sentence, namely "*... was das offizielle Kroatien aber in beträchtliche völkerrechtliche Schwierigkeiten bringen würde ...*"[4]. The predicate "*bringen* (to bring)" has been analysed as metaphorical; we focus on the literal (SOURCE) reading, described with a PLACING frame.[5]

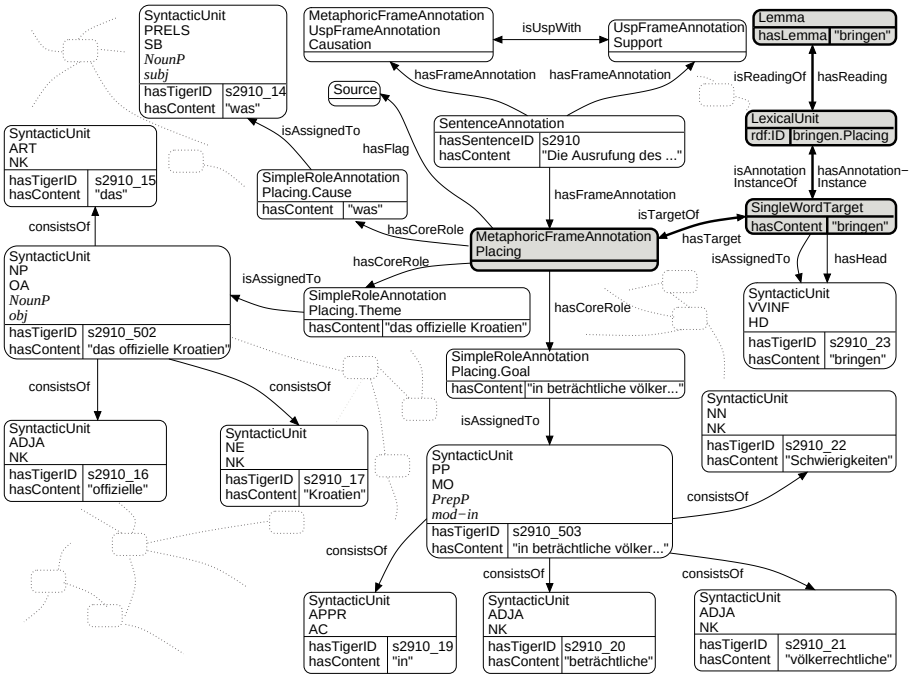

Figure 3: Partial lexicon representation of an annotated corpus sentence

The boxes in Figure 3 represent instances in the lexicon model, with the

[4] "*... which would, however, get Croatia into serious trouble with international law ...*"

[5] The understood (TARGET) reading is shown in the subgraph above PLACING. It has been analysed as an underspecification between aspectual support and a CAUSATION frame (see Burchardt et al., 2006 for details on the use of underspecification).

respective classes listed above the horizontal line, and datatype properties below it. The links between these instances indicate OWL object properties which have been defined for the instantiated classes. For example, the metaphorical PLACING frame is shown as the grey box in the middle.

Multiple inheritance is indicated in Figure 3 by instances carrying more than one class, such as the instance in the left centre, which instantiates the classes `SyntacticUnit`, `NP`, `OA`, `NounP` and `obj`. Multi-class instances inherit the properties of each of these classes, so that e.g. the metaphoric frame annotation of the PLACING frame in the middle of the figure has both the properties defined for *frames* (`hasCoreRole`) and for *frame annotations* (`hasTarget`). As discussed in Section 3.2, multiple inheritance is also used to define normalisations of the syntactic inventory provided by TIGER to derive generalised valence patterns from the annotated corpus data. These are indicated in italics (e.g., *NounP*).

The figure also highlights the model's graph-based structure with a high degree of interrelation between the lexicon entities. For example, the grey PLACING frame instance is directly related to its roles (left, bottom), its lexical anchor (right), the surrounding sentence (top), and a flag (top left).

Querying. Information is retrieved from the lexicon by stating queries which specify paths through the model graph. For example, the SeRQL query in Figure 4 extracts all lemmas which evoke the PLACING frame (cf. the grey boxes in Figure 3). Grouping of the results yields the frequency distribution shown in the table below.

```
SELECT LEMMA
FROM {LEMMA} salsa:hasReading {} salsa:hasAnnotationInstance {}
             salsa:isTargetOf {} serql:directType {salsa:Placing}
```

Lemma	*No. of instances*	*Lemma*	*No. of instances*
legen	38	ablegen	3
bringen	35	kippen	3
nehmen	13	einlagern	1
plazieren	4	einpflanzen	1

Figure 4: SeRQL query and results: lemmas evoking the PLACING frame

The normalisation of corpus categories discussed above allows for the querying of annotated data on different levels of granularity. Table 1 shows results for a query that retrieves the syntactic realisation patterns for individual semantic roles, contrasting the *specific* (as annotated in the corpus) and *normalised* levels. On the complete lexicon, the use of normalised categories

reduces the number of realisation patterns from 2,176 to 1,026, capturing generalisations which would have remained undetected otherwise.

Specific		**Normalised**	
Role	*Category/POS*	*Role*	*Category*
`Placing.Theme`	`NN`	`Placing.Theme`	`NounP`
`Placing.Theme`	`NE`	–	
`Placing.Theme`	`PPER`	–	
`Statement.Message`	`S`	`Statement.Message`	`Sent`
`Statement.Message`	`VROOT`	–	

Table 1: Results based on specific vs. normalised categories

Consistency control. Since we aim at providing a coherent representation of the annotated corpus sentences in the lexicon model, it is essential to ensure that the data fed into the lexicon are consistent in the first place. Here, we can make use of SALSA's distributed annotation practice, which includes the extraction of about 500 subcorpora for specific lemmas from the original TIGER corpus. This organisation of the data enables a rather local detection of inconsistencies in the annotations, such as uses of non-existing frames, typos or – on the level of DL – logical incoherences with respect to the definitions of the underlying framework. Once these have been removed, the respective corpora can be added incrementally to the consistent data that are already in the lexicon. The whole process of building the lexicon is schematised in Figure 5 below, including the construction of the model, abstractions and generalisations over corpus annotations, conversion of the corpora as well as consistency control and the final lexicon representation, which can then be accessed by users and/or applications.

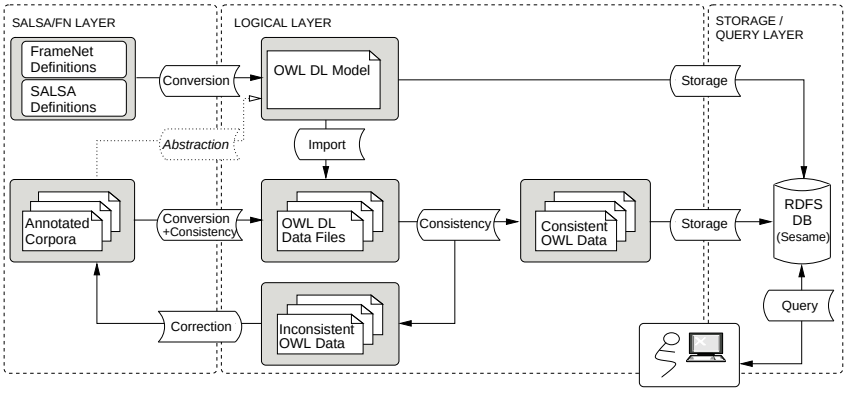

Figure 5: Workflow of the lexicon creation process

Size of the lexicon. As the corpus annotation is still work in progress, we have carried out a preliminary experiment based on a sample of more than 7,500 different sentences – about one third of the targeted size of the first SALSA release. The sample yielded a total of more than 150,000 instances in the lexicon, instantiating 185 different frame classes and 631 role classes. Table 2 provides more detailed figures.

Type	*No. of instances*	*Type*	*No. of instances*
Lemmas	337	Frame annotations	9,069
Lemma-frame pairs	727	Role annotations	17,082
Sentences	7,618	Syntactic units	114,441

Table 2: Instance count based on the current corpus data

5 Conclusion

In this paper, we have shown how a DL-based lexicon model derived from a corpus with multi-layer syntactic and semantic annotation can yield a clean formalisation of complex linguistic structures. Interfaced with a database and powerful query language, the model is easily accessible for human inspection, supporting among others the computation of frequency data and formulation of linguistically insightful queries. Moreover, the graph-based structure allows for incremental refinement and normalisation of the model.

While the work on the SALSA corpus (Burchardt et al., 2006) is being completed, finalised corpora are being imported into the lexicon model. Current and future work concentrates on the refinement of consistency control and normalisation, as well as quantitative evaluation, e.g., to identify unbalanced data that can later on be supplemented.

References

Sue Atkins, Charles J. Fillmore, and Christopher R. Johnson. Lexicographic relevance: Selecting information from corpus evidence. *International Journal of Lexicography*, 16(3), 2003.

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

Olga Babko-Malaya, Ann Bies, Ann Taylor, Szuting Yi, Martha Palmer, Mitch Marcus, Seth Kulick, and Libin Shen. Issues in Synchronizing the English Treebank and PropBank. In *Proceedings of the COLING/ACL Workshop on Frontiers in Linguistically Annotated Corpora*, Sydney, Australia, 2006.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the ACL/COLING*, Montreal, 1998.

Peter Baumgartner and Aljoscha Burchardt. Logic Programming Infrastructure for Inferences on FrameNet. In *Proceedings of the 9th JELIA*, 2004.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria, 2002.

Jeen Broekstra, Arjohn Kampman, and Frank van Hermelen. Sesame: A generic architecture for storing and querying RDF and RDF Schema. In *Proceedings of the 1st ISWC*, Sardinia, Italy, 2002.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th LREC*, Genoa, Italy, 2006.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. LMF for multilingual, specialized lexicons. In *Proceedings of the 5th LREC*, Genoa, Italy, 2006.

Anette Frank. Generalisations over corpus-induced frame assignment rules. In *Proceedings of the LREC Workshop on Building Lexical Resources From Semantically Annotated Corpora*, Lisbon, Portugal, 2004.

Catherine Lai and Steven Bird. Querying and updating treebanks: A critical survey and requirements analysis. In *Proceedings of the Australasian Language Technology Workshop*, Sydney, Australia, 2004.

Srini Narayanan, Charles J. Fillmore, Collin F. Baker, and Miriam R. L. Petruck. FrameNet Meets the Semantic Web: A DAML+OIL Frame Representation. In *Proceedings of the 18th AAAI*, Edmonton, 2002.

Alain Polguère. Structural properties of lexical systems: Monolingual and multilingual perspectives. In *Proceedings of the COLING/ACL Workshop on Multilingual Language Resources and Interoperability*, Sydney, Australia, 2006.

Jan Scheffczyk, Collin F. Baker, and Srini Narayanan. Ontology-based reasoning about lexical resources. In *Proceedings of the 5th OntoLex*, Genoa, Italy, 2006.

Dennis Spohr and Ulrich Heid. Modeling monolingual and bilingual collocation dictionaries in Description Logics. In *Proceedings of the EACL Workshop on Multiword Expressions in a Multilingual Context*, Trento, Italy, 2006.

Fei Xia, Martha Palmer, and Aravind Joshi. A Uniform Method of Grammar Extraction and Its Applications. In *Joint SIGDAT/EMNLP/VLC Conference*, Hong Kong, 2000.