Formalising Multi-layer Corpora in OWL DL – Lexicon Modelling, Querying and Consistency Control

Aljoscha Burchardt[1], Sebastian Padó[2*], Dennis Spohr[3*], Anette Frank[4*] and Ulrich Heid[3]

[1]Dept. of Comp. Ling.
Saarland University
Saarbrücken, Germany
albu@coli.uni-sb.de

[2]Dept. of Linguistics
Stanford University
Stanford, CA
pado@stanford.edu

[3]Inst. for NLP
University of Stuttgart
Stuttgart, Germany
spohrds,heid@ims.uni-stuttgart.de

[4]Dept. of Comp. Ling.
University of Heidelberg
Heidelberg, Germany
frank@cl.uni-heidelberg.de

Abstract

We present a general approach to formally modelling corpora with multi-layered annotation, thereby inducing a *lexicon model* in a typed logical representation language, OWL DL. This model can be interpreted as a graph structure that offers *flexible querying functionality* beyond current XML-based query languages and powerful methods for *consistency control*. We illustrate our approach by applying it to the syntactically and semantically annotated SALSA/TIGER corpus.

1 Introduction

Over the years, much effort has gone into the creation of large corpora *with multiple layers of linguistic annotation*, such as morphology, syntax, semantics, and discourse structure. Such corpora offer the possibility to empirically investigate the interactions between different levels of linguistic analysis.

Currently, the most common use of such corpora is the acquisition of statistical models that make use of the "more shallow" levels to predict the "deeper" levels of annotation (Gildea and Jurafsky, 2002; Miltsakaki et al., 2005). While these models fill an important need for practical applications, they fall short of the general task of *lexicon modelling*, i.e., creating an abstracted and compact representation of the corpus information that lends itself to 'linguistically informed' usages such as human interpretation or integration with other knowledge sources (e.g., deep grammar resources or ontologies). In practice, this task faces three major problems:

Ensuring consistency. Annotation reliability and consistency are key prerequisites for the extraction of generalised linguistic knowledge. However, with the increasing complexity of annotations for 'deeper' (in particular, semantic) linguistic analysis, it becomes more difficult to ensure that all annotation instances are consistent with the annotation scheme.

Querying multiple layers of linguistic annotation. A recent survey (Lai and Bird, 2004) found that currently available XML-based corpus query tools support queries operating on multiple linguistic levels only in very restricted ways. Particularly problematic are intersecting hierarchies, i.e., tree-shaped analyses on multiple linguistic levels.

Abstractions and application interfaces. A pervasive problem in annotation is granularity: The granularity offered by a given annotation layer may diverge considerably from the granularity that is needed for the integration of corpus-derived data in large symbolic processing architectures or general lexical resources. This problem is multiplied when more than one layer of annotation is considered, for example in the characterisation of interface phenomena. While it may be possible to obtain coarser-grained representations procedurally by collapsing categories, such procedures are not flexibly configurable.

Figure 1 illustrates these difficulties with a sentence from the SALSA/TIGER corpus (Burchardt et al., 2006), a manually annotated German newspaper corpus which contains role-semantic analyses in the FrameNet paradigm (Fillmore et al., 2003) on top of syntactic structure (Brants et al., 2002).[1] The se-

[*] At the time of writing, Sebastian Padó and Dennis Spohr were affiliated with Saarland University, and Anette Frank with DFKI Saarbrücken and Saarland University.

[1] While FrameNet was originally developed for English, the majority of frames has been found to generalise well to other

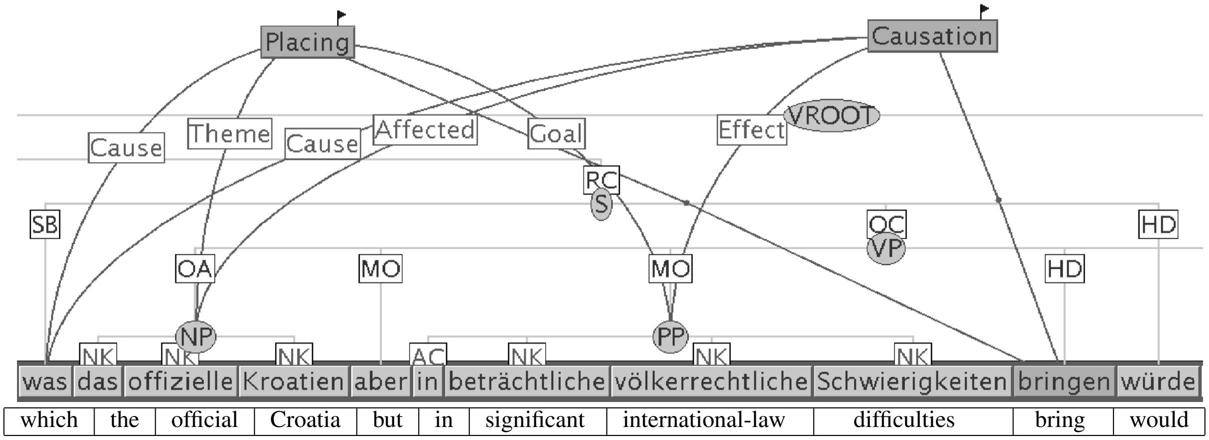
Placing
Causation
Cause
Theme
Cause
Affected
Goal
Effect
VROOT
RC
S
SB
OA
MO
MO
OC
VP
HD
HD
NP
PP
NK
NK
NK
AC
NK
NK
NK
was das offizielle Kroatien aber in beträchtliche völkerrechtliche Schwierigkeiten bringen würde
which the official Croatia but in significant international-law difficulties bring would

Figure 1: Multi-layer annotation of a German phrase with syntax and frame semantics *('which would bring official Croatia into significant difficulties with international law')*

mantic structure consists of *frames*, semantic classes assigned to predicating expressions, and the semantic roles introduced by these classes. The verb *bringen ('to bring')* is used metaphorically and is thus analysed as introducing one frame for the "literal" reading (Placing) and one for the "understood" reading (Causation), both with their own role sets.

The high complexity of the semantic structure even on its own shows the necessity of a device for consistency checking. In conjunction with syntax, it presents exactly the case of intersecting hierarchies which is difficult to query. With respect to the issue of abstraction, note that semantic roles are realised variously as individual words (*was ('which')*) and constituents (NPs, PPs), a well-known problem in deriving syntax-semantics mappings from corpora (Frank, 2004; Babko-Malaya et al., 2006).

Our proposal. We propose that the problems introduced above can be addressed by formalising corpora in an *integrated, multi-layered corpus and lexicon model* in a declarative logical framework, more specifically, the description logics-based OWL DL formalism. The major benefits of this approach are that all relevant properties of the annotation *and* the underlying model are captured in a uniform representation and, moreover, that the formal semantics of the model makes it possible to use general and efficient knowledge representation techniques for consistency control. Finally, we can extract specific *subsets* from a corpus by defining *task-specific views* on the graph.

After a short discussion of related approaches in Section 2, Section 3 provides details on our methodology. Sections 4 and 5 demonstrate the benefits of our strategy on a model of the SALSA/TIGER data. Section 6 concludes.

2 Related Work

One recent approach to lexical resource modelling is the Lexical Systems framework (Polguère, 2006), which aims at providing a highly general representation for arbitrary kinds of lexica. While this is desirable from a representational point of view, the resulting models are arguably too generic to support strong consistency checks on the encoded data.

A further proposal is the currently evolving Lexical Markup Framework (LMF; Francopoulo et al. (2006)), an ISO standard for lexical resource modelling, and an LMF version of FrameNet exists. However, we believe that our usage of a typed formalism takes advantage of a strong logical foundation and the notions of inheritance and entailment (cf. Scheffczyk et al. (2006)) and is a crucial step beyond the representational means provided by LMF.

Finally, the closest neighbour to our proposal is the ATLAS project (Laprun et al., 2002), which combines annotations with a descriptive meta-model. However, to our knowledge, ATLAS only models basic consistency constraints, and does not capture dependencies between different layers of annotation.

languages (Burchardt et al., 2006; Boas, 2005).

3 Modelling Multilevel Corpora in OWL DL

3.1 A formal graph-based Lexicon

This section demonstrates how OWL DL, a strongly typed representation language, can serve to transparently formalise corpora with multi-level annotation. OWL DL is a logical language that combines the expressivity of OWL[2] with the favourable computational properties of Description Logics (DL), notably decidability and monotonicity (Baader et al., 2003). The strongly typed, well-defined model-theoretic semantics distinguishes OWL DL from recent alternative approaches to lexicon modelling.

Due to the fact that OWL DL has been defined in the Resource Description Framework (RDF[3]), the first central benefit of using OWL DL is the possibility to conceive of the lexicon as a *graph* – a net-like entity with a high degree of interaction between layers of linguistic description, with an associated class hierarchy. Although OWL DL itself does not have a graph model but a model-theoretic semantics based on First Order Logic, we will illustrate our ideas with reference to a graph-like representation, since this is what we obtain by transforming our OWL DL files into an RDFS database.

Each node in the graph instantiates one or more classes that determine the *properties* of the node. In a straightforward sense, properties correspond to labelled edges between nodes. They are, however, also represented as nodes in the graph which instantiate (meta-)classes themselves.

The model is kept compact by OWL's support for *multiple instantiation*, i.e., the ability of instances to realise more than one class. For example, in a syntactically and semantically annotated corpus, all syntactic units (constituents, words, or even parts of words) can instantiate – in addition to a syntactic class – one or more semantic classes. Multiple instantiation enables the representation of information about several annotation layers within single instances.

As we have argued in Section 2, we believe that having one generic model that can represent all corpora is problematic. Instead, we propose to construct lexicon models for specific types of corpora. The design of such models faces two central design questions: (a) Which properties of the annotated instances should be represented?; (b) How are different types of these annotation properties modelled in the graph?

Implicit features in annotations. Linguistic annotation guidelines often concentrate on specifying the *linguistic data categories* to be annotated. However, a lot of linguistically relevant information often remains implicit in the annotation scheme. Examples from the SALSA corpus include, e.g., the fact that the annotation in Figure 1 is metaphorical. This information has to be inferred from the configuration that one predicate evokes two frames. As such information about different annotation types is useful in final lexicon resources, e.g. to define clean generalisations over the data (singling out "special cases"), to extract information about special data categories, and to define formally grounded consistency constraints, we include it in the lexicon model.

Form of representation. All relevant information has to be represented either as assertional statements in the model graph (i.e., nodes connected by edges), or as definitional axioms in the class hierarchy.[4]

This decision involves a fundamental trade-off between expressivity and flexibility. Modelling features as axioms in the class hierarchy imposes definitional constraints on all instances of these classes and is arguably more attractive from a cognitive perspective. However, modelling features as entities in the graph leads to a smaller class hierarchy, increased querying flexibility, and more robustness in the face of variation and noise in the data.

3.2 Modelling SALSA/TIGER Data

We now illustrate these decisions concretely by designing a model for a corpus with syntactic and frame-semantic annotation, more concretely the SALSA/TIGER corpus. However, the general points we make are valid beyond this particular setting.

As concerns implicit annotation features, we have designed a *hierarchy of annotation types* which now explicitly expresses different classes of annotation phenomena and which allows for the definition of annotation class-specific properties. For example, frame targets are marked as a multi-word target if

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

[3] `http://www.w3.org/RDF/`

[4] This choice corresponds to the DL distinction between TBox ("intensional knowledge") and ABox ("extensional knowledge").

Linguistic model	Annotation types
• Frames ⊒ Intentionally_affect ⊒ Placing ⊒ Motion, . . . • Roles ⊒ Intentionally_affect.Act ⊒ Placing.Means • TIGER edge labels and POS ⊒ SB, OA, PPER, ADJA, . . . • *Generalised* functions and categories ⊒ subj, obj, NounP, AdjP, . . .	• Frame Annotations ⊒ Simple ⊒ Metaphoric ⊒ Underspecified • Role Annotations ⊒ Simple ⊒ Underspecified • Target Annotations ⊒ Single-word targets ⊒ Multi-word targets • Sentences, syntactic units, . . .

Figure 2: Schema of the OWL DL model's class hierarchy ("TBox")

their span contains at least two terminal nodes. The hierarchy is shown on the right of Figure 2, which shows parts of the bipartite class hierarchy.

The left-hand side of Figure 2 illustrates the *linguistic model*, in which frames and roles are organised according to FrameNet's inheritance relation. Although this design seems to be straightforward, it is the result of careful considerations concerning the second design decision. Since FrameNet is a hierarchically structured resource with built-in inheritance relations, one important question is whether to model individual frames, such as SELF_MOTION or LEADERSHIP, and their relations either as instances of a general class `Frame` and as links between these instances, or as hierarchically structured classes with richer axiomatisation. In line with our focus on consistency checking, we adopt the latter option, which allows us to use built-in reasoning mechanisms of OWL DL to ensure consistency.

Annotation instances from the corpus instantiate multiple classes in both hierarchies (cf. Figure 2): On the annotation side according to their types of phenomena; on the linguistic side based on their frames, roles, syntactic functions, and categories.

Flexible abstraction. Section 1 introduced granularity as a pervasive problem in the use of multi-level corpora. Figure 2 indicates that the class hierarchy of the OWL DL model offers a very elegant way of defining *generalised* data categories that provide abstractions over model classes, both for linguistic categories and annotation types. Moreover, properties can be added to each abstracting class and then be used, e.g., for consistency checking. In our case, Figure 2 shows (functional) edge labels and part-of-speech tags provided by TIGER, as well as sets of (largely theory-neutral) grammatical functions and categories that subsume these fine-grained categories and support the extraction of generalised valence information from the lexicon.

An annotated corpus sentence. To substantiate the above discussion, Figure 3 shows a partial lexicon representation of the example in Figure 1. The boxes represent instance nodes, with classes listed above the horizontal line, and datatype properties below it.[5] 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 a grey box in the middle.

Multiple inheritance is indicated 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 has both the properties defined for *frames* (`hasCoreRole`) and for *frame annotations* (`hasTarget`). The generalised syntactic categories discussed above are given in italics (e.g., *NounP*).

The figure 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) indicating metaphorical use.

[5] For the sake of simplicity, we excluded explicit 'is-a' links.

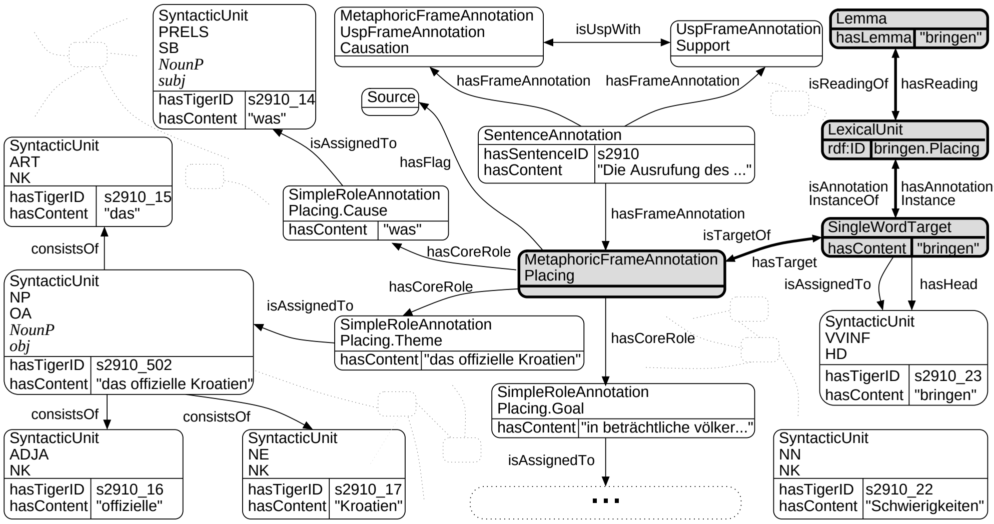

Figure 3: Partial lexicon representation of an annotated corpus sentence

Role	*Cat*	*Func*	*Freq*
Item	NounP	obj	26
Agent	NounP	subj	15
Difference	PrepP	mod-um	6
Cause	NounP	subj	4
Value_2	PrepP	mod-auf	3
Value_2	PrepP	pobj-auf	2
Value_1	PrepP	mod-von	1

Table 1: Role-category-function assignments for *senken* / CAUSE_CHANGE_OF_SCALAR_POSITION (CCSP)

Role set for senken / CCSP				*Freq*
Agent subj NounP		Item obj NounP		11
	Cause subj NounP	Item obj NounP		4
		Item obj NounP		4
Agent subj NounP		Item obj NounP	Difference mod-um PrepP	2

Table 2: Sample of role sets for *senken* / CCSP

4 Querying the Model

We now address the second desideratum introduced in Section 1, namely a flexible and powerful query mechanism. For OWL DL models, such a mechanism is available in the form of the Sesame (Broekstra et al., 2002) SeRQL query language. Since SeRQL makes it possible to extract and view arbitrary subgraphs of the model, querying of intersective hierarchies is possible in an intuitive manner.

An interesting application for this querying mechanism is to extract genuine *lexicon views* on the corpus annotations, e.g., to extract syntax-semantics mapping information for particular senses of lemmas, by correlating role assignments with deep syntactic information. These can serve both for inspection and for interfacing the annotation data with deep grammatical resources or general lexica. Applied to our complete corpus, this "lexicon" contains on average 8.5 role sets per lemma, and 5.6 role sets per frame. The result of such a query is illustrated in Table 1 for the lemma *senken ('to lower').*

From such view, frame- or lemma-specific *role sets*, i.e., patterns of role-category-function assignments can easily be retrieved. A typical example is given in Table 2, with additional frequency counts. The first row indicates that the AGENT role has been realised as a (deep) subject noun phrase and the ITEM as (deep) object noun phrase.

We found that generalisations over corpus categories encoded in the class hierarchies are central to the usefulness of the resulting patterns. For example, the number of unique mappings between semantic roles and syntactic categories in our corpus is 5,065 for specific corpus categories, and 2,289 for abstracted categories. Thus, the definition of an abstraction layer, in conjunction with a flexible query mechanism, allows us to induce lexical characterisations of the syntax-semantics mapping – aggregated and generalised from disparate corpus annotations.

Incremental refinements. Querying, and the resulting lexical views, can serve yet another purpose: Such aggregates make it possible to conduct a *data-driven* search for linguistic generalisations which might not be obvious from a theoretical perspective, and allow quick inspection of the data for counterexamples to plausible regularities.

In the case of semantic roles, for example, such a regularity would be that semantic roles are not assigned to conflicting grammatical functions (e.g., deep subject and object) within a given lemma. However, some of the role sets we extracted contained exactly such configurations. Further inspection revealed that these irregularities resulted from either noise introduced by errors in the automatic assignment of grammatical functions, or instances with syntactically non-local role assignments.

Starting from such observations, our approach supported a semi-automatic, incremental refinement of the linguistic and annotation models, in this case introducing a distinction between local and non-local role realisations.

Size of the lexicon. Using a series of SeRQL queries, we have computed the size of the corpus/lexicon model for the SALSA/TIGER data (see Table 3). The lexicon model architecture as described in Section 3 results in a total of more than 304,000 instances in the lexicon, instantiating 581 different frame classes and 1,494 role classes.

5 Consistency Control

The first problem pointed out in Section 1 was the need for efficient consistency control mechanisms. Our OWL DL-based model in fact offers two mechanisms for consistency checking: axiom-based and query-based checking.

Axiom-based checking. Once some constraint has been determined to be universally applicable, it can be formulated in Description Logics in the form of *axiomatic expressions* on the respective class in the model. Although the general interpretation of these axioms in DL is that they allow for inference of new statements, they can still be used as a kind of well-formedness "constraint". For example, if an individual is asserted as an instance of a particular class, the reasoner will detect an inconsistency if this instance does not adhere to the axiomatic class definition. For semantic role annotations, axioms can e.g. define the admissible relations between a particular frame and its roles. This is illustrated in the DL statements below, which express that an instance of PLACING may *at most* have the roles GOAL, PATH, etc.

Type	*No. of instances*
Lemmas	523
Lemma-frame pairs (LUs)	1,176
Sentences	13,353
Syntactic units	223,302
Single-word targets	16,268
Multi-word targets	258
Frame annotations	16,526
Simple	14,700
Underspecified	995
Metaphoric	785
Elliptic	107
Role annotations	31,704
Simple	31,112
Underspecified	592

Table 3: Instance count based on the first SALSA release

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

Relations between roles can be formalised in a similar way. An 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.

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

The restrictions are used in checking the consistency of the semantic annotation; violations of these constraints lead to inconsistencies that can be identified by theorem provers. Although current state-of-the-art reasoners do not yet scale to the size of entire corpora, axiom-based checking still works well for our data due to SALSA's policy of dividing the original TIGER corpus into separate subcorpora, each dealing with one particular lemma (cf. Scheffczyk et al. (2006)).

Query-based checking. Due to the nature of our graph representation, constraints can combine different types of information to control adherence to annotation guidelines. Examples are the assignment of the SUPPORTED role of support verb constructions, which ought to be assigned to the maximal syntactic constituent projected by the supported noun, or the exclusion of reflexive pronouns from the span of the target verb. However, the consistency of multi-level annotation is often difficult to check: Not only are some types of classification (e.g. assignment of semantic classes) inherently difficult; the annotations also need to be considered in context. For such cases, axiom-based checking is too strict. In practice, it is important that manual effort can be reduced by automatically extracting subsets of "suspicious" data for inspection. This can be done using SeRQL queries which – in contrast to the general remarks on the scalability of reasoners – are processed and evaluated very quickly on the entire annotated corpus data.

Example queries that we formulated examine suspicious configurations of annotation types, such as target words evoking two or more frame annotations which are neither marked as underspecified nor tagged as a pair of (non-)literal metaphorical frame annotations. Here, we identified 8 cases of omitted annotation markup, namely 4 missing metaphor flags and 4 omitted underspecification links.

On the semantic level, we extracted annotation instances (in context) for metaphorical vs. non-metaphorical readings, or frames that are involved in underspecification in certain sentences, but not in others. While the result sets thus obtained still require manual inspection, they clearly illustrate how the detection of inconsistencies can be enhanced by a declarative formalisation of the annotation scheme. Another strategy could be to concentrate on frames or lemmas exhibiting proportionally high variation in annotation (Dickinson and Meurers, 2003).

6 Conclusion

In this paper, we have constructed a Description Logics-based lexicon model directly from multi-layer linguistic corpus annotations. We have shown how such a model allows for explicit data modelling, and for flexible and fine-grained definition of various degrees of abstractions over corpus annotations.

Furthermore, we have demonstrated that a powerful logical formalisation which integrates an underlying annotation scheme can be used to directly control consistency of the annotations using general KR techniques. It can also overcome limitations of current XML-based search tools by supporting queries which are able to connect multiple levels of linguistic analysis. These queries can be used variously as an additional means of consistency control, to derive quantitative tendencies from the data, to extract lexicon views tailored to specific purposes, and finally as a general tool for linguistic research.

Acknowledgements

This work has been partly funded by the German Research Foundation DFG (grant PI 154/9-2). We also thank the two anonymous reviewers for their valuable comments and suggestions.

References

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

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

Hans C. Boas. 2005. Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18(4):445–478.

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

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

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

Markus Dickinson and W. Detmar Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proceedings of the 10th EACL*, Budapest.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

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

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

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

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

Christophe Laprun, Jonathan Fiscus, John Garofolo, and Sylvain Pajot. 2002. Recent Improvements to the ATLAS Architecture. In *Proceedings of HLT 2002*, San Diego.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*, Barcelona, Spain.

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

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