

# Corpus-based Induction of an LFG Syntax-Semantics Interface for Frame Semantic Processing

**Anette Frank**

Language Technology Lab  
DFKI GmbH  
Stuhlsatzenhausweg 3  
66123 Saarbrücken, Germany  
Anette.Frank@dfki.de

**Jiří Semecký**

Institute of Formal and Applied Linguistics  
Charles University  
Malostranské náměstí 25  
11800 Prague, Czech Republic  
semecky@ufal.ms.mff.cuni.cz

## Abstract

We present a method for corpus-based induction of an LFG syntax-semantics interface for frame semantic processing in a computational LFG parsing architecture. We show how to model frame semantic annotations in an LFG projection architecture, including special phenomena that involve non-isomorphic mappings between levels. Frame semantic annotations are ported from a manually annotated corpus to a “parallel” LFG corpus. We extract functional descriptions from the frame-annotated LFG corpus, to derive general frame assignment rules that can be applied to new sentences. We evaluate the results by applying the induced frame assignment rules to LFG parser output.<sup>1</sup>

## 1 Introduction

There is a growing insight that high-quality NLP applications for information access are in need of deeper, in particular, semantic analysis. A bottleneck for semantic processing is the lack of large domain-independent lexical semantic resources.

There are now efforts for the creation of large lexical semantic resources that provide information on predicate-argument structure. FrameNet (Baker et al., 1998), building on Fillmore’s theory of frame semantics, provides definitions of frames and their semantic roles, a lexical database and a manually annotated corpus of example sentences. A strictly corpus-based approach is carried out with ‘Prop-Bank’ – a manual predicate-argument annotation on top of the Penn treebank (Kingsbury et al., 2002).

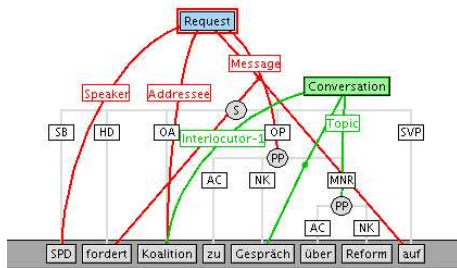
First approaches for learning stochastic models for semantic role assignment from annotated corpora have emerged with Gildea and Jurafsky (2002) and Fleischman et al. (2003). While current competitions explore the potential of shallow parsing

for role labelling, Gildea and Palmer (2002) emphasise the role of deeper syntactic analysis for semantic role labelling. We follow this line and explore the potential of deep syntactic analysis for role labelling, choosing Lexical Functional Grammar as underlying syntactic framework. We aim at a computational interface for frame semantics processing that can be used to (semi-)automatically extend the size of current training corpora for learning stochastic models for role labelling, and – ultimately – as a basis for automatic frame assignment in NLP tasks, based on the acquired stochastic models.

We discuss advantages of semantic role assignment on the basis of functional syntactic analyses as provided by LFG parsing, and present an LFG syntax-semantics interface for frame semantics, building on a first study in Frank and Erk (2004). In the present paper we focus on the corpus-based induction of a computational LFG interface for frame semantics from a semantically annotated corpus. We describe the methods used to derive an LFG-based frame semantic lexicon, and discuss the treatment of special (since non-isomorphic) mappings in the syntax-semantics interface. Finally, we apply the acquired frame assignment rules in a computational LFG parsing architecture.

The paper is structured as follows. Section 2 gives some background on the semantically annotated corpus we are using, and the LFG resources that provide the basis for automatic frame assignment. In Section 3 we discuss advantages of deeper syntactic analysis for a principle-based syntax-semantics interface for semantic role labelling. We present an LFG interface for frame semantics which we realise in a modular description-by-analysis architecture. Section 4 describes the method we apply to derive frame assignment rules from corpus annotations: we port the frame annotations to a “parallel” LFG corpus and induce general LFG frame assignment rules, by extracting syntactic descriptions for the frame constituting elements. We use LFG’s functional representations to distinguish local and non-local role assignments. The derived frame as-

<sup>1</sup>The research reported here was conducted in a cooperation project of the German Research Center for Artificial Intelligence, DFKI Saarbrücken with the Computational Linguistics Department of the University of the Saarland at Saarbrücken.



SPD requests that coalition talk about reform  
 Figure 1: SALSA/TIGER frame annotation

signment rules are reapplied to the original syntactic LFG corpus to control the results. In Section 5 we apply and evaluate the frame projection rules in an LFG parsing architecture. In Section 6 we summarise our results and discuss future directions.

## 2 Corpus and Grammar Resources

**Frame Semantic Corpus Annotations** The basis for our work is a corpus of manual frame annotations, the SALSA/TIGER corpus (Erk et al., 2003).<sup>2</sup> The annotation follows the FrameNet definitions of frames and their semantic roles.<sup>3</sup> Underlying this corpus is a syntactically annotated corpus of German newspaper text, the TIGER treebank (Brants et al., 2002). TIGER syntactic annotations consist of relatively flat constituent graph representations, with edge labels that indicate functional information, such as head (HD), subject (SB), cf. Figure 1.

The SALSA frame annotations are flat graphs connected to syntactic constituents. Figure 1 displays frame annotations where the REQUEST frame is triggered by the (discontinuous) frame evoking element (FEE) *fordert ... auf* (requests). The semantic roles (or frame elements, FEs) are represented as labelled edges that point to syntactic constituents in the TIGER syntactic annotation: the noun *SPD* for the SPEAKER, *Koalition* for the ADDRESSEE, and the PP *zu Gespräch über Reform* for the MESSAGE.

**LFG Grammar Resources** We aim at a computational syntax-semantics interface for frame semantics, to be used for (semi-)automatic corpus annotation for training of stochastic role assignment models, and ultimately as a basis for automatic frame assignment. As a grammar resource we chose a wide-coverage computational LFG grammar for German (developed at IMS, University of Stuttgart). This German LFG grammar has already been used for semi-automatic syntactic annotation of the TIGER corpus, with reported coverage of 50%, and 70%

precision (Brants et al., 2002). The grammar runs on the XLE grammar processing platform, which provides stochastic training and online disambiguation packages. Currently, the grammar is further extended, and will be enhanced with stochastic disambiguation, along the lines of (Riezler et al., 2002).

**LFG Corpus Resource** Next to the German LFG grammar, (Forst, 2003) has derived a ‘parallel’ LFG f-structure corpus from the TIGER treebank, by applying methods for treebank conversion. We make use of the parallel treebank to induce LFG frame annotation rules from the SALSA/TIGER annotations.

## 3 LFG for Frame Semantics

**Lexical Functional Grammar** (Bresnan, 2001) assumes multiple levels of representation. Most prominent are the syntactic representations of c(onstituent)- and f(unctional)-structure. The correspondence between c- and f-structure is defined by functional annotations of rules and lexical entries. This architecture can be extended to semantics projection (Halvorsen and Kaplan, 1995).

LFG f-structure representations abstract away from surface-syntactic properties, by localising arguments in mid- and long-distance constructions, and therefore allow for uniform reference to syntactic dependents in diverse syntactic configurations. This is important for the task of frame annotation, as it abstracts away from aspects of syntax that are irrelevant to frame (element) assignment.

In (1), e.g., the SELLER role can be uniformly associated with the *local* SUBJECT of *sell*, even though it is realized as (a.) a relative pronoun of *come* that controls the SUBJECT of *sell*, (b.) an implicit second person SUBJ, (c.) a non-overt SUBJ controlled by the OBLIQUE object of *hard*, and (d.) a SUBJ (*we*) in VP coordination.

- (1) a. *The woman who* had come in to *sell* flowers overheard their conversation.
- b. Don’t *sell* the factory to another company.
- c. It would be hard for *him* to *sell* newmont shares.
- d. .. *we* decided to sink some of our capital, buy a car, and *sell* it again before leaving.

**LFG Semantics Projection for Frames** As in a standard LFG projection architecture, we define a frame semantics projection  $\sigma_f$  from the level of f-structure. We define the  $\sigma_f$ -projection to introduce elementary frame structures, with attributes FRAME, FEE (frame-evoking element), and frame-specific role attributes. Figure 2 displays the  $\sigma_f$ -projection for the sentence in Figure 1.<sup>4</sup>

<sup>2</sup><http://www.coli.uni-sb.de/lexicon>

<sup>3</sup>See <http://www.icsi.berkeley.edu/~framenet>

<sup>4</sup>The MESSAGE role is coindexed with a lower frame, the frame projection introduced by the noun *Gespräch*.

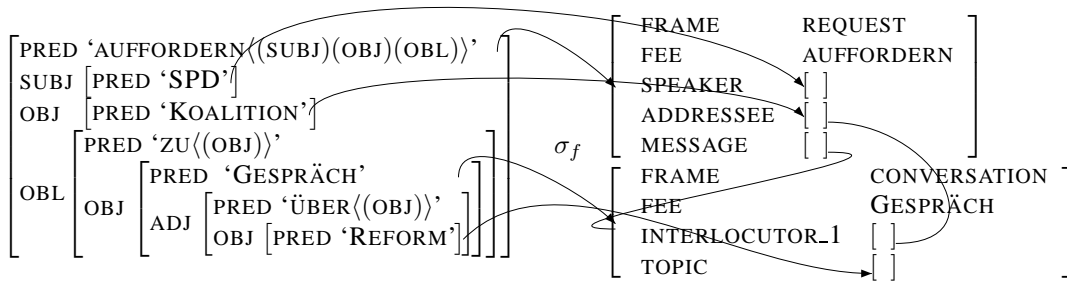


Figure 2: LFG projection architecture for Frame Semantics

auffordern V,  
 $(\uparrow \text{PRED}) = \text{'AUFFORDERN'} \langle (\uparrow \text{SUBJ}) (\uparrow \text{OBJ}) (\uparrow \text{OBL}) \rangle$   
 $\dots$   
 $(\sigma_f(\uparrow) \text{ FRAME}) = \text{REQUEST}$   
 $(\sigma_f(\uparrow) \text{ FEE}) = (\uparrow \text{ PRED FN})$   
 $(\sigma_f(\uparrow) \text{ SPEAKER}) = \sigma_f(\uparrow \text{ SUBJ})$   
 $(\sigma_f(\uparrow) \text{ ADDRESSEE}) = \sigma_f(\uparrow \text{ OBJ})$   
 $(\sigma_f(\uparrow) \text{ MESSAGE}) = \sigma_f(\uparrow \text{ OBL OBJ})$

Figure 3: Frame projection by co-description

Figure 3 states the lexical entry for the REQUEST frame.  $\sigma_f$  is a function of f-structure. The verb *auffordern* introduces a node  $\sigma_f(\uparrow)$  in the semantics projection of  $\uparrow$ , its local f-structure, and defines its attributes FRAME and FEE. The frame elements are defined as  $\sigma_f$ -projections of the verb's SUBJ, OBJ and OBL OBJ functions. E.g. the SPEAKER role, referred to as  $(\sigma_f(\uparrow) \text{ SPEAKER})$ , the SPEAKER attribute in the projection  $\sigma_f(\uparrow)$  of  $\uparrow$ , is defined as identical to the  $\sigma_f$ -projection of the verb's SUBJ,  $\sigma_f(\uparrow \text{ SUBJ})$ .

**Frames in Context** The projection of frames in context can yield connected frame structures. In Figure 2, *Gespräch* fills the MESSAGE role of REQUEST, but it also introduces a frame of its own, CONVERSATION. Thus, the CONVERSATION frame, by coindexation, is an instantiation, in context, of the MESSAGE of REQUEST.

**Co-description vs. description-by-analysis** In the co-description architecture we just presented f- and s-structure equations jointly determine the valid analyses of a sentence. Analyses that do not satisfy *both* f- and s-structure constraints are inconsistent and ruled out.

An alternative to co-description is semantics construction via description-by-analysis (DBA) (Halvorsen and Kaplan, 1995). Here, semantics is built on top of fully resolved f-structures. F-structures that are consistent with semantic mapping constraints are semantically enriched – remaining analyses are left untouched.

Both models are equally powerful – yet while co-

$\text{pred}(\text{X}, \text{auffordern}),$   
 $\text{subj}(\text{X}, \text{A}), \text{obj}(\text{X}, \text{B}), \text{obl}(\text{X}, \text{C}), \text{obj}(\text{C}, \text{D})$   
 $\Rightarrow$   
 $+ 's_f::'(\text{X}, \text{SemX}), + \text{frame}(\text{SemX}, \text{request}),$   
 $+ \text{fee}(\text{X}, \text{auffordern}),$   
 $+ 's_f::'(\text{A}, \text{SemA}), + \text{speaker}(\text{SemX}, \text{SemA}),$   
 $+ 's_f::'(\text{B}, \text{SemB}), + \text{addressee}(\text{SemX}, \text{SemB}),$   
 $+ 's_f::'(\text{D}, \text{SemD}), + \text{message}(\text{SemX}, \text{SemD}).$

Figure 4: Frame projection by DBA (via transfer)

description integrates the semantics projection into the grammar and parsing process, DBA keeps it as a separate module. Thus, with DBA, semantics does not interfere with grammar design and can be developed separately. The DBA approach also facilitates the integration of external semantic knowledge sources (such as word senses or named entity types).

**DBA by transfer** We realise the DBA approach by way of a term-rewriting transfer system that is part of the XLE grammar processing platform. The system represents f-structures as sets of predicates which take as arguments variables for f-structure nodes or atomic values. Transfer is defined as a sequence of ordered rules. If a rule applies to an input set of predicates, it defines a new output set. This output set is input to the next rule in the cascade. A rule applies if all terms on its left-hand side match some term in the input set. The terms on the right hand side (prefixed '+') are added to the input set. There are obligatory ( $\Rightarrow$ ) and optional ( $? \Rightarrow$ ) rules. Optional rules introduce two output sets: one results from application of the rule, the other is equal to the input set.

Figure 4 displays a transfer rule that corresponds to the co-description lexical entry of Figure 3. For matched f-structure nodes (pred, subject, object, oblique object) it defines a  $\sigma_f$ -projection (by predicate ' $s_f::$ ') with new s-structure nodes. For these, we define the frame information (FRAME, FEE) and the linking of semantic roles (e.g., the  $\sigma_f$ -projection SemA of the SUBJ is defined as the SPEAKER role of the head's semantic projection SemX).

Frame	FeeID	Role(s)	FeID(s)
Request	2 (from {2, 8})	Speaker Addressee Message	1 3 501

Figure 5: Core frame information for Fig. 1

```
% projecting frame information for FEE
project_fee(FeeID, Frame) ::
ti-id(X, FeeID), pred(X, Pred) ==>
+'s::'(X, S_X), +frame(S_X, Frame), +fee(S_X, Pred).
```

```
% semantic projection for (each) FE of FEE
project_fe_of_fee(FeeID, Frame, FeID, Role) ::
ti-id(X, FeeID), 's::'(X, S_X), frame(S_X, Frame),
ti-id(Y, FeID), pred(Y, Pred) ==>
+'s::'(Y, S_Y), +Role(S_X, S_Y), +rel(S_Y, Pred).
```

Figure 6: SALSAs-2-LFG-TIGER transfer

## 4 Corpus-based induction of an LFG frame semantics interface

### 4.1 Porting SALSAs annotations to LFG

A challenge for corpus-based induction of a syntax-semantics interface for frame assignment is the transposition of the corpus annotations from a given syntactic annotation scheme to the target syntactic framework. The basis for our work are annotations of the SALSAs/TIGER corpus (Erk et al., 2003), encoded in an XML annotation scheme that extends the syntactic TIGER XML annotation scheme.

The TIGER treebank has been converted to a parallel LFG f-structure corpus (Forst, 2003). The SALSAs/TIGER and LFG-TIGER corpora could be used to learn corresponding syntactic paths in the respective structures. Thus, we could establish the paths of frame constituting elements in the SALSAs/TIGER corpus, and port the annotations to the corresponding path in the LFG-TIGER corpus.

However, we could apply a more precise method, by exploiting the fact that the LFG-TIGER corpus preserves the original TIGER constituent identifiers, as f-structure features TI-ID (see Fig. 7). We use these 'anchors' to port the SALSAs annotations to the parallel LFG-TIGER treebank. Thus, in a first step we extend the latter to an LFG corpus with frame semantics projection. From the extended corpus we induce general LFG frame assignment rules. This will be described in more detail in Section 4.2.

**Porting annotations by transfer** For each sentence we extract the constituent identifiers of frame constituting elements in the SALSAs XML annotations (cf. Figure 5). This information is coded into transfer rules, where we refer to the corresponding TI-ID features in the f-structure as anchors to project the frame information for a given frame annotation

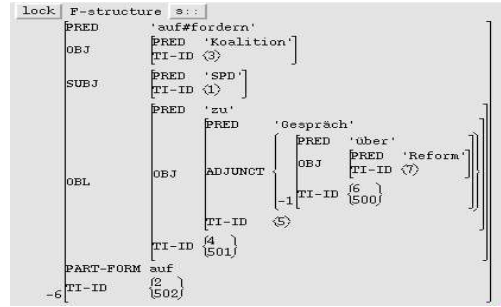


Figure 7: LFG-TIGER f-structure (w/ TI-ID)

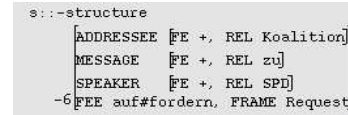


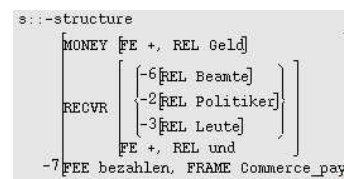
Figure 8: Frame projection from f-str of Fig. 7

instance. The first transfer rule (template) in Figure 6 defines the semantic projection of the FEE, where the correct f-structure location is referenced by the feature TI-ID. Subsequent rules – one for each role to be assigned – define the given semantic role as an argument of the FEE's semantic projection, again using the TI-IDs of the FEE and FE as anchors.

We generate these frame projection rules for each sentence in the SALSAs/TIGER corpus, and apply them to the corresponding f-structure in the LFG-TIGER corpus. The result is an LFG corpus with frame semantic annotations (cf. Figures 7 and 8).

The basic structure of frame-inducing rules in Figure 6 was refined to account for **special cases**:

**Coordination** For frame elements that correspond to coordinated constituents, as in Figure 9, we project a semantic role that records a *set* of semantic predicates (REL), one for each of the conjuncts.



*Beamten, Politikern und Geschäftsleuten wird Schmiergeld bezahlt – Clerks, politicians and businessmen are paid bribes*

Figure 9: Frame with coordinated RECVR role

**Underspecification** The SALSAs annotation scheme allows for *underspecification*, to represent unresolved word sense ambiguities or optionality (Erk et al., 2003). In a given context, a predicate may evoke alternative frames (i.e. word senses), where it is impossible to decide between them. E.g. the verb *verlangen* (demand) may convey the meaning of REQUEST, but also COMMERCIAL TRANSACTION. Such cases are annotated with

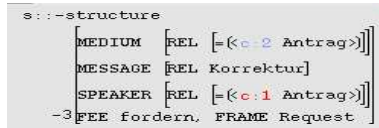
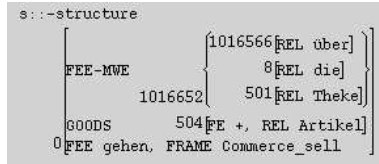


Figure 10: Underspecification as disjunction



4 Artikel gingen über die Ladentheke—4 items were sold  
 Figure 11: Multiword expressions

alternative frames, which are marked as elements of an ‘underspecification group’. Underspecification may also affect frame elements of a single frame. A motion (*Antrag*), e.g., may be both MEDIUM and SPEAKER of a REQUEST. Finally, a constituent may or may not be interpreted as a frame element of a given frame. It is then represented as a single element of an underspecification group.

We model underspecification as disjunction, which is encoded by optional transfer rules that create alternative (disjunctive) contexts. Optionality is modeled by a single optional rule. Figure 10 displays the result of underspecified frame element assignment in an f-structure chart (Maxwell and Kaplan, 1989). Context *c1* displays the reading where *Antrag* is assigned the SPEAKER role, alternatively, in context *c2*, it is assigned the role MEDIUM.

In a symbolic account disjunction doesn’t correctly model the intended meaning of underspecification. Yet, a stochastic model for frame assignment should render the vagueness involved in underspecification by close stochastic weights. Thus, underspecified annotation instances provide alternative frames in the training data and can be used for fine-grained evaluation of frame assignment models.

**Multiword Expressions** The treatment of multiword expressions (idioms, support constructions) requires special care. For idioms, the constituting elements are annotated as multiple frame evoking elements (cf. Figure 11 for *über die Ladentheke gehen* – go over the counter (being sold)). We define semantic projections for the individual components: the main frame evoking predicate (FEE) and the idiom-constituting words, which are recorded in a set-valued feature FEE-MWE. Otherwise, idioms are treated like ordinary main verbs. E.g., like *sell*, the expression triggers a COMMERCE\_SELL frame with the appropriate semantic roles, here GOODS.

**Asymmetric Embedding** Another type of non-isomorphism between syntactic and semantic rep-

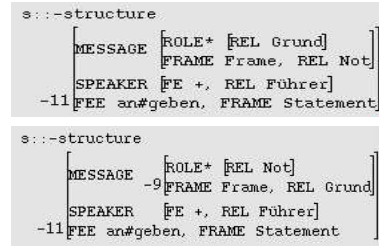


Figure 12: Asymmetric embedding (example (2))

resentation occurs in cases where distinct syntactic constituents are annotated as instantiation of a single semantic role. In (2), PP and NP are annotated as the MESSAGE of a STATEMENT, since they jointly convey its content. Projecting distinct constituents to a single semantic node can, however, lead to inconsistencies, especially if both constituents independently project semantic frames.

- (2) Der Geschäftsfuehrer gab [*PP-MO* als Grund fuer die Absage] [*NP-OBJ* Terminnoete] an.  
 The director mentioned [time conflicts] [as a reason for cancelling the appointment]

In the SALSA annotations *asymmetric* embedding at the semantic level is the typical pattern for such double-constituent annotations. I.e., for (2), we assume a target frame structure where the MESSAGE of STATEMENT points to the PP – which itself projects a frame REASON with semantic roles CAUSE for *Terminnöte*, and EFFECT for *Absage*.

Such multiple-constituent annotations arise in cases where frame annotations are *partial*: since corpus annotation proceeds frame-wise, in (2) the REASON frame may not have been treated yet. Moreover, annotators are in general not shown complete(d) sentence annotations.

We account for these cases by a simulation of *functional uncertainty* equations, which accommodate for a potential embedded frame within either one of the otherwise re-entrant constituents. We apply a transfer rule set that embeds one (or the other) of the two constituent projections as an *embedded* role of an unknown frame, to be evoked by the respective ‘dominating’ node. We introduce an ‘unknown’ role ROLE\* for the embedded constituent, which is to be interpreted as a functional uncertainty path over variable semantic roles.

Figure 12 displays the alternative (hypothetical) frame structures for (2), where the second one – with FRAME instantiated to REASON and ROLE\* to CAUSE – corresponds to the actual reading.

**Overview of data** Our current data set comprises 12436 frame annotations for 11934 sentences. Table 1 gives frequency figures for the special phe-

	coord	usp	mwe	asym	>dbl	all
abs	467	395	1287	421	97	12436
in %	3.76	3.18	10.34	3.39	0.78	100

Table 1: Overview of special annotation types

nomena: coordination, underspecification, multi-word expressions and double constituents (asym).<sup>5</sup>

We successfully ported 11713 frame annotations to the LFG-TIGER corpus, turning it into an LFG corpus with frame annotations.

## 4.2 Inducing frame projection rules

From the enriched corpus we extract lexical frame assignment rules that – instead of node identifiers – use f-structure descriptions to identify constituents and map them to frame semantic roles. These rules can then be applied to the f-structure output of free LFG parsing, i.e. to novel sentences.

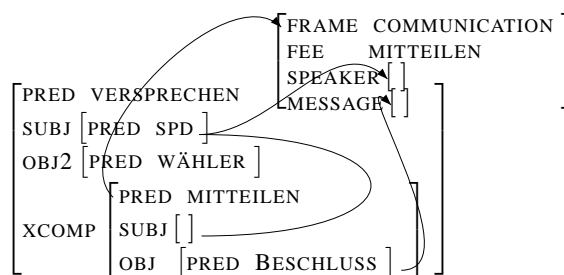
We designed an algorithm for extracting f-structure paths between pairs of f-structure nodes that correspond to the s-structure of the frame evoking element and one of its semantic roles, respectively. Table 2 gives an example for the frame projection in Figure 13. Starting from the absolute f-structure path (f-path) for (the f-structure projecting to) the FEE MITTEILEN we extract relative f-paths leading to the roles MESSAGE and SPEAKER. The f-path for the MESSAGE ( $\uparrow$ OBJ) is *local* to the f-structure that projects to the FEE. For the SPEAKER we identify two paths: one local, the other non-local. The *local* f-path ( $\uparrow$ SUBJ) leads to the local SUBJ of *mitteilen* in Figure 13. By co-indexation with the SUBJ of *versprechen* we find an alternative *non-local* path, which we render as an inside-out functional equation ((XCOMP $\uparrow$ ) SUBJ).

Since f-structures are directed acyclic graphs, we use graph accessibility to distinguish local from non-local f-paths. In case of alternative local and non-local paths, we choose the local one. From alternative non-local paths, we chose the one(s) with shortest inside-out subexpression.

**Generating frame assignment rules** We extracted f-path descriptions for frame assignment from the enriched LFG-TIGER corpus. We compiled 9707 lexicalised frame assignment rules in the format of Figure 4. The average number of distinct assignment rules per FEE is 8.38. Abstracting over the FEEs, we obtain 7317 FRAME-specific rules, with an average of 41.34 distinct rules per frame.

Due to the surface-oriented TIGER annotation format, the original annotations contain a high number of non-local frame element assignments that

<sup>5</sup>Role assignment to more than two constituents (>dbl) constitute a rather disparate set of data we do not try to cover.



SPD verspricht Wählern, Beschlüsse mitzuteilen  
SPD promises voters to report decisions

Figure 13: Local and non-local frame elements

	absolute f-path	relative f-path	
FEE	XCOMP <u>PRED</u>	$\uparrow$	
MSG	XCOMP <u>OBJ</u>	$\uparrow$ OBJ	local
SPKR	SUBJ	(XCOMP $\uparrow$ )SUBJ	nonlocal
	XCOMP <u>SUBJ</u>	$\uparrow$ SUBJ	local

Table 2: Local and nonlocal path equations

are localised in LFG f-structures. The f-paths extracted from the enriched LFG corpus yield 12.82% non-local (inside-out) vs. 87.18% local (outside-in) frame element assignment rules.

As an alternative rule format, we split frame assignment into separate rules for projection of the FEE and the individual FEs. This allows assignment rules to apply in cases where the f-structure does not satisfy the functional constraints for some FE. This yields improved robustness, and accounts for syntactic variability when applied to new data. For this rule format, we obtain 960 FEE assignment rules, and 8261 FEE-specific FE assignment rules. Abstracting over the FEE, this reduces to 4804 rules.<sup>6</sup>

## 4.3 Reapplying frame assignment rules

We reapplied the induced frame assignment rules to the original syntactic LFG-TIGER corpus, to control the results. The results are evaluated against the frame-enriched LFG-TIGER corpus that was created by explicit node anchoring (Sec. 4.1). We applied 'full frame rules' that introduce FEE and all FEs in a single rule, as well as separated FEE and FE rules. We applied all rules for a given frame to any sentences that had received the same frame in the corpus. We obtained 93.98% recall with 25.95% precision (full frame rules), and 94.98% recall with 45.52% precision (split rules), cf. Table 3.a. The low precision is due to overgeneration of the more general abstracted rules, which are not yet controlled by statistical selection. We measured an ambiguity of 8.46/7.83 frames per annotation instance.

<sup>6</sup>In the future we will experiment with assignment rules that are not conditioned to FEEs, but to frame-specific syntactic descriptions, to assign frames to 'unknown' lexical items.

	full frame rules			FEE and FE rules		
	rec	prec	amb	rec	prec	amb
(a)	93.98	25.95	8.46	94.98	45.52	7.83
(b)	52.21	6.93	13.35	76.41	18.32	9.00

Table 3: Evaluation of annotation results:

(a) on TIGER corpus, (b) on LFG parses

## 5 Applying frame assignment rules in an LFG parsing architecture

We finally apply the frame assignment rules to original LFG *parses* of the German LFG grammar. The grammar produces f-structures that are compatible with the LFG-TIGER corpus, thus the syntactic constraints can match the parser’s f-structure output. In contrast to the LFG-TIGER corpus, the grammar delivers f-structures for alternative syntactic analyses. We don’t expect frame projections for all syntactic readings, but where they apply, they will create ambiguity in the semantics projection.

We applied the rules to the parses of 6032 corpus sentences. Compared to the LFG-TIGER corpus we obtain lower recall and precision (Table 3.b) and a higher ambiguity rate per sentence. Drop in precision and higher ambiguity are due to the higher ambiguity in the syntactic input. Moreover, we now apply the complete rule set to any given sentence. The rules can thus apply to new annotation instances, and create more ambiguity. The drop in recall is mainly due to overgenerations by automatic lemmatisation and functional assignments to PPs in the TIGER-LFG corpus, which are not matched by the LFG parser output. These mismatches will be corrected by refinements of the TIGER-LFG treebank.

## 6 Summary and Future Directions

We presented a method for corpus-based induction of an LFG syntax-semantics interface for frame semantic processing. We port frame annotations from a manually annotated corpus to an LFG parsing architecture that can be used to process unparsed text. We model frame semantic annotations in an LFG projection architecture, including phenomena that involve non-isomorphic mappings between levels.

In future work we will train stochastic models for disambiguation of the assigned frame semantic structures. We are especially interested in exploring the potential of deeper, functional syntactic analyses for frame assignment, in conjunction with additional semantic knowledge (e.g. word senses, named entities). We will set up a bootstrapping cycle for learning increasingly refined stochastic models from growing training corpora, using semi-supervised learning methods. We will explore multi-lingual aspects of frame assignment, using

English FrameNet data and an English LFG grammar with comparable f-structure output. Finally, we will investigate how similar methods can be applied to syntactic frameworks such as HPSG, which already embody a level of semantic representation.

**Acknowledgements** We thank the IMS Stuttgart for allowing us to use the German LFG grammar. Special thanks go to Martin Forst who provided us with the TIGER-LFG corpus and added special features to support our work. Finally, thanks go to Dick Crouch who greatly enhanced the transfer system.

## References

- C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL 1998*, Montr´eal, Canada.
- S.Brants, S.Dipper, S.Hansen, W.Lezius, G.Smith. 2002. The TIGER Treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.
- J. Bresnan. 2001. *Lexical-Functional Syntax*. Blackwell Publishers, Oxford.
- K. Erk, A. Kowalski, S. Pad´o, and M. Pinkal. 2003. Towards a Resource for Lexical Semantics: A Large German Corpus with Extensive Semantic Annotation. In *Proceedings of the ACL 2003*, Sapporo, Japan.
- M. Fleischman, N. Kwon, and E. Hovy. 2003. Maximum entropy models for FrameNet classification. In *Proceedings of EMNLP’03*, Sapporo, Japan.
- M. Forst. 2003. Treebank Conversion – Establishing a testsuite for a broad-coverage LFG from the TIGER treebank. In A. Abeill´e, S. Hansen, and H. Uszkoreit (eds), *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora*, Budapest.
- A. Frank and K. Erk. 2004. Towards an LFG Syntax–Semantics Interface for Frame Semantics Annotation. In A. Gelbukh (ed), *Computational Linguistics and Intelligent Text Processing*, Springer, Heidelberg.
- D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3).
- D. Gildea and M. Palmer. 2002. The Necessity of Parsing for Predicate Argument Recognition. In *Proceedings of ACL’02*, Philadelphia, PA.
- P.-K. Halvorsen and R.M. Kaplan. 1995. Projections and Semantic Description in Lexical-Functional Grammar. In M. Dalrymple, R.M. Kaplan, J.T. Maxwell, A. Zaenen (eds), *Formal Issues in Lexical-Functional Grammar*, CSLI Lecture Notes, Stanford.
- P. Kingsbury, M. Palmer, and M. Marcus. 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of the HLT Conference*, San Diego.
- J. T. III Maxwell and R. M. Kaplan. 1989. An overview of disjunctive constraint satisfaction. In *Proceedings of IWPT*, pages 18–27.
- S. Riezler, T. H. King, R. M. Kaplan, R. Crouch, J. T. III Maxwell, and M. Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the ACL’02*, Philadelphia, PA.