Generalisations over Corpus-induced Frame Assignment Rules

Anette Frank

German Research Center for Artificial Intelligence
DFKI GmbH
Saarbrücken, Germany
frank@dfki.de
and
Computational Linguistics Department
Saarland University
Saarbrücken, Germany
frank@coli.uni-sb.de

Abstract

In this paper we discuss motivations and strategies for generalising over instance-based frame assignment rules that we extract from frame-annotated corpora. Corpus-induced syntax-semantics mapping rules for frame assignment can be used for automatic semantic role labelling of unparsed text, but further, to extract linguistic knowledge for a lexical semantic resource with a general syntax-semantics interface. We provide a data analysis of a comprehensive rule set of corpus-induced frame assignment rules, and discuss the potential of applying different types of generalisations and filters, to obtain a uniform extended data set for the extraction of linguistic knowledge.

1. Introduction

Various research groups are currently concerned with the creation of large-scale lexical semantic resources that provide information about predicate-argument structure. The Berkeley FrameNet project (Baker et al., 1998), following Fillmore's theory of frame semantics (Fillmore, 1976), is building a large semantic lexicon, including the definition of frames and semantic roles, and a corpus of manually annotated sentences. A strictly corpus-based approach is carried out with 'PropBank' (Kingsbury et al., 2002) – a manual semantic role annotation on top of the PennII Treebank.

There are first approaches for learning stochastic models for semantic role assignment from annotated corpora; e.g. (Gildea and Jurafsky, 2002; Fleischman et al., 2003). Probabilistic models for semantic role assignment systems will eventually be used for automated semantic annotation in NLP applications, but they can also be used, in a bootstrapping architecture, to learn increasingly refined probabilistic models from extended training sets, by application of meta-learning strategies, such as active learning.

The current models for stochastic role assignment models are essentially corpus-based. Yet, besides the development of systems for automated role labelling, there is also interest in a general lexical semantics resource that can be formalised and integrated into alternative NLP systems.

In our work we investigate techniques for automated induction of rules for automatic semantic role assignment from semantically annotated corpora.[1] In this paper we discuss strategies for generalising over corpus-induced frame assignment rules. We provide a data analysis of a comprehensive rule set, and discuss the potential of applying different types of generalisations and filters, to obtain a uniform extended data set – for semi-automatic acquisition of new training data, and the extraction of linguistic knowledge.

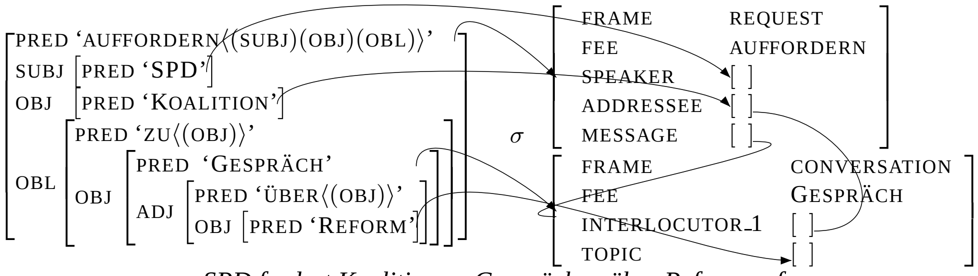
2. Deep syntactic analysis for semantic role labelling

Since semantic role assignment is based on a syntactic annotation layer, automated processing for semantic role assignment on unparsed text requires an interface between a syntactic analyser and the targeted semantic annotation. Current competitions explore the potential of shallow parsing as a basis for semantic role labeling. However, (Gildea and Palmer, 2002) have emphasised the role of deeper syntactic analysis for semantic role assignment. We follow this line, and explore the potential of deep syntactic analysis for semantic role labelling, choosing Lexical Functional Grammar (Bresnan, 2001) as underlying syntactic framework.

In a first study, (Frank and Erk, 2004) discuss advantages of semantic role assignment on the basis of functional syntactic analyses as provided by LFG parsing, and present an LFG projection architecture for frame semantics. In this architecture, frames are projected from f-structure representations, as displayed in Figure 1. The semantic projection is defined by lexical entries of frame evoking predicates, which map f-structure nodes for grammatical functions to frame semantic roles in a frame semantics projection. The projection of frames in context can yield partially connected frame structures. In Figure 1, *Gespräch* projects to 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. Figure 2 displays how these mappings can defined in a classical LFG co-description projection architecture, by use of functional descriptions; see (Frank and Erk, 2004) for details.

As an alternative to the co-description approach, we implemented frame projection in a description-by-analysis (DBA) architecture. In co-description, semantics projection is tightly intervowen with grammar definitions and the parsing process. The DBA approach, by contrast, is more

[1] The work is conducted in the context of the SALSA project; see (Erk et al., 2003) and http://www.coli.uni-sb.de/lexicon.



SPD fordert Koalition zu Gesprächen über Reform auf.
'SPD requests coalition to talk about Reform'

Figure 1: LFG projection architecture for Frame Annotation

auffordern V,
(↑PRED)='AUFFORDERN⟨(↑SUBJ)(↑OBJ)(↑OBL)⟩'
...
(σ(↑) FRAME) = REQUEST
(σ(↑) FEE) = (↑ PRED FN)
(σ(↑) SPEAKER) = σ(↑ SUBJ)
(σ(↑) ADDRESSEE) = σ(↑ OBJ)
(σ(↑) MESSAGE) = σ(↑ OBL OBJ)

Figure 2: Frame projection in lexical entry (co-description)

```
pred(X,auffordern),
subj(X,A), obj(X,B), obl(X,C), obj(C,D)
==>
+'s::'(X,SemX), +frame(SemX,request), +fee(X,auffordern)
+'s::'(A,SemA), +speaker(SemX,SemA),
+'s::'(B,SemB), +addressee(SemX,SemB),
+'s::'(D,SemD), +message(SemX,SemD).
```

Figure 3: Frame projection rule (as a transfer rewrite rule)

modular. Here, frame projection rules apply to completed f-structure representations produced by the LFG parser.

The DBA approach is realised by use of a transfer rewrite system.[2] The system allows the definition of rewrite rules that apply to an f-structure context and introduce, on their right-hand side, a semantic projection for frames: the specific FRAME evoked by the frame evoking element (FEE), i.e., the triggering predicate in the f-structure. The rules further define the projection of frame-specific semantic roles from particular local (or sometimes non-local) functional paths (such as SUBJ, OBJ, OBL OBJ), starting from the f-structure node of the frame evoking predicate. The example of Figure 3 is equivalent to the co-description variant in Figure 2, and thus yields the same frame projection, displayed in Figure 1.

3. Corpus-based induction of an LFG–frame semantics interface

(Frank and Semecky, 2004) present a method for the automatic induction of LFG-based frame assignment rules from semantically annotated corpora. This method was first applied to the SALSA corpus (Erk et al., 2003), a German newspaper corpus enriched with frame semantic annotations. The SALSA annotations are built on, and extend the syntactically annotated TIGER corpus (Brants et al., 2002). In (Frank and Semecky, 2004) the frame semantic annotations of the SALSA/TIGER corpus were ported to a 'parallel' TIGER corpus of corresponding LFG f-structure analyses (Forst, 2003). Figure 3 displays an example of a frame assignment rule that was extracted from the resulting frame-extended LFG SALSA/TIGER corpus. (Frank and Semecky, 2004) further present first experiments to apply the resulting computational syntax-semantics interface for frame semantics in an LFG parsing architecture, using a wide-coverage LFG grammar of German.[3]

A similar architecture for corpus-based induction of a frame semantics interface was recently developed in the context of the Senseval-3 task on semantic role labeling for English.[4] Here, the basis was a subset of the English frame annotated sentences of the FrameNet project (Baker et al., 1998), and the wide-coverage stochastic English LFG grammar developed at Parc (Riezler et al., 2002). The grammar provided a 'parallel' LFG corpus with most-probable analyses for the annotated sentences. Similar to the methods applied for SALSA/TIGER, we port the frame annotations to the LFG parsed sentences, and extract frame assignment rules that can be applied to new sentences in an LFG parsing-transfer architecture.

In both scenarios, the next steps towards an automated system for LFG-based frame assignment involve the design of probabilistic models to select the most probable frame assignments from the choice of possible assignments that are generated by application of the corpus-induced frame assignment rules proper – as well as generalisations of these rules, which account for unseen configurations.

Besides the development of a probabilistic semantic role labelling system, the aim of the SALSA project is to acquire generalised linguistic knowledge, i.e. a frame semantic lexicon with a well-defined syntax-semantics interface, from a large frame-annotated German corpus. It is also in view of this more ambitious aim that we are concerned with a closer inspection of the corpus-induced syntax-semantic mapping rules for frame assignment.

[2] The system comes as a module of the grammar development platform XLE (http://www2.parc.com/istl/groups/nltt/). It was designed and implemented by Martin Kay (Xerox Parc) for a Machine Translation prototype; see (Frank, 1999). Recent enhancements to the system were realised by Richard Crouch.

[3] The German LFG grammar is being developped at the IMS, University of Stuttgart.

[4] This was done in joint work with Katrin Erk and Ulrike Baldewein.

4. Generalisations over corpus-induced frame assignment rules

In this section we discuss motivations and strategies for generalising over sets of instance-based frame assignment rules that we extract from frame annotated corpora. In Section 5 we provide a quantitative evaluation of the rule set we extracted from the English FrameNet corpus sentences that were provided as training data in the Senseval-3 semantic role labeling task.

On the basis of this evaluation, Section 6 reviews the potential of the proposed generalisations over corpus-induced frame assignment rules: for abstraction of a general linguistic knowledge base, and for the targeted acquisition of training material in an active learning scenario, to develop increasingly refined stochastic models for frame assignment on the basis of continuously extended training corpora.

4.1. Motivations

Corpus-based extraction of frame assignment rules is confronted with two problematic issues: quality and coverage.

Quality It is well-known from treebank-based grammar induction that corpus-based acquisition and formalisation of linguistic knowledge is confronted with the problem of noise in the data. In our case, noise can be imported from various sources: (i) mistakes and inconsistencies in the manual syntactic or semantic annotations; (ii) problems in the automated mapping from corpus specific syntactic annotation schemes to the LFG f-structure encoding; (iii) problems in the extraction of frame assignment rules from the frame-enriched LFG corpora, and finally (iv) parsing errors or missing coverage of the underlying LFG grammars.

Coverage The problem of coverage is specific to the nature of lexical semantic corpus annotation. Lexical semantic annotation is confronted with a severe sparse data problem, since we may not encounter a large-enough variety of predicates in specific senses and constructions within manageable sizes of manually annotated corpora. E.g., while the SALSA corpus is comparable, in size, to the Penn Treebank, of the 4185 verbs (types), 1457 (34.81%) occur only once, and 3307 (79.02%) occur with frequency 1-10.

This sparse data problem is even more serious if we consider, as we do in SALSA, semi-automatic annotation of new corpus instances and learning of a principled syntax-semantics interface from corpus annotations: since there are multiple sources of noise in the data (see above), we may miss out a number of (already rare) corpus instances.

’Filling Gaps’ In order to address these problems, we investigate the potential of various generalisations or ’filters’ over instance-based rule sets, which can be used to identify and ‘fill gaps’ in the base of corpus samples.

Targeted acquisition of new corpus data to fill these gaps will enable the extraction of more homogeneous syntax-semantics mapping constraints for the final semantic lexcion resource. Most importantly, though, this way of acquiring new corpus material can be used to support active learning techniques, by providing a selection of ‘informative’ novel annotation instances, i.e. novel training instances that are promising candidates for improving stochastic models for automated frame assigment.

In the following we present different aspects of generalisations over corpus-based frame annotation instances. These range from linguistically motivated generalisations to distributional criteria regarding the densitiy of annotation samples for different classes of annotation events.

4.2. Linguistic generalisations

LFG f-structures provide a level of representation that abstracts away from surface-syntactic variations that are irrelevant for frame assignment (such as word order, long-distance phenomena or coordination). On the other hand, f-structures are genuine *syntactic representations* that differ from semantic predicate argument structures in that they do represent *functional syntactic* variants that are not distinguished in the semantic representation.

Diatheses A prominent example is the active-passive diathesis. Due to the sparseness of data we encounter with current sizes of annotated corpora, we may or may not encounter both active and passive constructions for a given frame evoking predicate and its specific semantic role configuration. This ‘gap’ in the training data may be compensated by the use of a greater variety of features in stochastic modelling for role assignment, but the lack of generalisation will be problematic for automated methods in building a final lexicon resource from the corpus-induced rule sets.

In order to fill such gaps in the training corpus we can generate missing active or passive variants of frame projection rules, and apply them to candidate sentences extracted from unparsed corpora. Sentences that receive the targeted annotation can be presented to annotators for acknowledgement, and – on approval – can be added to the set of training samples. On the basis of the extended corpus, we can extract more general frame assignment rules, with disjunctive constraints to account for active and passive constructions (see Figure 4). This will lead to a more homogeneous frame semantic lexicon resource, and will increase the coverage of automated frame assignment models when applied to unseen text.

```
pred(X,auffordern),
{ passive(X,−), subj(X,A), obj(X,B)
| passive(X,+), subj(X,B), obl_ag(X,A) },
obl(X,C), obj(C,D)    ==>
+’s::’(X,SemX), +frame(SemX,request), +fee(X,auffordern)
+’s::’(A,SemA), +speaker(SemX,SemA),
+’s::’(B,SemB), +addressee(SemX,SemB),
+’s::’(D,SemD), +message(SemX,SemD).
```

Figure 4: Generalisation over active-passive diathesis

Non-local frame element assignments Another source of gaps in the annotation samples are frames that occur in non-local syntactic contexts. In case the evoking predicate is not, alternatively, found in a local syntactic context, the extracted rules will not be able to annotate the same frame in a more general, local context.

The LFG formalism provides a significant capacity for argument localisation (in long-distance, coordination, raising and control constructions). However, there are constructions where arguments cannot be localised on syntactic grounds. A classical example are constructions involving anaphoric control, such as gerunds.

In example (1), from the FrameNet data, the THEME role of the frame evoking predicate, *disappear*, was annotated as the passive SUBJ of the main clause, while the FEE is contained in the clausal ADJUNCT phrase (cf. Figure 5), while the local subject of the adjunct clause is a non-overt pronominal SUBJ. The functional path from the f-structure node of the frame evoking predicate to the f-structure of the THEME role is inside-out and non-local: ((ADJUNCT $ ↑) SUBJ).[5] Starting out from the local f-structure ↑ of the frame evoking element *disappear* the path leads inside-out via the set-valued ADJUNCT function to the dominating node (ADJUNCT $ ↑). From this node, the path leads outside-in via the function SUBJ to the f-structure of *sword*.

(1) *The Solland Sword was lost for many years, having disappeared during the destruction of Solland by Gorbad Ironclaw's Orcs .*

Similar to the active-passive distinction, in cases were our rule set does not comprise the corresponding local variant of the identified non-local frame assignment rule, we can generate an alternative local assignment rule, here looking for a local SUBJ of the frame evoking predicate in active voice. We can use such rules to automatically annotate sentences from unparsed corpora, again presenting the targeted instances to annotators for acknowledgement. With this method, we systematically extend the set of general, local frame assignment rules.

The identified patterns of typical non-local path descriptions can, moreover, serve as a 'functional bridge' in non-local annotation contexts. That is, we can state generic frame assigning rules that account for such 'bridging' non-local functional paths for frame element assignment. These can be triggered as fallback rules, to identify novel annotation instances in non-local configurations.

4.3. Abstractions from frame assignment rules

Finally, we can apply similar methods for acquiring novel annotation instances, by analysing the distribution of role assignments for a given frame, abstracting over the specific frame evoking elements that were found to invoke the frame. That is, from the FEE-specific annotations in the corpus we abstract classes of 'non-lexicalised frames' with syntactic mapping constraints. We can apply these generic frame assignment rules to novel corpus instances, where we condition the application to the set of FEEs that can trigger the given frame. We will further experiment with frame assignment rules that define clusters (instead of specific instances) of role-preposition correspondences.

5. Investigating corpus-induced samples of frame assignment rules

In this section we provide a data analysis of LFG frame assignment rules that we acquired from frame-annotated corpora. For this analysis, we concentrate on the rule set we induced from the FrameNet corpus data (Section 3).[6]

5.1. Coverage

Due to the lexicographic approach of the FrameNet project, the English FrameNet data can be assumed to be rather homogeneous and balanced as to the quantitative distribution of frame evoking predicates and their constructional variants. By contrast, the mapping from the FrameNet annotations to LFG representations is currently based on the most probable analysis of the English LFG grammar, which may still feature wrong selections. Moreover, a number of frame element bracketings in the FrameNet annotations did not map to a unique f-structure node in the corrsponding LFG analysis, and hence did not yield frame assignments in the LFG-based frame-enriched corpus.[7]

These (interrelated) challenges are reflected in the coverage figures of Table 1, with 90.19% of sentences that receive frame element annotations, yet only 67.41% coverage at the level of overall frame element assignments, measured against the target annotations in the original FrameNet corpus.[8] We obtain 1.77 frame element assignments per sentence in average, against 2.33 in the FrameNet data.

	abs no	in %	avg/s
s(entences)	24274	100	-
s with extracted fpaths	21893	90.19	-
target fes	57325	100	2.33 fe/s
extracted fpaths for fes	38643	67.41	1.77 fe/s

Table 1: Coverage: extracted fe-assignment paths

Table 2 gives an overview of the distribution of different *types* of functional path equations (fpaths) that lead from (the f-structure of) the frame evoking element (FEE) to (the f-structure of) its frame element (or semantic role) – for distinct FEEs, or abstracting over the FEE of a given frame. As expected, taking the assigned semantic roles into account (in fpath-role) leads to a greater variety of distinct fpath-role assignments, both for FEE-specific and – proportionally higher – for frame-specific assignment paths.

per FEE	all	min	max	avg.
fpath	11465	1	67	8.10
fpath-role	13477	1	79	9.52

per Frame	all	min	max	avg.
fpath	4211	22	292	105.28
fpath-role	5497	24	385	137.43

Table 2: Distribution of fpath types (per FEE, per Frame)

5.2. Active-passive diathesis

The above figures are not really informative as to how complete the distribution of the acquired frame assignment

[5] ADJUNCTs are represented as set-valued f-structures. In functional path descriptions, reference to an element of a set is made by the path symbol '$' for 'in_set'.

[6] As the SALSA corpus is still under construction, our rule set is considerably smaller, and relatively unbalanced over frames. A data analysis on the basis of the more balanced and sufficiently varied FrameNet data therefore seemed to prove more indicative.

[7] We will further improve the mapping procedures from corpus annotations to LFG parses, so we expect the figures to improve.

[8] We lost 284 sentences of the original corpus that we could not map to f-structures for technical reasons. These sentences have not been subtracted from the FrameNet data counts in Table 1.

rules is for specific syntactic variants (i.e. fpath-role assignments) over the different classes – whether FEEs or frames.

A closer look is provided by Table 3, for the distributional patterns of fpath-role assignments in active-passive alternations. Almost half of the verb types do only appear in either active or passive constructions - and it is not clear from the counts whether there are missed-out alternations, or whether there are genuinely non-alternating verbal predicates.[9] Moreover, as is seen on the right-hand side, the proportion of local (subj,obj, obl_ag) fpaths found in active and passive constructions is very low (11.89–15.09% for active, and 12.09-20.48% for passive constructions).

Table 4 views the active-passive alternation from a different angle, by looking at passive-invariant semantic roles, i.e. the roles whose functional path assignment is (for given a frame, or a given FEE) never affected by the active-passive alternation. The frequency of such invariant fpath-role pairs (i.e. identical fpath-role assignments in a passive and active constructions) is very low.

verbs (types)		all vs. local fpaths			
		active		passive	
nonfragmented	590	all fp	7118	all fp	3028
active/passive	321	subj	1072	subj	620
passive only	24	obj	846	obl_ag	366
active only	245	obl_ag	2	obj	4

Table 3: Active-passive diathesis: distribution and fpaths

	all	passive-invariant	
FEE-fpath-role	4827	224	4.64%
Frame-fpath-role	2210	206	9.32%

Table 4: Passive-invariant fpath-role assignments

Closer inspection of the data underlying Table 4 shows that many fpath-role pairs are wrongly classified as passive-invariant due to a rare active or passive occurance that is produced by noise in the data (e.g. a wrong parse). Typical examples of such misclassifications are cases like *mumble*, occurring with SUBJ-SPEAKER assignment in both active and passive, yet with a distribution of 28 vs. 3. While these are rather clear weighted distributions, there are cases where the distribution is more unmarked (e.g. *murder* with a SUBJ-VICTIM distribution of 1 vs. 3 active vs. passive occurrences), and thus become difficult to distinguish from correct, but still infrequent distributions of correct instances of passive-invariant fpath-role pairs, in particular adjuncts.

This kind of noise in the data does clearly not only affect the identification of passive-invariant fpath-role assignments, but also the identification of active-passive alternating verbs in Table 3. That is, we observe a high number of instances that are identified as active-passive alternating, but on the basis of erroneous active or passive occurrences.

Filtering noise In order to filter such misclassifications, we computed a confidence weight for fpath-role assignments on the basis of their proportional distribution in passive vs. active assignments. The weight for a given fpath-role assignment in an active or passive construction, respectively, is computed by its relative frequency wrt. the overall number of fpath-role assignments in the respective voice, for a given FEE (or frame). This value we then used to experiment with different thresholds for computing counts on the active-passive distribution of fpath-role assignments.

As seen in Table 5, this filter reduces the number of active-passive alternating verb (type)s, by filtering erroneous instances from the base of counts. While the number of instances drastically reduces, only a small number of verb types are eliminated from consideration. On the other hand, the proportion of correct *local* functional subcategorisation paths in the retained set of fpath-role assignments increases with the threshold. For active verbs, the culmination point for positive filtering effects seems to be around .6. For passive verbs, we obtain the best filtering effect for subj with threshold .6, and for obl_ag with .7. Thus, the filters eliminate erroneous or otherwise rare occurrences.

verbs (types)		all vs. local fpaths					
		active			passive		
nonfrag	590	all fp	7118	in %	all fp	3028	in %
act/pass	321	subj	1072	15.06	subj	620	20.48
pass only	24	obj	846	11.89	obl_ag	366	12.09
act only	245	obl_ag	2		obj	4	
thresh .6	581	all fp	1470		all fp	741	
act/pass	309	subj	386	26.26	subj	211	28.48
pass only	25	obj	332	22.59	obl_ag	167	22.58
act only	247	obl_ag	1		obj	1	
thresh .7	580	all fp	1470		all fp	677	
act/pass	307	subj	386	26.26	subj	166	24.52
pass only	24	obj	332	22.59	obl_ag	160	23.63
act only	249	obl_ag	1		obj	1	

Table 5: Filters on active-passive diathesis

As a filter of noise in the computation of passive-invariant fpath-role assignments, we compute a weight for each fpath-role pair based on the relative frequency of passive as opposed to active occurrences (per FEE or frame). As seen in Table 6, this results in a radical reduction of passive-invariant fpath-role assignments, since many fpath-role occurrences do not show a sufficiently unbalanced distribution over active and passive, and thus do not exceed the threshold. This holds in particular for adjuncts and obliques which are clearly non-alternating functions. Selected application of the filter to functions that participate in the active-passive alternation, such as SUBJ and OBJ, shows moderate filtering effects that produce satisfactory results.[10]

threshold (.6)	filter on all fpaths		filter on subj/obj	
FEE-fpath-role	141/224	62.95	54/71	76.06%
frame-fpath-role	157/206	76.21	69/82	84.15%
threshold (.7)				
FEE-fpath-role	86/224	38.39	40/71	56.34%
frame-fpath-role	110/206	53.40	52/82	63.41%

Table 6: Filters on passive-invariant fpath-roles

[9] We only consider verbs whose functional context is not affected by fragmentary parses (nonfragmented).

[10] We will further experiment with weights that are parameterised for specific functional roles and patterns of argument structure variation, along the lines of (Merlo and Stevenson, 2001).

	all (w/o fragmented)				outside-in				inside-out (and outside-in)			
	abs	in %	types	in %	abs	in %	types	in %	abs	in %	types	in &
all lengths	38034	100/100	1582	100/100	31568	83/100	431	27/100	6466	17/100	1151	73/100
length 1	27567	72.48	97	6.13	27567	87.33	97	22.51	0	0.00	0	0.00
length 2	5967	15.69	218	13.78	3577	11.33	75	17.40	2390	36.96	143	12.42
length 3	3460	9.10	610	38.56	314	0.99	158	36.66	3146	48.65	452	39.27
length 4	820	2.16	456	28.82	63	0.20	57	13.23	757	11.71	399	34.67
length 5	187	0.49	169	10.69	47	0.15	44	10.21	140	2.17	125	10.86
length 6	29	0.08	28	1.77	0	0.00	0	0.00	29	0.45	28	2.43
length 7	4	0.00	4	0.25	0	0.00	0	0.00	4	0.00	4	0.35

Table 7: Path types

outside-in		inside-out (and outside-in)	
path	frequency	path	frequency
↑	9213	((OBJ ADJUNCT $ ↑) OBJ)	548
(↑ SUBJ)	5030	((OBJ $ ↑) SUBJ)	497
(↑ SPEC POSS)	3228	((ADJUNCT $ ↑) SUBJ)	240
(↑ OBJ)	3176	((SUBJ ADJUNCT $ ↑) SUBJ)	228
(↑ ADJUNCT)	2835	(($ ↑) $)	195
(↑ MOD)	2556	(($ ADJUNCT $ ↑) $)	160
(↑ ADJUNCT_OF)	1001	(($ OBJ ↑) $)	135
(↑ OBL_AG)	499	((ADJUNCT $ ↑) $)	133
(↑ ADJUNCT_IN)	314	(($ OBJ ↑) SUBJ)	123
(↑ OBL_WITH)	297	((XCOMP ADJUNCT $ ↑) XCOMP)	121

Table 8: Top ten frequent path types

5.3. Local and non-local frame assignment paths

Another issue that affects the homogeneity of the corpus-induced syntax-semantics interface for frame semantics is the nature and variety of functional paths that are extracted from frame-annotated sentences. As seen in Table 5, only a small proportion of fpaths involved in active-passive alternations is found to be local, i.e. involve a locally subcategorised SUBJ, OBJ, or OBL_AG grammatical function.

Path types Table 7 gives an overview of the distribution of path lengths in the fpath assignments we extracted from the FrameNet data. With increasing path length, the frequency of occurrences decreases, while the variety of fpath types increases. We further differentiate between *outside-in* paths (the path leads from the f-structure of the FEE downwards to an embedded f-structure node) and *inside-out* paths (leading from the FEE inside-out and outside-in to an f-structure node that is not dominated by the FEE).

Infrequent path occurrences are susceptible of noise in the data or are not expected to contribute valuable information in stochastic training. So, both for the extraction of linguistic knowledge and for stochastic training, we could set a frequency-based threshold on the length of paths to consider. A general cut-off for all paths to length ≤3 retains 97.27% of the coverage, and yields a reduction of path types to 58.47%. However, the frequency distributions for inside-out and outside-in path types are quite different. Also, the variety of fpaths is significantly higher for inside-out paths (73%) as opposed to outside-in paths (27%). A selective cut-off, restricting path length to ≤2 for outside-in, and ≤3 for inside-out paths leaves 96.44% coverage and 48.48% of path types; including path length 3 for inside-out yields 98.43% coverage with 73.70% of the path types.

As seen in Figure 8, inside-out fpaths of length 3 occur most frequently among inside-out fpaths, and two of them range among the top ten frequent fpaths overall.[11]

Thus, as an alternative to a cut of data based on path length, a cut-off on the basis of frequencies for individual fpaths could be more adequate for cautious filtering.

Generalising over non-local assignment paths Among the top ten inside-out fpaths we also find the non-local fpath described in Section 4.2. This fpath occurs with 135 verb types (210 tokens). For 4 verb types we do not find a corresponding local fpath in the extracted rule set. However, there are 501 verbs (4385 tokens) with local subject fpaths, while we have seen the non-local configuration only for 135 types. These remaining 370 types can be caught by generalised fall-back rules for the non-local variant, if in new corpus data they occur in the identified non-local context.

On the other hand, there are less frequent non-local paths that account for general syntactic configurations that we may encounter in new data, such as the coordination construction in (2). Here the FEE *occupants*, which triggers the RESIDENCE frame, takes as its LOCATION role the coordinated adjunct PP *of .. flats*. The coordinated adjunct is attached high to the coordinated noun heads *owners and occupants*. This high attachment is reflected in the f-structure, which differs from non-coordination.[12] The fpath we obtain is (($ ↑) ADJUNCT), crossing coordination inside-out.

(2) *give greater protection to the [[owners and occupants] [of shops, commercial premises, houses and flats]]*

We identified 97 instances of this pattern, for 38 predicates in 13 frames and for 16 roles. The corresponding local

[11] The element relation of set-valued ADJUNCTs does not contribute to the path length, but it does for coordination: (($ ↑) $).

[12] The grammar does not distribute ADJUNCTs in coordination.

fpath (adjunct_of) occurs in 1013 instances of 340 predicates in 33 frames and for 62 roles. Again, we can provide alternative local/non-local annotation rules, to account for non-local configurations that are not in the data set.

6. Implications

There are several conclusions that can be drawn from the data analysis in Section 5.

Filtering noise In order to be able to extract a lexical semantic resource with a general syntax-semantics interface from corpus annotations, we must acquire sufficiently large and varied corpus samples. We have seen for various examples that reliable generalisations can only be obtained if noise in the data can be eliminated by various kinds of frequency-based filters. Where appropriate, these should be combined to yield reliable confidence measures.

Targeted data acquisition On the basis of quantititive evaluations and an automated frame-assignment architecture, we can identify candidate sentences in unparsed text to 'fill gaps' in the pruned set of annotations, or to provide additional 'evidence' in cases of indiscriminative data counts. Thus, we can pursue a process of targeted data acquistion in an effective, and semi-automated way.

Rule generalisations As seen in Table 7, and in the analysis of the active-passive diathesis, there is a great variety of fpaths in the mapping to semantic roles, due to constructional varieties in the underlying corpus sentences. We identified related local and non-local fpath assignments, and more of these need to be established by data inspection. For such regular alternations, we can identify gaps for local variants, which we can fill with newly acquired data, for the extraction of a frame semantic lexicon with well-defined syn-sem mappings.

For the purpose of active learning techniques in stochastic model building, regular alternations and constructional variants in frame projection can be modeled by generalising frame assignment rules to account for the respective variants. This extends the coverage of automated frame assignment, and the stochastic models that are built on top of it.

Corpus-driven vs. lexicographic The SALSA project – a primarily corpus-driven annotation effort – will be confronted with additional challenges. In contrast to FrameNet data, assembled in a lexicographic effort, the TIGER corpus is less balanced and features novel annotation problems (idioms, support constructions, or metaphors). The need to acquire additional data by generalisations over existing annotations will be even more important in this scenario, to extend the base of annotations in a targeted way.

However, the TIGER annotations will provide a significant boost, for construction of an initial set of frame assignment rules and models for probabilistic selection. Acquisition of novel informative training data can be steered by data analysis and generalisations over existing annotations.

Interplay of statistical and symbolic techniques In sum, we propose to combine statistical techniques with a symbolic syntax-semantics interface for frame assignment, to support both the targeted acquisition of 'informative' training data and the extraction of a semantic lexicon with a well-defined syntax-semantics interface.

Acknowledgements The research reported here was conducted in the project LFG4SALSA, a cooperation project of the German Research Center for Artificial Intelligence, DFKI Saarbrücken and the Computational Linguistics Department of the University of the Saarland at Saarbrücken.

7. References

Baker, C. F., C. J. Fillmore, and J. B. Lowe, 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL 1998*. Montréal, Canada.

Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith, 2002. The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.

Bresnan, J., 2001. *Lexical-Functional Syntax*. Oxford: Blackwell Publishers.

Erk, K., A. Kowalski, S. Padó, 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.

Fillmore, Charles J., 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280.

Fleischman, M., N. Kwon, and E. Hovy, 2003. Maximum entropy models for FrameNet classification. In *Proceedings of EMNLP'03*. Sapporo, Japan.

Forst, M., 2003. Treebank Conversion – Establishing a testsuite for a broad-coverage LFG from the TIGER treebank. In A. Abeillé, S. Hansen, and H. Uszkoreit (eds.), *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC '03)*. Budapest.

Frank, A., 1999. From Parallel Grammar Development towards Machine Translation. A Project Overview. In *Proceedings of Machine Translation Summit VII "MT in the Great Translation Era"*. Singapore.

Frank, A. and K. Erk, 2004. Towards an LFG Syntax–Semantics Interface for Frame Semantics Annotation. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, Vol. 2945. Springer Verlag, Heidelberg.

Frank, A. and J. Semecky, 2004. Corpus-based Induction of an LFG Syntax-Semantics Interface for Frame Semantic Processing. *to appear*.

Gildea, D. and D. Jurafsky, 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3).

Gildea, D. and M. Palmer, 2002. The Necessity of Parsing for Predicate Argument Recognition. In *Proceedings of ACL'02*. Philadelphia, PA.

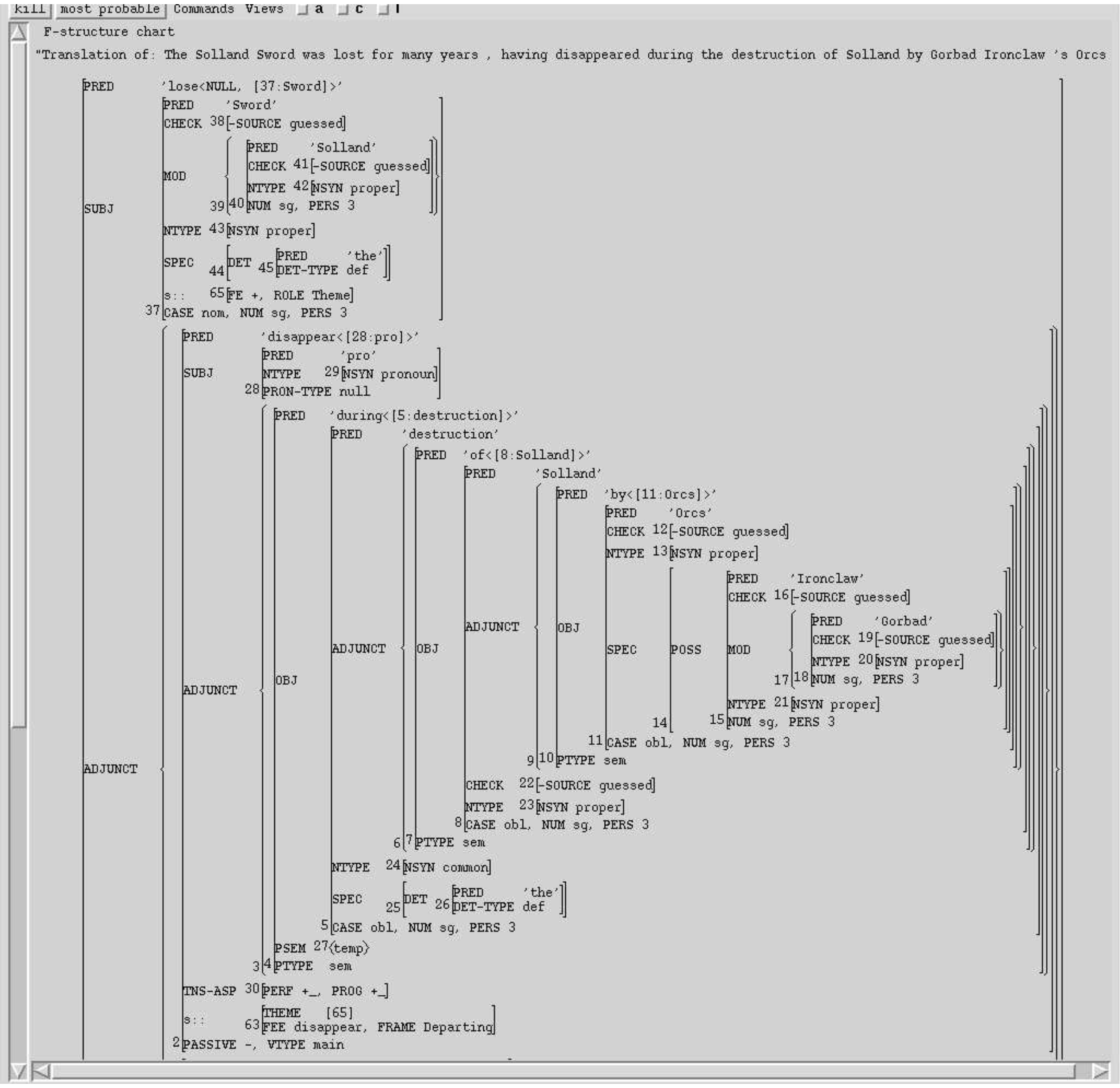
Kingsbury, P., M. Palmer, and M. Marcus, 2002. Adding semantic annotation to the Penn TreeBank. In *Proceedings of the HLT Conference*. San Diego.

Merlo, P. and S. Stevenson, 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–406.

Riezler, S., 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.

Appendix

(1) *The Solland Sword was lost for many years , having disappeared during the destruction of Solland by Gorbad Ironclaw’s Orcs .*



```
 F-structure chart
"Translation of: The Solland Sword was lost for many years , having disappeared during the destruction of Solland by Gorbad Ironclaw 's Orcs

     PRED      'lose<NULL, [37:Sword]>'
               PRED    'Sword'
               CHECK 38[-SOURCE guessed]
                         PRED    'Solland'
               MOD       CHECK 41[-SOURCE guessed]
                         NTYPE 42[NSYN proper]
     SUBJ            39 40 NUM sg, PERS 3
               NTYPE 43[NSYN proper]
               SPEC   44 DET 45 PRED      'the'
                                DET-TYPE def
               s::    65[FE +, ROLE Theme]
            37 CASE nom, NUM sg, PERS 3
                 PRED      'disappear<[28:pro]>'
                           PRED      'pro'
                 SUBJ      NTYPE   29[NSYN pronoun]
                        28 PRON-TYPE null
                             PRED   'during<[5:destruction]>'
                                    PRED     'destruction'
                                                PRED  'of<[8:Solland]>'
                                                      PRED     'Solland'
                                                                    PRED  'by<[11:Orcs]>'
                                                                          PRED    'Orcs'
                                                                          CHECK 12[-SOURCE guessed]
                                                                          NTYPE 13[NSYN proper]
                                                                                          PRED    'Ironclaw'
                                                                                          CHECK 16[-SOURCE guessed]
                                                                                                    PRED    'Gorbad'
                                                      ADJUNCT       OBJ                             CHECK 19[-SOURCE guessed]
                                    ADJUNCT     OBJ                       SPEC     POSS   MOD       NTYPE 20[NSYN proper]
                             OBJ                                                               17 18 NUM sg, PERS 3
                 ADJUNCT                                                                  NTYPE 21[NSYN proper]
                                                                                   14  15 NUM sg, PERS 3
                                                                       11 CASE obl, NUM sg, PERS 3
                                                                  9 10 PTYPE sem
                                                      CHECK  22[-SOURCE guessed]
                                                      NTYPE  23[NSYN proper]
                                                    8 CASE obl, NUM sg, PERS 3
                                             6 7 PTYPE sem
                                    NTYPE  24[NSYN common]
                                    SPEC   25 DET 26 PRED      'the'
                                                     DET-TYPE def
                                  5 CASE obl, NUM sg, PERS 3
                             PSEM 27<temp>
                          3 4 PTYPE  sem
                 TNS-ASP 30[PERF +_, PROG +_]
                          THEME   [65]
                 s::    63 FEE disappear, FRAME Departing
     ADJUNCT   2 PASSIVE -, VTYPE main
```

Figure 5: F-structure for example (1), with partial s-projection for frames