Cross-Lingual Annotation Projection Models for Role-Semantic Information

Dissertation zur Erlangung des Grades
eines Doktors der Philosophie
der Philosophischen Fakultäten
der Universität des Saarlandes

vorlegt von

Sebastian Padó

aus Stuttgart

Saarbrücken, 2007

Gleichzeitig wirft die Notwendigkeit, auf der Grundlage einer kleinen Stichprobe zu allgemeinen Schlußfolgerungen zu kommen, schwierige methodologische Fragen auf.

At the same time, the necessity to come to general conclusions on the basis of a small sample poses difficult methodological questions .

Par ailleurs, la nécessité de tirer des conclusions générales en fonction d'échantillons réduits pose de sérieux problèmes méthodologiques.

(Europarl s14236202)

Letzte Promotionsleistung:	31. Mai 2007
Dekan:	Prof. Dr. Ulrike Demske
Berichterstatter:	Prof. Dr. Manfred Pinkal
	Dr. Mirella Lapata
	Prof. Dr. Matthew W. Crocker

Abstract

Due to the high cost of manual annotation, resources with role-semantic annotation exist only for a small number of languages, notably English. This thesis addresses the resulting *resource scarcity problem* for languages where such resources are not available by developing methods which automatically induce role-semantic annotations for these languages.

We address the induction task by taking advantage of the *resource gradient* between languages, extracting annotations from existing resources (e.g., for English) and transferring them to new languages. We effect the transfer using *annotation projection*, a general procedure to exchange linguistic information between aligned sentences in a parallel corpus. Basic annotation projection is a knowledge-lean approach, and thus applicable even to resource-poor languages. Specifically, we apply projection to semantic annotation in the *frame semantics* paradigm. Frame-semantic annotation consists of two annotation layers: semantic classes for predicates, and semantic roles linking predicates to their arguments. We evaluate our approach by using *FrameNet*, a large English resource for frame semantics, to induce frame-semantic annotation for two target languages, German and French.

In the first part of this thesis, we assess a prerequisite for annotation projection, namely the degree of *parallelism* between monolingual reference annotations in a parallel corpus. Inspection of a manually annotated sample corpus shows that frame-semantic annotation in fact exhibits a substantial degree of parallelism, both with respect to semantic classes and semantic roles. This result holds for both language pairs we consider, namely English–French and English–German.

The two central parts of the thesis are concerned with the actual projection step for individual predicates. We project semantic classes and roles in two separate steps, since the two tasks have different profiles. The *projection of semantic classes* can be realised using simply by using correspondences between predicates, which are usually single words. We

identify translational shifts as the central problem of this task, i.e., translations which change the semantic class (frame) of the original predicate. We demonstrate that knowledge-lean filtering mechanisms relying on distributional properties are sufficient to induce high-precision seed lexicons. In contrast, the *projection of semantic roles* relies mainly on clean correspondences between sentential constituents (i.e., role-bearing phrases). The latter are difficult to obtain due to errors and omissions in the word alignment. We show that better constituent alignments can be obtained by formalising the task as a graph matching problem which can integrate knowledge about syntactic bracketings. The linguistic information encoded in the bracketings alleviates word alignment errors and results in high-precision projections even for noisy input data (e.g., resulting from automatic shallow semantic parsing).

In the last part of the thesis, we approach the problem of translational shifts. We identify a subclass of cases for which parallelism can be restored by considering groups of more than one frame. These *frame group paraphrases* are amenable to a generalised version of annotation projection, and we provide a semi-supervised algorithm for their corpus-based acquisition. In a manual pilot study, we show that the acquisition algorithm results in linguistically plausible frame group paraphrases which can furthermore account for a large portion of translational shifts in our sample.

In sum, the results of this thesis indicate that the semantic generalisations made by frame semantics carry over to a considerable degree from English to other languages. The projection methods we have developed can be applied to robustly and automatically create frame-semantic resources for new languages.

Zusammenfassung

Zum jetzigen Zeitpunkt existieren nur für wenige Sprachen große Korpora mit hochwertiger semantischer Annotation, was primär auf den hohen Arbeitsaufwand für manuelle Bedeutungsannotation zurückzuführen ist. Ziel dieser Arbeit ist es, diesen *Ressourcenmangel* zu beheben, indem wir Methoden zur automatischen Zuweisung von rollensemantischer Annotation für Sprachen entwickeln, für die keine entsprechenden Ressourcen verfügbar sind.

Wir machen uns für diese Aufgabe den *Ressourcengradienten* zwischen Sprachen zunutze, indem wir Information aus bereits existierenden Ressourcen (z.B. für das Englische) in die Zielsprache überführen. Die konkrete Umsetzung findet mithilfe von *Annotationsprojektion* statt, einer allgemeinen Strategie zur Übertragung sprachlicher Information über Sprachgrenzen hinweg. Annotationsprojektion arbeitet auf Satzpaaren in parallelen, alinierten Korpora (d.h., Übersetzungen), bei denen ein Satz (der Quellsatz) annotiert ist, der andere (der Zielsatz) jedoch nicht. Die Quellannotation wird auf den Zielsatz übertragen, indem jedes Element der Annotation auf das beste Übersetzungsäquivalent der zugrundeliegenden sprachlichen Einheit im Zielsatz abgebildet wird. Da Annotationsprojektion (in der einfachsten Form) kein weiteres Wissen benötigt, ist diese Strategie auch für die Anwendung auf Sprachen geeignet, für die wenig computerlinguistische Infrastruktur zur Verfügung steht.

Konkret projizieren wir *frame-semantische Annotation*, der eine zweistufige Theorie der semantischen Rollen zugrunde liegt. In der Frame-Semantik erhält jedes Prädikat zunächst eine semantische Klasse, Frame genannt, die die Menge der zulässigen semantischen Rollen bestimmt. In einem zweiten Schritt werden dann semantische Rollen zugewiesen, die das Prädikat mit seinen Argumenten verbinden. Wir evaluieren unseren Ansatz, indem wir ausgehend von *FrameNet*, einer großen frame-semantischen Ressource für das Englische, vergleichbare Ressourcen für das Deutsche und das Französische induzieren.

Der erste Teil der Arbeit betrachtet den Grad an *Parallelismus* zwischen der frame-semantischen Annotation für die verschiedenen Sprachen. Dies ist notwendig, da Annotationsprojektion die Annotation der Zielsprache direkt nach dem Vorbild der Quellsprache konstruiert. Infolgedessen führt fehlende Übereinstimmung zwischen monolingualen Referenzannotationen in beiden Sprachen zu fehlerhaften Projektionen. Zur Untersuchung dieses Problems wurde ein Testkorpus von 1000 parallelen Sätzen für alle drei Sprachen (English, Französisch, und Deutsch) manuell frame-semantisch annotiert. Unsere Analyse ergibt, daß eine deutliche Mehrheit der alinierten Prädikate Parallelismus aufweisen, und zwar sowohl im Hinblick auf ihre semantische Klasse als auch auf ihre semantischen Rollen.

Der zweite und dritte Teil der Arbeit befaßt sich mit der eigentlichen Projektion. Da semantische Klassen und Rollen bei der Übertragung unterschiedliche Probleme mit sich bringen, betrachten wir die beiden Schritte unabhängig voneinander.

- Die *Projektion semantischer Klassen* (Teil 2) benötigt ausschließlich Alinierungen zwischen Prädikaten, d.h., typischerweise einzelnen Wörtern. Die Hauptschwierigkeit besteht bei diesem Schritt dabei, *translational shifts* zu erkennen, d.h. Übersetzungen, bei denen sich die semantische Klasse des Quellprädikates ändert. Wir zeigen, daß wissensarme Methoden, die vor allem auf die distributionalen Eigenschaften von Übersetzungspaaren zurückgreifen, ausreichen, um Kernlexika zu erzeugen. Dies sind vergleichsweise kleine Lexika, die jedoch eine hohe Qualität aufweisen und daher monolingual erweitert werden können.

- Die *Projektion semantischer Rollen* (Teil 3) erfordert hingegen die Identifikation ganzer Phrasen im Zielsatz, die mit semantischen Rollen etikettiert werden sollen. Die Schwierigkeit dieser Aufgabe liegt vor allem in Lücken und Fehlern in der Wortalinierung begründet, die zu unvollständigen oder falschen Projektionen führen. Wir zeigen, daß diese Aufgabe als Matchingproblem in einem bipartiten Graphen mit gewichteten Kanten formuliert werden kann, dessen Knoten den syntaktischen Konstituenten des Quell- und Zielsatzes entsprechen. Dieser Ansatz kann Fehler in der Wortalinierung mithilfe des linguistischen Wissens ausgleichen, das in der syntaktischen Analyse enthalten ist. Der resultierende robuste-

re Projektionsprozess weist selbst für fehlerhafte Quelldaten eine hohe Präzision auf, was den Einsatz automatischer flacher semantischer Parser zur Annotation einer großen Menge von Quellsätzen ermöglicht.

Im vierten Teil dieser Arbeit nähern wir uns dem Problem der Übersetzungen, bei denen sich die semantischen Klassen der Prädikate nicht entsprechen. Wir zeigt, dass es möglich ist, eine Unterklasse dieser Fälle zu identifizieren, für die der Parallelismus wiederhergestellt werden kann, indem Frames nicht einzeln, sondern gruppenweise betrachtet werden. Auf diese sogenannten *frame group paraphrases* kann dann eine verallgemeinerte Version von Annotationsprojektion angewende werden. Wir geben einen halbüberwachten Algorithmus zur Identifikation von neuen frame group paraphraseses an, evaluieren ihn erfolgreich in einer manuellen Pilotstudie, und zeigen, wie er in einem Bootstrapping-Zyklus mit Standard-Annotationsprojektion für die parallelen Fälle kombiniert werden kann.

Zusammengefaßt zeigen die Ergebnisse dieser Arbeit, dass die semantischen Generalisierungen von Frame-Semantik nicht auf das Englische begrenzt sind und zu großen Teilen auf andere Sprachen übertragen werden können. Die Methoden, die wir entwickelt haben, können verwendet werden, um in der Praxis automatisch und robust frame-semantische Ressourcen für Sprachen zu entwickeln, für die diese bisher nicht existieren.

Acknowledgements

First of all, I would like to thank three people whose ideas and influence have decisively shaped this thesis. The first one is my *Doktorvater* Manfred Pinkal. He gave me a lot of freedom to pursue my own ideas, but kept questioning me regularly about my results, gently pushing me to think about the implications of my work in a wider context. My other advisor, Mirella Lapata, taught me to distinguish the interesting from the trivial, and the infeasible from the interesting. In our collaboration, I have benefited enormously from her huge experience in empirical computational linguistics. The third is my friend and longtime officemate Katrin Erk. Her enthusiasm for our line of work is infectious, and our intensive discussions have time and again changed the way I think about particular issues.

I was lucky to write this thesis in a very productive and friendly research environment. Thanks go to my colleagues in the SALSA project, Aljoscha Burchardt, Anette Frank, and Andrea Kowalski, where I always felt part of a team. Alexander Koller and Marco Kuhlmann generously shared their knowledge (and literature) about graph theory with me. The Internationales Graduiertenkolleg "Language Technology and Cognitive Systems" gave me the opportunity to meet many interesting people, and provided funds for a number of my conference visits.

A number of people were involved with this thesis in less direct but nevertheless important ways. First mention is due to the annotators who prepared my manual gold standards: I am grateful to Beata Kouchnir, Paloma Kreischer, and Garance Paris for their diligent work. Next, Guillaume Pitel and Susanne Salmon-Alt at INRIA Nancy collaborated with me on the annotation of a French corpus with semantic roles. Finally, Chris Callison-Burch, Ana-Maria Giuglea, and Alessandro Moschitti made their software tools available for me, saving me a lot of effort.

Last, but definitely not least, I thank Ulrike for all her love and support, and for reading innumerable versions of the thesis.

Contents

I Introduction and Background

1 Introduction 3
- 1.1 Role Semantics . . . 4
- 1.2 Shallow Semantic Parsing (Automatic Frame-Semantic Analysis) . . . 11
- 1.3 Cross-lingual Annotation Projection . . . 15
- 1.4 Thesis Overview . . . 19

2 Technical Background 23
- 2.1 The EUROPARL Corpus . . . 23
- 2.2 Bilingual Alignment . . . 26
- 2.3 Corpus Preprocessing . . . 35
- 2.4 Summary . . . 40

3 Cross-lingual Parallelism of Role-Semantic Annotation 43
- 3.1 Two Types of General Cross-lingual Parallelism . . . 43
- 3.2 Parallelism of Role-Semantic Analyses . . . 50
- 3.3 Assessing the Cross-lingual Parallelism of Frame-Semantic Annotation . . . 60
- 3.4 Summary . . . 69

II Cross-lingual Induction of Frame-Semantic Predicate Classes

4 A Framework for the Projection of Frame-Semantic Predicate Classes 73
- 4.1 Motivation . . . 73
- 4.2 Constructing Frame-Semantic Predicate Classifications . . . 76

4.3 Projection with a Generate-and-filter Strategy 81
4.4 Error Sources and Filtering Procedures 85
4.5 Summary 92

5 Experimental Evaluation **95**
5.1 Experimental Setup 95
5.2 Experiment 1: Language pair English–German 102
5.3 Experiment 2: Language pair English–French 108
5.4 Dictionary-based Predicate Class Induction 113
5.5 Related Work 114
5.6 General Discussion 116
5.7 A Closer Look at Translational Shifts 122
5.8 Summary 129

III Cross-lingual Projection of Frame-Semantic Roles

6 A Framework for Cross-lingual Role Projection **133**
6.1 Motivation 133
6.2 Decomposing Projection into Alignment and Transfer . . . 134
6.3 Framework Formalisation 137
6.4 Summary 151

7 Experimental Evaluation **153**
7.1 Experimental Setup 153
7.2 Experiment 1: Language Pair English–German 158
7.3 Experiment 2: Language Pair English–French 169
7.4 Related Work 177
7.5 General Discussion 181
7.6 Summary 185

IV Further Directions and Conclusions

8 Beyond Frame Instance Parallelism: Frame Group Paraphrases **189**
8.1 Motivation 189
8.2 Frame Group Paraphrases 190

8.3 Corpus-based Acquisition of Frame Group Paraphrases . . 197
8.4 Experimental Evaluation 203
8.5 Cross-lingual Transfer as a Bootstrapping Cycle 212
8.6 Related Work 215
8.7 General Discussion 218
8.8 Summary 220

9 Conclusions **223**
9.1 Contributions 223
9.2 Projection vs. Manual Resource Creation 226
9.3 Avenues for Future Work 227

V Appendix

A Guidelines for the Evaluation of Projected FEE Candidates **233**
A.1 Introduction 233
A.2 The Annotation Task 234

B Guidelines for Frame-Semantic Annotation **243**
B.1 Introduction 243
B.2 Annotation Procedure 247
B.3 Linguistic difficulties – Frame Choice 255
B.4 Linguistic difficulties – Frame Elements 260
B.5 Reference: Important Syntactic Categories 262
B.6 Reference: (Some) Difficult Frame Distinctions 263

Bibliography **265**

List of Figures

1.1 Example output of a frame-semantic parser (adapted from Erk and Padó (2006)) 12
1.2 Projection of part-of-speech annotation on an English–German sentence pair, idealised version. 16
1.3 Projection of part-of-speech annotation on an English–German sentence pair, less idealised version. Projection errors are marked in grey. 18

2.1 Examples of sentence alignment (left) and word alignment (right) 27
2.2 Word alignment grid: Alignment links provided by the GIZA++ intersective alignment (black) and added by human annotator (grey). 34
2.3 Toolchain for the automatic linguistic analysis of bitexts . . 36

3.1 A short English–German bi-sentence with part-of-speech analyses 47
3.2 A short English–German bi-sentence with frame-semantic analysis 53

4.1 Noisy induction of German FEEs for the frame STATEMENT, using translation pairs in a parallel corpus for an English FEE of the same frame. 80

5.1 English–German: Precision–Recall tradeoff and mean average precision (map) for consistency filters (type-based evaluation). The grey horizontal line corresponds to the size of FrameNet (Recall level ≈ 30%). 106

5.2 English–French: Precision/recall tradeoff and mean average precision (map) for consistency filters (type-based evaluation). The grey horizontal line indicates the size of FrameNet (Recall level ≈ 40%). 111
5.3 Translation of predicates: The case of frame instance parallelism (left) and frame instance non-parallelism (right) . 124

6.1 Cross-lingual projection of semantic role information with optimal alignments. 135
6.2 Bi-sentence (left) with matrices of constituent similarities (top) and edge weights (bottom). Similarities computed according to Eq. (6.12). 140
6.3 Constituent alignments modelled as bipartite graphs and role projections resulting from different specifications of the class of admissible alignments (U_s, U_t: sets of source and target constituents; r_1, r_2: two semantic roles). 145
6.4 Filtering of unlikely arguments (predicate in boldface, argument candidates after filtering in boxes). 151

8.1 Analysis of the bi-sentence from Example (8.1): Non-parallelism on the level of individual frames 192
8.2 Analysis of the bi-sentence from Example (8.1): Parallelism on the level of frame groups 194
8.3 Iterative matching for frame paraphrase acquisition 199
8.4 Sentence 1 of example corpus to illustrate the iterative matching algorithm 200
8.5 Sentences 2 and 3 of example corpus to illustrate the iterative matching algorithm 201
8.6 Problematic bi-sentence: A German frame group involving more than two frames 218

A.1 Decision Tree for the FEE candidate annotation 236

B.1 English syntactic tree (Collins parser) 246
B.2 Choosing a corpus and adding frames 249
B.3 Error in the syntactic structure 255

List of Tables

1.1 Frame COMMITMENT: Definition and annotated example sentences. 10

2.1 Evaluation of English-German statistical word alignment . 32

3.1 Three synonymous English sentences with PropBank-style analyses. 58
3.2 Monolingual inter-annotator agreement on the calibration set (English and German) and on the complete dataset (French) 63
3.3 A quantification of the cross-lingual instance-level parallelism for the language pairs English–German (above) and English–French (below) 66

4.1 Senses, corresponding frames, and German translations of the verb *ask* 75
4.2 Notation overview for Chapter 4 82

5.1 Frame frequency bands (TP: translation pair instances; FNr: number of frames; AvgC: average number of candidate FEEs per frame) 97
5.2 Frames from different frequency bands selected for evaluation, with total number of unfiltered candidates for German and French 98
5.3 English–German: Type-based evaluation of binary filters and relative frequency of error classes (100% = all candidate types) 103
5.4 English–German: Token-based evaluation of binary filters and relative frequency of error types (100% = all candidate tokens) 103

5.5 English–German: Average number of true positives per frequency band for filter combinations (type-based evaluation) . . . 107
5.6 English–French: Type-based evaluation of binary filters and relative frequency of error types (100% = all candidate types) . . . 109
5.7 English–French: Token-based evaluation of binary filters and relative frequency of error types (100% = all candidate tokens) . . . 109
5.8 English–French: Average numbers of true positives per frequency band for filter combinations (type-based evaluation) . . . 112
5.9 Evaluation of dictionary-based induction of semantic predicate classes. . . . 114
5.10 Incomplete coverage of FrameNet: the case of *admit* . . . 121

7.1 Experiment 1, Condition 1: Model comparison for word-based models (intersective word alignment, development set) . . . 159
7.2 Experiment 1, Condition 1: Model comparison for chunk-based models (intersective word alignments, development set) . . . 159
7.3 Experiment 1, Condition 1: Model comparison for full constituent-based models (intersective word alignments, development set) . . . 162
7.4 Experiment 1, Condition 1: Comparison of the best full constituent-based models on the test set (intersective and manual word alignments) . . . 164
7.5 Evaluation of Giuglea and Moschitti's (2004) shallow semantic parser on the English side of our parallel corpus (test set) . . . 167
7.6 Experiment 1, Condition 2: Performance of best constituent-based model on the test set, using automatically labeled semantic roles as input . . . 168

7.7 Experiment 1, Condition 2: PerfMatch's performance by errors rate in automatic semantic role labelling per frame (Error 0: no labelling errors, Error 1: one labelling error, Error 2+: two or more labelling errors) 169
7.8 Experiment 2, Condition 1: Model comparison for word-based models (intersective word alignment, development set) 170
7.9 Experiment 2, Condition 1: Model comparison for full constituent-based models (intersective word alignments, development set) 172
7.10 Experiment 2, Condition 1: Comparison of the best full constituent-based models on the test set (intersective word alignment) 174
7.11 Experiment 2, Condition 2: Performance of best constituent-based model on the test set, using automatically labeled semantic roles as input 175
7.12 Experiment 2, Condition 2: PerfMatch's performance by errors rate in automatic semantic role labelling per frame (Error 0: no labelling errors, Error 1: one labelling error, Error 2+: two or more labelling errors) 176
7.13 Framework instantiations ((x): depending on word alignment) 182

8.1 Most important FrameNet frames used in Chapter 8 and their definitions (names of semantic roles for each frame printed in small caps) 191
8.2 Definition of self-constructed frames 205
8.3 Cross-lingual breakdown of single frame pairs evoked by *increase/höher*. 206
8.4 Cross-lingual breakdown of frames and frame groups evoked by *increase/höher* (FG: as base frame of frame group; n.c.: instances without realised Cause role; Success: Successfully acquired Frame Group Paraphrase). 207
8.5 Identified frame paraphrases for Ccpos which contribute a Cause role (* = self-defined frame) 208

8.6 English and German paraphrases for CCPOs identified by iterative matching (C = CAUSE; I = ITEM; * = self-defined frame). 209

A.1 Example definition for the frame ARRIVING 234

B.1 The frame REQUEST 244

Part I.

Introduction and Background

1. Introduction

This thesis is concerned with the development of methods for the automatic induction of role-semantic annotation in languages for which no resources are available on this level of description.

Recent advances in Natural Language Processing (NLP) suggest that many applications (such as information access) can benefit greatly from the mapping of surface text onto a *semantic representation*. Two interesting types of information provided by such representations are (a), word sense disambiguation for predicates (to distinguish *pass the house* from *pass the bucket*), and (b), characterisations of the semantic relations between predicates and their arguments (to account for the synonymy of *Peter sells the house* and *The house is sold by Peter*).

Contemporary theories of *role semantics* provide exactly these two types of information. At the same time, they are still amenable to current data-driven modelling techniques, in contrast to semantic representations further removed from the linguistic surface, such as predicate logics-based ones. As a result, recent years have seen substantial work on *shallow semantic parsing*, i.e., the assignment of senses and semantic roles to free text.

Unfortunately, shallow semantic parsing relies heavily on the use of supervised learning techniques. These require large annotated corpora as training data, a problem known as the *lexical-semantic bottleneck*. While such resources are available for English (notably FrameNet and PropBank), the prohibitive cost of manual semantic annotation has impeded the development of comparable resources for almost all other languages. The result is a serious *resource scarcity* problem.

In this thesis, we propose to *automatically induce role-semantic annotations* for languages affected by the resource scarcity problem. To do so, we exploit the cross-lingual *resource gradient*, using the annotation from an English resource as a starting point and employing cross-lingual *annotation projection* in a parallel corpus to induce corresponding anno-

tations for another language. The resulting resource can serve to bootstrap shallow semantic parsers for broad-coverage semantic analysis.

The main contribution of this thesis is the development of two general statistical frameworks for (a), the cross-lingual induction of semantic classes (such as STATEMENT or MOVEMENT); and (b), the cross-lingual induction of semantic roles (such as SPEAKER or GOAL). Our approach provides an inexpensive way of inducing role-semantic resources for new languages, or at least to greatly reduce the human effort involved. The models we construct are knowledge-lean, and thus applicable also for resource-poor target languages, but make use of linguistic information where available. We demonstrate that these models can be used to induce resources with high precision, even if the original data is noisy.

This chapter outlines the state of the art in the research areas relevant for this thesis. Section 1.1 introduces role semantics, the level of linguistic description we will consider. Then, Sections 1.2 and 1.3 discuss shallow semantic parsing and annotation projection. Finally, Section 1.4 presents the structure of this thesis.

1.1. Role Semantics

Role semantics is the level of linguistic analysis that describes the relationship between predicates and their *semantic arguments* or *semantic roles*, i.e., the participants and objects involved in the event or state described by the predicate. The crucial difference from a syntactic description of arguments (e.g., as subjects or objects) is that semantic roles are defined in terms of their *semantic properties*, abstracting to a large degree from their syntactic realisation.

The need for a description of arguments on the semantic level was first recognised in the 1960s in theoretical linguistics by Gruber (1965) and Fillmore (1968). Gruber's motivation was the limitations of Transformation Grammar (Chomsky, 1957), which describes the subcategorisation of predicates purely in terms of phrase types. This representation, Gruber argued, is inadequate for modelling many lexical phenomena, for example variations in the surface realisation of arguments, today called *diathesis alternations* (Levin, 1993). Transformation Grammar can represent quasi-universal diathesis alternations (such as active-passive) by means

of general grammatical transformations. However, consider the following example for the verb *pierce*, where the object being pierced can be realised either as NP or PP[1]:

(1.1) (a) $[\text{Peter}]_{\text{NP}}$ **pierces** [the paper]$_{\text{NP}}$ [with the pencil]$_{\text{PP}}$.

(b) $[\text{Peter}]_{\text{NP}}$ **pierces** [through the paper]$_{\text{PP}}$ [with the pencil]$_{\text{PP}}$.

The first sentence is an instance of *pierce* with the subcategorisation (NP, NP, PP), the second with (NP, PP, PP). In order to relate these two sentences, Transformation Grammar needs to postulate a predicate-specific alternation which applies to *pierce*, but not to related verbs such as *puncture*. This case-by-case solution leads to a large number of transformation rules with little predictive power.

Fillmore (1968) proposed to model the paraphrastic nature of such sentence pairs by introducing a representational device called *deep case* which characterises arguments at a semantic level and corresponds to semantic roles. Fillmore posited a set of six universal semantic roles (Agentive, Instrumental, Dative, Factitive, Locative, Objective). Using these, Example (1.1) is analysed as follows:

(1.2) (a) $[\text{Peter}]_{\text{Agentive}}$ **pierces** [the paper]$_{\text{Dative}}$ [with the pencil]$_{\text{Instrumental}}$.

(b) $[\text{Peter}]_{\text{Agentive}}$ **pierces** [through the paper]$_{\text{Dative}}$ [with the pencil]$_{\text{Instrumental}}$.

Fillmore's analysis shows that the same set of arguments, namely {Agentive, Dative, Instrumental}, is realised in both sentences. In this manner, it represents a generalisation over surface variation which cannot be derived easily from purely syntactic description.

The fact that this generalisation is driven by semantic considerations has one very important side effect: the assignment of semantic roles to arguments provides a coarse-grained characterisation of the arguments' semantic properties. Consider Fillmore's original definitions (1968) for the roles from Example (1.2):

> Agentive: the [...] typically animate perceived instigator of the action

[1]Example taken from Gruber (1965).

> INSTRUMENTAL: the [...] inanimate force or object causally involved in the action
> DATIVE: the [...] being affected by the action

Information of this kind has the potential of supporting inferences about the described events, such as *Peter was responsible for the piercing event*, *The pencil effected the piercing event*, or *Peter is (probably) animate*. In consequence, much research was carried out during the 1970s to determine a unique, universally applicable set of semantic roles. Unfortunately, no such attempt succeeded convincingly: there appears to be a number of choice points in the design of semantic role inventories where each path leads into problems (see Davis (1996) for a detailed discussion). One such design decision is granularity. Coarse-grained, universal role sets like Fillmore's original proposal cannot distinguish between the (semantically very similar) arguments of symmetrical predicates like *resemble*, and are thus not truly universal. The most obvious alternative, namely to define fine-grained roles specific to predicates, can model symmetrical predicates successfully, but has its own serious drawback: the large number of roles significantly reduces the predictive power of the theory. It was due to difficulties like this one that it was suggested to give up the notion of a discrete set of roles altogether. Dowty (1991) argued that the only clear-cut distinction that exists in the domain of semantic roles is between "Proto-Agent", a role cluster subsuming all agentlike roles, and "Proto-Patient", a cluster for all patientlike roles. While this account avoids many of the traditional pitfalls of semantic roles, it is too coarse-grained for practical purposes.

In the 1990s, new interest in semantic roles arose in computational linguistics. At that time, the field found itself faced with the challenge of processing and interpreting large amounts of textual data (e.g., from the Internet). This situation called for robust and scalable linguistic analysis models, which, as it turned out, could be acquired statistically from large, annotated corpora. An important initial achievement was the syntactic annotation of the Penn Treebank (PTB) for English (Marcus, Santorini, and Marcinkiewicz, 1993) with a context-free, comparatively flat syntactic structure. The availability of this resource led to considerable advances in the automatic construction of wide-coverage parsers (Collins, 1997; Charniak, 1997). Constituent-based syntactic analyses, such as adopted

by the PTB, however remain primarily surface-oriented: similar to the transformation grammar analyses described above, no common analysis is given to diathesis alternations; also, nonlocal arguments (such as subjects in control constructions, or extraposed arguments) are nontrivial to identify.

In this situation, role semantics reemerged as a representation that could provide a normalisation beyond syntactic structure, and from which both general data-driven models of lexical semantics and NLP tasks requiring semantic knowledge should be able to benefit. The particularly interesting types of information provided by semantic roles in this context are (a), information about the type of event or state referred to by a predicate, and (b), a semantic specification of the relationship between the event and the participants expressed by arguments of the predicate. In addition, the prominent function of semantic roles in many theories of syntax-semantics linking (see Davis (1996) for a discussion) indicates the practical feasibility of mapping from syntactic structure onto semantic roles. Note in this context that role semantics does not, in contrast to logics-based semantic theories such as Montague semantics (Montague, 1974), attempt to model the complete meaning of a sentence. Concentrating on the types of information described above, it typically disregards structural meaning aspects such as quantification, modality, or negation.[2]

Today, three frameworks for semantic roles with different backgrounds are in general use in computational linguistics: *PropBank* (Palmer, Gildea, and Kingsbury, 2005), the Tectogrammatical Layer of the *Prague Dependency Treebank* (Hajičová, 1998), and *Frame Semantics* (Fillmore, 1982). All three are coupled with large-scale annotation projects which have demonstrated the feasibility of these frameworks for practical annotation. Importantly, it has been shown that semantic roles are amenable to data-driven modelling, which enables the construction of *shallow semantic parsers* to assign semantic roles to free text (see Section 1.2 for details). In this endeavour, the annotated corpora have figured as valuable training data.

This development has led to a considerable number of studies investigating how semantic roles can be used beneficially in natural lan-

[2] See Carlson (1984) for a proposal which integrates semantic roles in the syntax-semantics interface of a truth-conditional interpretation.

guage processing contexts. They have been applied to information extraction (Surdeanu, Harabagiu, Williams, and Aarseth, 2003), question answering (Narayanan and Harabagiu, 2004), machine translation (Boas, 2002), the modelling of entailment relations between sentences (Tatu and Moldovan, 2005; Burchardt and Frank, 2006), telephone call classification (Hakkani-Tür, Tur, and Chotimongkol, 2004), and the creation of visual scenes from text (Johansson, Williams, Berglund, and Nugues, 2004). Semantic roles have been also found application as a language-independent representation mediating between surface and ontology (Frank, Krieger, Xu, Uszkoreit, Crysmann, Jörg, and Schäfer, 2007), and for representing prepositional information in biomedicine (Cohen and Hunter, 2006). First steps also exist towards using semantic roles for modelling formal inferences (Baumgartner and Burchardt, 2004) and for constructing Semantic Web representation (Narayanan, Fillmore, Baker, and Petruck, 2002).

1.1.1. Frame Semantics and FrameNet

Frame Semantics (Fillmore, 1982, 1985) is a usage-based theory of meaning which focuses on the role of conceptual background knowledge in comprehension:

> I thought of each [...] frame as characterizing a small abstract "scene" or "situation", so that to understand the semantic structure of the verb, it was necessary to understand the properties of such schematized scenes. (Fillmore, 1982, p. 115)

For example, Frame Semantics argues that a hearer cannot understand the predicate *promise* without being familiar with the typical "committment" situation in which one party makes a pledge to a second party to follow some future course of action.

Building on this observation, Frame Semantics describes the meaning of predicates by reference to *frames*. On the conceptual level, a frame models a specific schematised situation and characterises the background knowledge attached to it. On the linguistic level, a frame is a semantic class containing all predicates which are capable of expressing the situation in question. Occurrences of these predicates in text are said to *evoke* the frame; thus, the predicates are called *frame-evoking elements (FEEs)*.

Each frame also specifies the "participants and props" of the situation it describes, the so-called *frame elements (FEs)*. Frame elements are the Frame Semantics instantiation of semantic roles. Frame Semantics assumes that under normal circumstances, there is a direct parallelism between the conceptual and the linguistic level: the participants of the situations are realisable as arguments of the frame-evoking elements. In sum, the frame-evoking elements of one frame share two important properties: (a), they are able to express the same semantic arguments; and (b), they are able to describe the same state of affairs.

Note that the definition of semantic roles on the intermediate level of frames (i.e., per semantic class) crucially distinguishes frame semantics from other theories of role semantics: Frame-specific roles allow controlled generalisations over all FEEs of the current frame without the need for a universally appropriate set of roles, and thus avoid the granularity problem outlined above. Generalisations across frames can be specified, but are not implied by default.

The Berkeley FrameNet project (Baker, Fillmore, and Lowe, 1998; Fillmore, Johnson, and Petruck, 2003) has been compiling a Frame Semantics-based *semantic lexicon* for English since 1997. Its current release (1.3) contains about 800 frames and 10,000 lexical items. Each frame is provided with a definition in natural language; these often list properties of the described situation and presuppositions of the situation or the frame elements.[3] As an example, Table 1.1 shows the FrameNet entry for the COMMITMENT frame which corresponds to the situation outlined informally above.

In FrameNet, the FEEs are represented as lemma-dot-part of speech (e.g., *promise.n*). Currently, only individual lexical items can evoke frames (note that particle verbs count as one lexical item), and FrameNet furthermore focuses on verbs, nouns, adjectives, and prepositions. However, a more general definition of FEEs within the framework of Construction Grammar (Kay and Fillmore, 1999) is envisaged, which would, for example, also allow superlative morphemes to evoke frames. Many FEEs are polysemous with respect to frames, i.e., can serve as FEEs for more than one frame. For example, the predicate *pass* can introduce either the

[3] FrameNet also provides a hierarchy which relates individual frames; however it is at present only skeletal.

Frame: Commitment		
Def.	A Speaker makes a commitment to an Addressee to carry out some future action. This may be an action desirable (as with promise) or not desirable (as with threaten) to the Addressee.	
Frame Elements	Speaker	The Speaker is the person who commits him/herself to do something.
	Addressee	The Speaker's commitment can be made to an Addressee.
	Message	An expression of the commitment made by the Speaker.
	Topic	The topic about which the Speaker makes a promise.
	Medium	The Medium is the physical entity or channel used to transmit the Message.
FEEs	consent.v, covenant.n, covenant.v, oath.n, vow.n, pledge.n, pledge.v, promise.n, promise.v, swear.v, threat.n, threaten.v, undertake.v	
Annotation	[Democratic audiences]$_{\text{Speaker}}$ had to **consent** [to this approach]$_{\text{Message}}$.	
	[The politicians]$_{\text{Speaker}}$ made vague **promises** [about independence]$_{\text{Topic}}$.	
	["I'll be back , "]$_{\text{Message}}$ [he]$_{\text{Speaker}}$ **threatened**.	

Table 1.1.: Frame Commitment: Definition and annotated example sentences.

Giving frame (as in *Pass me the butter.*) or the Path_shape frame (as in *We passed the station.*). In addition to the definitional content, FrameNet gives example sentences for almost all FEEs. These amount to an annotated corpus of more than 100,000 example sentencess extracted from the BNC (Burnard, 1995), and selected according to lexicographic criteria for the purpose of demonstrating important syntactic alternations of each predicate. FrameNet is under continual development and presently covers neither the complete core vocabulary of English, nor the complete "frame space".

The frame construction process results in a considerable degree of *cross-lingual interpretability* of the frames. Being defined mainly on the conceptual level, they are often appropriate to describe not only English predicate-argument structures, but also those of other languages. In fact, it has been proposed that frames can be seen as a largely language-independent meaning representation (Boas, 2005). Several projects are currently investigating the use of the English FrameNet frames for the semantic analysis of other languages, usually coupled with the annotation of corpora for the new language. Currently, active development takes place for German (Erk, Kowalski, Padó, and Pinkal, 2003; Burchardt, Erk, Frank, Kowalski, Padó, and Pinkal, 2006b), Japanese (Ohara, Fujii, Ohori, Suzuki, Saito, and Ishizaki, 2004), Spanish (Subirats and Petruck, 2003), and a common FrameNet for Romance languages (Pallotta, 2005); efforts for French (Pitel, 2006), Hebrew (Petruck, 2005), and Swedish (Viberg, 2006) are getting underway.

In sum, the conceptual nature of Frame Semantics offers the best prospects for the cross-lingual interpretability of role-semantic information. Therefore, the rest of this thesis will focus on Frame Semantics. We will consider the language-independence of linguistic analyses in detail in Chapter 3. This will also include an overview and discussion of the two other semantic role frameworks, PropBank and Tectogrammatical Structure.

1.2. Shallow Semantic Parsing (Automatic Frame-Semantic Analysis)

The use of semantic role information for any open-domain application requires robust models for the role-semantic analyses of predicates in free text, a task known as *shallow semantic parsing*. In the Frame Semantics paradigm, shallow semantic parsing of consists of two steps:

Frame assignment: Which frame (i.e., semantic class) is evoked by the predicate?

Role assignment: What constituents in the sentence realise which frame elements (i.e., semantic roles) of this frame?

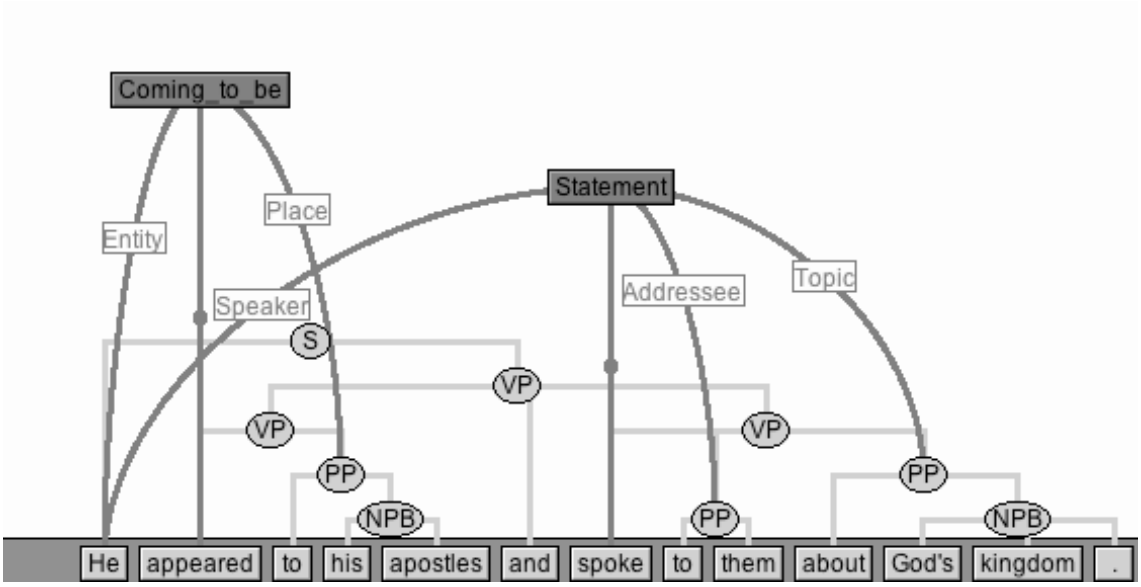


Figure 1.1.: Example output of a frame-semantic parser (adapted from Erk and Padó (2006))

Figure 1.1 illustrates the output of a shallow semantic parser for an example sentence. Light grey angular lines and oval boxes indicate the syntactic structure; frames and frame elements are added on top in dark grey rectangular boxes and curved lines. The sentence contains two predicates, *appeared* and *spoke*. The first of these, *appeared*, evokes the frame COMING_TO_BE, with the roles ENTITY (assigned to the subject pronoun *he*) and PLACE (assigned to the prepositional phrase *to his apostles*). The second predicate, *spoke*, evokes the frame STATEMENT, with three roles (SPEAKER, ADDRESSEE, and TOPIC).

Starting with the seminal study by Gildea and Jurafsky (2002), a considerable number of studies has investigated the task. The interest within the computational linguistics community is witnessed by the adoption of shallow semantic parsing as shared task at SENSEVAL 3 (Mihalcea and Edmonds, 2004) and CoNLL (Carreras and Màrquez, 2004, 2005). The predominant paradigm has been supervised statistical modelling, so that the annotated example sentences provided by FrameNet proved instru-

mental in the construction of shallow semantic parsers.[4] Both frame and role assignment were almost exclusively modelled as classification tasks, the major properties of which have now emerged.

Frame assignment. Frame assignment is a disambiguation step: Given all frames that a predicate can evoke, and one particular instance of this predicate in running text, frame assignment determines the unique frame that is appropriate for this instance. Since the set of appropriate frame elements is dependent on the frame (cf. Section 1.1.1), frame assignment has to precede role assignment. In the example above, the predicate *appeared* had to be disambiguated between the two frames listed in FrameNet, namely COMING_TO_BE, which can be paraphrased with *emerge* or *materialise*, and APPEARANCE, *to exhibit certain salient characteristics*. Only when COMING_TO_BE was chosen, it became clear that the two appropriate frame elements were ENTITY and PLACE.

Frame assignment is closely related to the well-studied problem of word sense disambiguation (WSD): frames can be seen as coarse-grained senses for predicates, and thus the set of frames that a predicate can evoke can be seen as a sense inventory (Erk, 2005). It follows that all major strategies available for word sense disambiguation (see Agirre and Edmonds (2006) for an overview) are also applicable to frame assignment. Unfortunately, WSD is known as a hard problem, in the sense that even with supervised methods, it is difficult to improve significantly on a simple frequency baseline that assigns each predicate the most frequent sense (McCarthy, Koeling, Weeds, and Carroll, 2004). This may contribute to the fact that frame assignment is considerably less well studied than role assignment: often, the frame is simply assumed as given (e.g., at SENSEVAL 3).

To our knowledge, Erk (2005) is the only study of frame assignment as an independent task. Erk uses a classical WSD approach based on a Naive Bayes classifier that employs a rich feature set combining a bag-of-words context with n-grams and local syntactic information. On a held-out dataset of the English FrameNet 1.2 data, it was evaluated at 93.2% F-Score, just above the frequency baseline (93.0% F-Score).[5] Thompson, Levy, and

[4] A small number of first studies towards unsupervised shallow semantic parsing exist; however, all of these assume a small, fixed set of semantic roles (Swier and Stevenson, 2004, 2005; Grenager and Manning, 2006).

[5] The height of the baseline is due to the *incompleteness of the FrameNet sense inventory*

Manning (2003), who also address the task, develop a joint generative model for frames and roles; optimisation of this model yields the best combination of frame and roles. Again on a held-out dataset of FrameNet 1.2, they report a frame assignment accuracy of 89%.

Role assignment. Role assignment is the identification of semantic roles in the context of the predicate. Usually, this step is modelled as the classification of syntactic constituents which are obtained from robust, PCFG-based parsers.[6] In the example in Figure 1.1, this amounts to assigning each constituent (shown as grey ovals) either a semantic role or a pseudolabel NONE. For the frame COMING_TO_BE, all constituents but the subject and the prepositional object are labelled with NONE.

Since typically only a small portion of all constituents realises semantic roles, role assignment is often subdivided into two separate classification steps: *argument recognition*, a binary division of all constituents into roles and non-roles; and *argument labelling*, the distinction of individual roles, which is only applied to those constituents which were classified as roles in the first step. A number of classification frameworks has been applied to solve these classification tasks. Examples include the direct estimation of probability distributions (Gildea and Jurafsky, 2002), Maximum Entropy models (Fleischman and Hovy, 2003), Support Vector Machines (Moschitti and Bejan, 2004; Giuglea and Moschitti, 2006), and Hidden Markov Models (Thompson et al., 2003).

Argument recognition is a relatively simple problem, since it only requires a binary decision between arguments and non-arguments, a distinction that generalises sufficiently well across predicates. It can be solved mainly on the basis of syntactic features, notably the parse tree path from the predicate to the constituent. F-Scores routinely exceed 90% even with comparatively little training data, and remaining errors are often due to nonlocal arguments for which training data is sparse. Argument labelling, in contrast, is a more difficult problem. The mapping between syntactic positions and semantic roles is highly lexicalised and idiosyncratic (Padó

for many predicates, which leads to an artificially low average polysemy. This problem will be discussed in Section 5.6. On a more realistic German dataset, the system significantly outperformed the baseline (74.7% vs. 69.4% F-Score).

[6] See Frank and Erk (2004) for an account of role assignment as symbolic projection rules in Lexical Functional Grammar.

and Boleda Torrent, 2004), so that syntactic features alone are insufficient. Both practical experiences and theoretical studies (Erk and Padó, 2005) indicate that lexical features such as the predicate itself and the head word of the constituent provide important clues. Since these are typically sparse, argument labelling has been found to require much more training data (Fleischman and Hovy, 2003), and accuracies are still worse than for argument recognition, with the best systems obtaining F-Scores in the high 80s.

1.3. Cross-lingual Annotation Projection

Over the last decade, supervised learning techniques, as described above for shallow semantic parsing, have become one of the mainstays of computational linguistics. These techniques exploit the information inherent in large annotated corpora to construct models for linguistic analysis. Unfortunately, the manual construction of corpora with linguistic annotation is an expensive process which requires detailed specifications and safeguards to guarantee consistency. This problem becomes especially pressing on "deeper" levels of linguistic analysis such as lexical semantics. For example, lexical-semantic annotation systematically involves problems of vagueness and ambiguity that cannot be avoided (Kilgarriff and Rosenzweig, 2000). At the same time, statistical models of such tasks typically require larger amounts of training data (cf. the description of argument labelling above).

In fact, only research for English and a small number of other, mostly European, languages commands the resources necessary to hand-craft large corpora with linguistic annotation. In contrast, the majority of languages worldwide are so-called *resource-poor* languages for which no such resources are available. This *resource scarcity* precludes that low-density languages profit from the advances in supervised learning and modelling. Thus, any method which can help to reduce the manual effort involved with resource creation for new languages constitutes an important step toward cashing out the results of research on English for wider range of languages.

A promising method to overcome resource scarcity is *annotation projection* (Yarowsky, Ngai, and Wicentowski, 2001). Its fundamental idea is

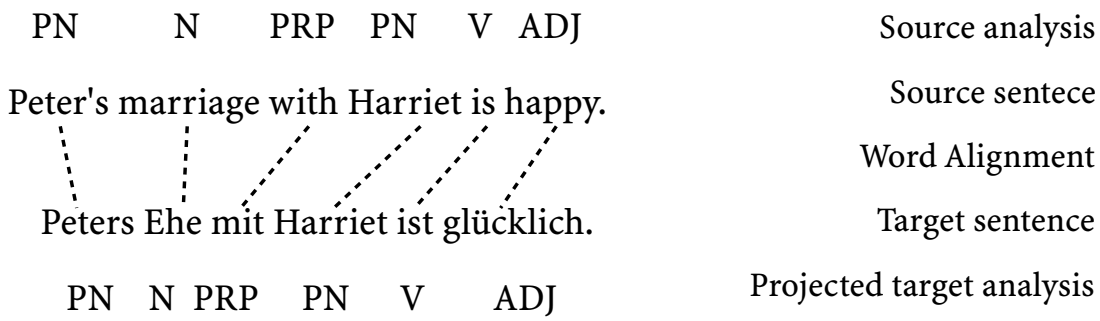


Figure 1.2.: Projection of part-of-speech annotation on an English–German sentence pair, idealised version.

to consider aligned sentence pairs in a parallel corpus for one language where some resource is available (the source), and one language where it is not (the target). For each sentence pair, annotation projection directly transfers the linguistic annotation available on the source side over to the (unannotated) sentence on the target side. This strategy effectively exploits existing knowledge about the source language to construct a dataset for the resource-poor target language, which can subsequently be used as training set for supervised learning.

The central problem of annotation projection is establishing reliable *alignments* between units of linguistic annotation to guide the projection process. Such units can be words, but also chunks, constituents, or even larger units (e.g., in the case of discourse annotation). Alignments for subsentential units rely crucially on the availability of *word alignments*, links between individual words of either sentence that indicate translational equivalence. Thus, annotation projection has been greatly facilitated by the improvements in unsupervised word alignment induction (Och and Ney, 2003). State-of-the-art word alignment models can automatically produce relatively accurate word alignments for sentence pairs, given a sufficiently large parallel corpus of the two languages. No manual annotation is required, which is an important point with respect to the applicability to resource-poor languages.

Figure 1.2 shows a successful case of annotation projection for part-of-speech information, using English as source and German as target

language. For parts of speech, annotation projection simply follows the word alignment links (shown as dotted lines), assigning the part of speech of each English word to its German counterpart. In this example, all German words receive a sensible tag, and thus provide a clean dataset for training a German POS tagger.

Over the last years, the feasibility of annotation projection to induce linguistic resources for resource-poor languages has been demonstrated for a variety of levels of linguistic description (morphology, syntax, and semantics) and a considerable number of language pairs. In morphology, projection has been used for part-of-speech tags (English–French (Yarowsky and Ngai, 2001) and English–Chinese (Hi and Hwa, 2005)). In syntax, applications include NP chunks (English–French and English–Chinese (Yarowsky et al., 2001)) and dependency trees (English–Chinese and English–Spanish (Hwa, Resnik, Weinberg, and Kolak, 2002; Hwa, Resnik, Weinberg, Cabezas, and Kolak, 2005)). Finally, in semantics, studies include word sense (English–Italian (Bentivogli and Pianta, 2005)), coreference chains (English–Romanian (Postolache, Cristea, and Orasan, 2006)), and information extraction annotation (English–French (Riloff, Schafer, and Yarowsky, 2002)).

Unfortunately, annotation projection relies on a strong, problematic assumption, which can be demonstrated on the less idealised sentence pair in Figure 1.3. Here, the projection results in wrong part-of-speech tags for two words, which are marked in grey. For example, the last German word, *verheiratet (married)*, is a participle and should receive the part of speech V (verb), but is assigned N (noun) by the projection. Note that these errors do not result from alignment problems: *verheiratet* is clearly the best translational equivalent of the English noun *marriage* in the German sentence; but the projection is still wrong. The problem is rather that basic annotation projection, as described above, presumes *perfect cross-lingual parallelism of linguistic analyses*. Since the original annotation is just copied between the source and target sentences, it cannot deal with *translational shifts* (van Leuven-Zwart, 1989) which break the parallelism between the two. Such shifts can result from free translations, or differences in preferred lexical realisation patterns (Dorr, 1995), to name only two possible sources for divergences. It is evident that the degree of parallelism determines an *upper bound* for the accuracy of annotation projection proper without postprocessing. However, in this

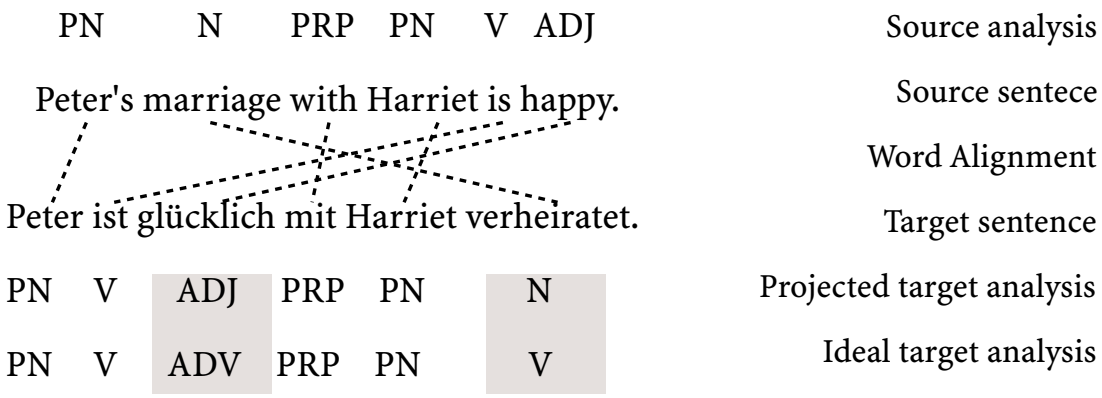


Figure 1.3.: Projection of part-of-speech annotation on an English–German sentence pair, less idealised version. Projection errors are marked in grey.

domain no universal statements are possible: for each language pair, and for each linguistic framework, the degree of parallelism has to be assessed individually.

We will provide a thorough discussion of these issues in Chapter 3. At this point, we note only that many studies on annotation projection have emphasised the need for subsequent filtering even when using high-quality or gold-standard word alignments; this filtering step usually leverages monolingual information about the target language to rule out implausible projection results. Unfortunately, such filtering procedures are often ad-hoc and difficult to motivate. An alternative direction for research, which is however still in its infancy, is the use of high-density language annotation as additional knowledge source in the unsupervised induction of models for resource-poor languages. In this scenario, the linguistic analyses do not have to be exactly parallel, but merely informative for one another. In the area of syntax, this idea has been pursued by Smith and Smith (2004), who train a bilinugal grammar for English and Korean on a parallel corpus, and Kuhn (2004), who uses evidence from a parallel corpus to restrict the rules proposed by a monolingual EM-based PCFG induction method. In lexical semantics, Diab and Resnik (2002) present an unsupervised system for word sense disambiguation based on translational correspondences.

1.4. Thesis Overview

This thesis addresses the resource scarcity that exists for many languages on the level of frame-semantic annotation. Our goal is the development of methods to automatically construct corpora with high-quality frame-semantic annotation in such languages. We approach this task by developing annotation projection-based models that leverage information from the English FrameNet database and transfer it across languages in a parallel corpus. The resulting annotation data can be used as training data for shallow semantic parsers using standard learning methods.

We approach the modelling task statistically, for two main reasons. First, statistical models offer a flexible framework for quantitative reasoning over multiple knowledge sources; this allows us to phrase our models in an extensible way: in their basic state, they are knowledge-lean, thus language-independent and applicable even to very resource-poor languages; however, they can benefit of additional information wherever it is available. Second, a statistical formulation of the projection problem makes it possible to apply well-understood and efficient optimisation techniques that guarantee the scalability of the models to large corpora.

The thesis is structured as follows: The first part starts out with methodological considerations (Chapters 2 and 3). The two central parts address the projection of the two levels of frame-semantic annotation (cf. Section 1.2): In Part II, we describe a model for the induction of frame-semantic predicate classes (Chapter 4) and its evaluation (Chapter 5). Part III developes a model for the projection of semantic roles (Chapter 6) and evaluates it (Chapter 7). In Part IV, Chapters 8 and 9 conclude the thesis with a general discussion and outlook.

Chapter 2 provides information on the parallel corpus we are using, and on the preprocessing steps necessary to prepare it for the projection of role-semantic information. In particular, we spell out the techniques and tools we use for the creation of word alignments as well as for morphological and syntactic analysis.

Chapter 3 addresses the cross-lingual parallelism of linguistic analyses. The first part of the chapter gives a theoretical discussion, concluding

that frame-semantic classes and roles are likely to show considerable cross-lingual parallelism, due to their construction method. To verify this judgement, the second half of the chapter presents a study in which a 1000-sentence parallel corpus for three languages (English, French, and German) is annotated manually with class and role information. Analysis of this corpus shows that parallelism for frames exceeds two thirds, and roles are parallel in around 90% of all cases. These results represent solid empirical support for the appropriateness of applying annotation projection to frame-semantic data. Results from this chapter have been published in Padó and Lapata (2005b).

Chapters 4 and 5 consider the first projection task, namely inducing a predicate classification for a target language which lists the appropriate FrameNet frames for predicates of the target language. This resource is necessary for frame assignment (cf. Section 1.2) in the target language, as well as for determining the applicability of cross-lingual role projection. We show that annotation projection based on word alignment can be used to solve this task. The main problem of this approach is the identification and removal of predicate pairs which are aligned, but evoke non-parallel frames, i.e., translational shifts (cf. Section 1.2). We show that filtering schemes which assess the level of *reliability* of the predicate pair over the complete corpus on the basis of distributional characteristics and shallow linguistic features can be used to remove such cases to a high degree. This results in small, but very clean seed lexicons that be used to bootstrap larger predicate classifications for the target language. Results from this chapter have been published in Padó and Lapata (2005a).

Chapters 6 and 7 focus on the second projection task, the projection of semantic roles between sentences with aligned predicates that evoke the same frame. Again, we approach this task on the basis of word alignment information. We find, however, that the correct projection of roles relies crucially on the quality of correspondences between sentential constituents (i.e., role-bearing phrases), which are difficult to obtain from word alignments directly, due to errors and omissions. To obtain clean constituent alignments, we define a general framework that phrases the task as an optimisation problem and can integrate syntactic bracketing

information and linguistically motivated filtering steps. We assess the impact of a wide range of parameters, notably different kinds of bracketing information (none vs. chunks vs. full parses), filtering schemes, word alignment (automatic vs. manual), and the quality of the semantic role annotation in the source language (automatic vs. manual). Our results demonstrate that linguistic information encoded in the bracketings can effectively alleviate noise and makes it possible to create high-precision projections even for noisy input data and word alignments. Results from this chapter have been published in Padó and Lapata (2005b, 2006).

Chapter 8 develops a perspective for the processing of the class of aligned predicates whose frames are non-parallel (cf. Chapter 3), and which are thus excluded from processing by the methods developed in Chapters 4 to 7. We address these cases by introducing the concept of *frame group paraphrases*, contiguous groups of frames which can be parallel even if individual frames are not. We show that annotation projection can also apply at the level of frame group paraphrases, thus extending the coverage of the methods developed earlier, and provide a semi-supervised algorithm for the identification of frame groups. A pilot study on a small parallel corpus sample with manual frame-semantic analysis indicates that frame group paraphrases in fact capture almost all of the non-parallel cases. Results from this chapter have been published in Padó and Erk (2005).

Chapter 9 summarises and discusses the main results of this thesis and outlines suggestions for future work.

2. Technical Background

As described in Chapter 1, corpora are crucial as data sources in all areas of (empirical) computational linguistics, in particular for the training and evaluation of data-driven models. Cross-lingual investigations rely mostly on a particular class of corpora, namely parallel corpora, which consist of parallel texts in several languages which are translations of one another.

This chapter describes EUROPARL (Koehn, 2005), the parallel corpus used in this thesis. Since word alignments form the cornerstone of projection methods, we also discuss the state of the art in word alignment methods. We sketch the preprocessing toolchain that we use to obtain morphological and syntactic analyses for the English, German and French EUROPARL material.[1] For each aspect of the data and preprocessing, we discuss possible systematic confounds for the subsequent evaluation.

2.1. The EUROPARL Corpus

The current release of the EUROPARL corpus, Release 2, contains around seven years' worth (1996-2003) of professionally translated debates of the European Parliament (Koehn, 2005). This parallel corpus constitutes a *multitext* which provides the same content in the 11 official languages of the European Union at that time[2]. It comprises about 30 million words per language, which makes it, at the time of writing, the largest multilingual corpus that is readily available for research purposes. The corpus can be queried online, using the OPUS (open source parallel corpus) interface at the URL `http://logos.uio.no/opus/`. We concentrate on *bitexts*,

[1] All preprocessing steps were performed at Saarland University, with the exception of syntactic analysis of the French corpus, which was provided by the French FrameNet group at INRIA Lorraine, whose help we gratefully acknowledge.

[2] The languages are Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish, and Swedish.

i.e., parallel texts in two languages which consist of *bi-sentences*, pairs of sentences that are mutual translations. The 11 languages in EUROPARL combine into 55 bitexts. The 10 bitexts corresponding to combinations of English with the other languages can be downloaded in sentence-aligned form from the URL `http://www.statmt.org/europarl`.

In this thesis, EUROPARL data serves both for the evaluation of hypotheses about cross-lingual parallelism (Chapter 3), and as the primary basis of cross-lingual annotation projection methods (Chapters 4, 6, 8) which involves evaluation and model selection. An important issue, therefore, is the influence which the properties of EUROPARL as a corpus have on the methods and evaluation results presented in this thesis.

Corpus representativity. The first crucial consideration in this respect is the profile of the corpus in terms of both genre (text sort) and domain (subject matter), since it is well known in computational linguistics that the performance of models induced on corpora of specific genres or domains (e.g., parsers) deteriorates when these models are applied to different genres or domains (Gildea, 2001). Consequently, an "ideal" corpus would be balanced with respect to both genre and domain, i.e. mirror the distribution of all (relevant) genres and domains in the "real world". However, the impossibility of defining the complete population from which to sample , and the dynamic nature of language over time make the construction of such corpora extremely tedious and expensive. To date, only two large balanced monolingual corpora exist, and no balanced bilingual corpora are available at all.[3]

With respect to genre, EUROPARL is not balanced: it consists exclusively of transcribed spoken language, and its contents respect the special speech conventions of the parliamentary setting, such as frequent personal addresses ("Esteemed colleague, . . . ") followed by the use of first and second person singular, as well as recurring occurrences of of formulaic phrases ("I declare resumed the session of the European Parliament adjourned on . . . "). However, the corpus can be considered to be open-domain, with discussion topics ranging from personal addresses to fishing

[3]The two monolingual corpora are the British National Corpus (BNC) of British English from the 1980s (Burnard, 1995) and the Brown corpus of American English from the 1960s (Kučera and Francis, 1967).

quotas and technical debates, with a slight predominance of political vocabulary.

These imbalances are problems which EUROPARL shares with many other, commonly used corpora such as the Penn Treebank (Marcus et al., 1993). On the other hand, EUROPARL is the largest available parallel corpus of controlled origin (i.e., as opposed to noise-prone internet-harvested corpora), and is commonly used for multilingual research such as statistical Machine Translation (Koehn, 2005). In addition, we find that our application of EUROPARL – annotation projection of semantic class and role information – is comparatively robust to EUROPARL's imbalances. In particular, the induction of semantic classes (Chapter 4) only requires an open domain corpus, which EUROPARL provides. The methods for the projection of semantic roles (Chapter 6) might conceivably suffer from inferior performance of preprocessing software such as parsers on a special-genre corpus (cf. above); however, we can still regard our results as a lower bound of what should be obtainable with a less restricted corpus. Further details on these points are provided in Chapters 4 and 6.

Translation. In a multilingual corpus, the translation process is another central consideration, since its properties fundamentally shape the form of the corpus. For example, consistent, asymmetric translation from one source language into one target language can lead to an imprint of source language features onto the target language text. Similarly, in multilingual contexts a small set of proxy languages is often used through which all translations are routed; this is likely to result in freer translations. Fortunately, the problems apparently do not apply to EUROPARL: Each speaker in the European Parliament addresses the house in his or her native language. The session transcripts are later translated into all other languages and corrected offline[4]. This dataset forms the basis of EUROPARL. Since there is no preferred direction of translation, we expect that there are no major systematic biases which arise from translation. "Online" translation effects which often arise in simultaneous interpreting, such as the insertion of light verbs in translations of sentences from verb-final languages, can also be disregarded.

[4]Source: `http://www.europarl.europa.eu/cre/info_en.htm`

2.2. Bilingual Alignment

Central to the use of bilingual corpora is the construction of *bilingual alignments*. Bilingual alignments in the widest sense are links between structural units on either side of a bitext that express translational equivalence. In other words, alignments link up portions of text that are translations of one another, and thus make it possible in the first place to relate utterances, and parts thereof, across languages.

The two most common levels of alignment are *sentence alignments* and *word alignments*, both shown in Figure 2.1. The upper part of the figure shows an example of sentence alignment links. In non-literary bitexts like EUROPARL, where conserving propositional content is a priority of translation, the vast majority of sentence alignments are one-to-one, as is the case for the second English sentence. The first English sentence demonstrates a case where English realises some complex semantic content as a single sentence, whereas this is split into two sentences in German, resulting in a one-to-two sentence alignment. It is rarely the case that a sentence remains unaligned, since this would mean that it is not translated at all.

The lower part of Figure 2.1 shows the word alignment links between the second English and third German sentences. Note that word alignment appears much more varied than sentence alignments: while many one-to-one alignment links exist, such as *I* $\rightsquigarrow_1$ *ich* or *the* $\rightsquigarrow_1$ *die*, many words remain unaligned. Examples are the English infinitive marker *to*, or the German prepositional adverb *darum*, which are both not translated due to the subcategorisation differences between the matrix verbs: In English, *ask* requires a *to*-infinitive, while in German *bitten* can take *darum* and an embedded sentence. Another frequent phenomenon are one-to-many word alignment links such as between the English *would ask* and the German *bitte*.

The example illustrated that alignments for larger structural units are generally easier to obtain i.e., paragraph alignments can be constructed more easily than word alignments. Indeed, Gale and Church (1993) have shown that accurate paragraph and sentence alignments can be constructed for arbitrary bitexts, solely based on the heuristics that (a), order is mostly preserved in the translation of paragraphs, and (b), translations tend to have *similar lengths*. Naturally, this approach cannot be used to construct alignments on the word level, since reordering of words within

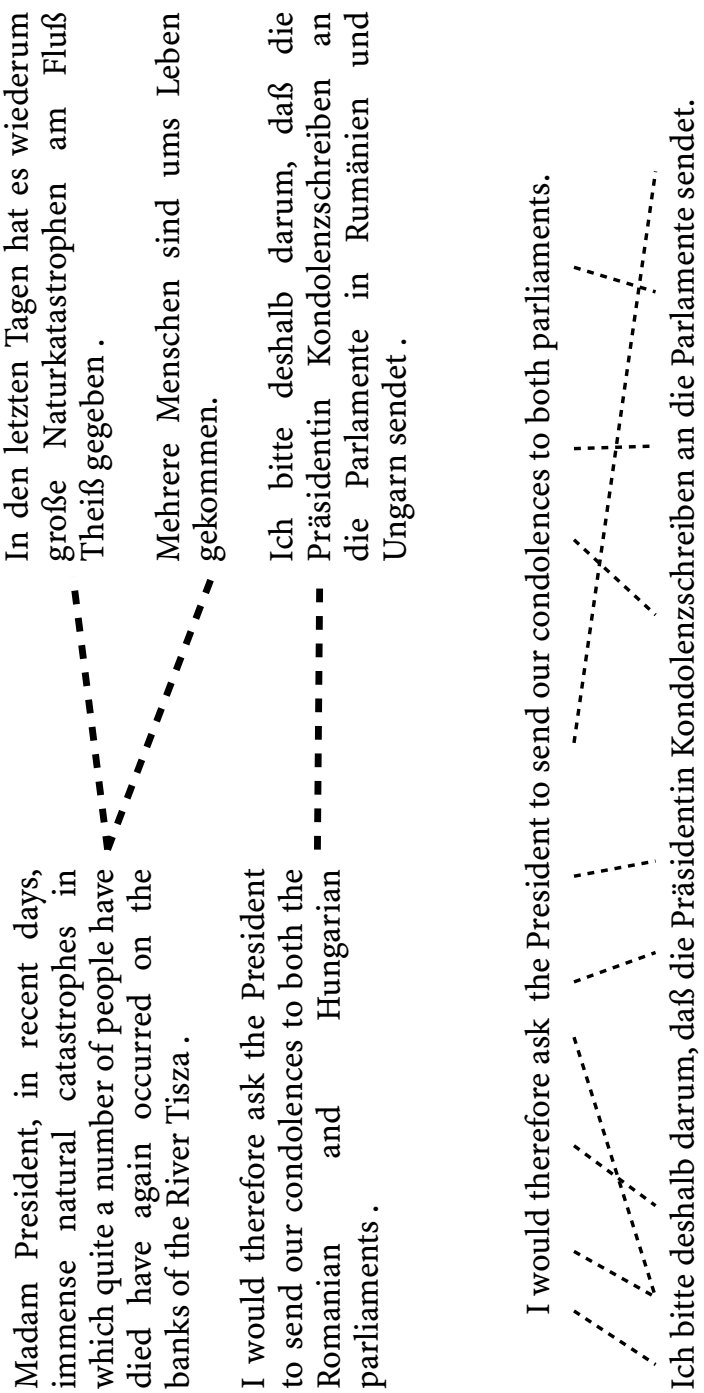


Figure 2.1.: Examples of sentence alignment (left) and word alignment (right)

sentences is very frequent, and the length of words is not a reliable indicator of translational equivalence. Since the EUROPARL corpus is available in a sentence-aligned form, we concentrate in the following on the task of modelling word alignments.

2.2.1. Models for Word Alignment

Assume that we have two aligned stretches of text f (e.g., French) and e (e.g., English), which are both sequences of words. f and e are usually sentences, but they can also be paragraphs. Then, the task of word alignment is to find a relation $a \subseteq e \times f$ which contains all pairs of words which are mutual translations. a is called a word alignment. Approaches to solve this task fall into two major families, namely *association-based* (or heuristic) models and *statistical* (or estimation-based) models.

Association-based word alignment. *Association-based word alignment models* compute a cross-lingual binary *association function* between word types of either language in a first step. Arbitrary linguistic knowledge, such as bilingual lexicons, can be used for the computation of the association function; however, most models restrict themselves to using standard association measures, applied to co-occurrence frequencies obtained from training corpora. Popular choices are the t-test, the Dice coefficient, or pointwise mutual information (Tiedemann, 2003b). In a second step, a concrete pair (e, f) is considered, and a decision procedure is applied to the association measures for the token pairs in (e, f) to obtain an actual alignment. Greedy search is a simple and frequent choice for the decision procedure. Various extensions have been proposed, such as competitive linking, a more sophisticated search strategy (Melamed, 2000). Another option is the use of multi-pass alignment (Pianta and Bentivogli, 2004): in a first pass, only content words are linked up, which results in a reliable, though incomplete, alignment, whose links are called *pivots*. Function words, which are translated in a more varied way, are aligned in a second pass so as to minimise the number of crossing alignment links.

Statistical word alignment models. *Statistical word alignment models* follow a different approach, treating word alignment as a subproblem

of statistical machine translation (SMT). The fundamental idea of SMT is to find a probabilistic model of the *translation probability* $\Pr(f|e)$, i.e. the probability for any sentence pair (f, e) (e.g. French-English) that f is the translation of e (see Brown, Pietra, Pietra, and Mercer (1993) for a detailed discussion).[5] The simplest definition for $\Pr$ is directly as an (empirically estimated) conditional probability p_θ:

$$\Pr(f|e) = p_\theta(f|e) \tag{2.1}$$

Note that p_θ, in this form, only explicitly represents the global probability of translating e as f. However, intuitively, the probability of a given translation depends greatly on which French words are supposed to be translations of which English words. *Statistical alignment models* (Brown, Cocke, Della Pietra, Della Pietra, Jelinek, Lafferty, Mercer, and Roossin, 1990) introduce this notion into the structure of the model itself, by incorporating an additional variable a, the alignment. The alignment specifies which words in f and e are translations of one another. Using alignments, the probability of translating e into f can be defined as the sum of translation probabilities p_θ over all possible word alignments A:

$$\Pr(f|e) = \sum_{a \in A} p_\theta(f, a|e) \tag{2.2}$$

The training of the model consists in finding a optimal set of parameters, $\hat{\theta}$. These are usually estimated by optimising the likelihood of the given translations in some parallel, aligned training corpus Tr, summing over all possible word alignments a, and assuming independence between the individual bi-sentences:

$$\hat{\theta} = \operatorname*{argmax}_\theta \prod_{(f,e) \in Tr} \sum_{a \in A} p_\theta(f, a|e) \tag{2.3}$$

A number of instantiations for the translation model $p_\theta(f, a|e)$ are routinely used, which differ mainly in the independence assumptions they make. Common choices are sequence-based models (such Hidden Markov Models) and fertility-based models (such as the so-called "IBM models" 3–5 (Brown et al., 1993)). See Och and Ney (2003) for a comparative evaluation of different word alignment models.

[5] We assume that f and e are ordered lists of words.

All state-of-the-art models have in common that they include *hidden parameters*, i.e. parameters which cannot be observed directly in the training data. The alignment variable a is exactly such a hidden parameter. Fortunately, methods to deal with missing data such as the Expectation Maximisation (EM) algorithm (Dempster, Laird, and Rubin, 1977) can be used to estimate optimal parameters for $p_\theta(f, a|e)$, although such methods usually cannot guarantee to converge to the global maximum.

A central advantage of the SMT-based approach to word alignment is that the word alignment forms part of the translation model: Once an optimal translation model has been estimated, the optimal alignment for any new bi-sentence (f, e) is simply the single alignment which optimises the translation probability of the bi-sentence:

$$\hat{a} = \operatorname*{argmax}_a p_{\hat{\theta}}(f, a|e) \tag{2.4}$$

This alignment is called the *Viterbi alignment*.

Discussion. While association-based models can potentially integrate more linguistic knowledge, they do not construct word alignments by maximising a global optimisation criterion. This makes their integration into more general statistical processing frameworks problematic, since this usually means expressing the output as the optimal solution to a joint optimisation problem. Statistical alignment models, on the other hand, do have an optimisation criterion, but require a large bilingual corpus for training, and external linguistic information is difficult to integrate. Both model families share the problem of choosing from a wealth of possible instantiations.

For the purpose of this thesis, we chose to use a statistical word alignment model. As mentioned in the previous section, EUROPARL is a corpus with a considerable amount of technical vocabulary. This is a problem for association-based models, which rely on bilingual lexicons to obtain state-of-the-art performance (Pianta and Bentivogli, 2004); on the other hand, the data-driven construction of statistical alignment models makes them naturally adaptive. Also, comparable word alignment models for new language pairs can be constructed very simply from the corresponding bitexts.

The cross-lingual projection strategies we present in subsequent chapters are independent of the alignment model in the sense that they can equally use alignments obtained from association-based models. However, note that the choice of alignment model can influence the types of observed errors to a certain degree. Since the methods we apply for the induction of frame-semantic classes (Section 4) rely on the correctness of individual alignment links, the use of a different alignment model might make it necessary to modify or extend the filtering procedures. Our constituent-based methods for the projection of role information (Section 6) are more robust in this respect by generalising over individual alignment links.

2.2.2. Automatic Word Alignment of EUROPARL

Throughout our experiments, we have used Och's GIZA++ implementation of the statistical word alignment models presented and evaluated in Och and Ney (2003). To alleviate the tendency of Expectation Maximisation to get stuck in local extrema, GIZA++ offers the possibility of stacked training: It starts by training simpler models, and iteratively uses the learnt parameters from one model to initialise the next model, instead of starting from random parameters. This results in a better convergence towards the global maximum. The standard setting involves five iterations of the IBM Model 1, five iterations of the HMM model, five iterations of the IBM Model 3, and five iterations of the IBM Model 4.[6]

For both language pairs (English-German and English-French), we first used the corresponding EUROPARL bitexts to train statistical alignment models with GIZA++. We trained two models for each bitext, one for each direction. In addition, we followed common practice in Machine Translation (Koehn, Och, and Marcu, 2003) and computed the intersection of the source-target and target-source alignments, which is known to exhibit high precision. Finally, we also produced manual word alignments for 1000 German-English bi-sentences (see Section 3 for details), using the intersective automatically produced GIZA++ alignments as a starting point. These were manually corrected by one annotator, who followed the

[6]For more details on the individual models, see Och and Ney (2003) (IBM models) and Vogel, Ney, and Tillmann (1996) (HMM model).

Matching method	Precision	Recall
One-to-one only	89.5	64.2
Strict match	81.4	51.7
Individual links	98.6	52.9

Table 2.1.: Evaluation of English-German statistical word alignment

Blinker annotation guidelines (Melamed, 1998a,b).[7]

In sum, three alignment models were available for each bitext (source-target,target-source, and intersection); for the English-German bitext, we additionally created a manual alignment. When we refer to word alignments or alignment links in subsequent chapters, we mean the Viterbi alignments of these models; the individual chapters specify which alignment is used.

Our manual alignment can also be used to assess the quality of the automatic alignment (at least for the English-German bitext). For reasons of simplicity, we only evaluated the the GIZA++ intersection model, using the manual alignment as a gold standard. We chose this model since it will be used for the final experiments in both Chapters 4 and 6. Note that the intersective word alignment only contains one-to-one alignment links; on the other hand, the Blinker guidelines for manual annotation encourage one-to-many links. To examine the impact of this problem, we evaluated the automatically produced alignments in three conditions against the manually constructed word alignments (see Table 2.1). The first condition (one-to-one only), exclusively considers one-to-one alignment links. The results show that 90% of the one-to-one links posited by GIZA++ are correct, but that the model managed to retrieve only two thirds of all one-to-one links. The next condition (strict match) compares all alignments, be they one-to-one or one-to-many. Not surprisingly, the numbers drop both for precision and recall, since the manual word alignment contains one-to-many alignments which are either not found at all (decrease in recall), or only identified partially (decrease in precision). The last condition, (individual links) makes it possible to assess the impact of these

[7] We would like to thank Chris Callison-Burch for providing us with the `linearb` graphical frontend for this task.

partial alignments: it evaluates individual alignment links by splitting one-to-many links ($s_i \rightsquigarrow (t_j, t_k)$) into their individual one-to-one links ($s_i \rightsquigarrow_1 t_j$, $s_i \rightsquigarrow_1 t_k$). The very high precision number shows that indeed almost all individual links in the intersective GIZA++ alignment are correct; the improvement over the one-to-one only condition consists of alignments links which form part of one-to-many links in the manual alignment. The recall remains low at 53%; again, the difference between this condition and one-to-one only corresponds to one-to-many links not retrieved by GIZA++. These account for almost 10% of all links in the corpus.

To illustrate the problems of word alignment on a concrete example, Figure 2.2 shows a word alignment grid for an example sentence from EUROPARL. Black cells denote word alignment links found by GIZA++, grey cells word alignment links added by the human annotator (in this sentence, no erroneous alignments were found that had to be deleted). As discussed above, an intersective alignment can contain (at most) one alignment link for each row and each column. Nevertheless, one-to-many links are clearly necessary in both directions, at least according to the maximal alignment strategy of the Blinker annotation style guide (Melamed, 1998a), which stipulates the annotation of phrasal alignments whenever no direct correspondence between individual words can be established. An example for a single English word translated as a German phrase is *clearly* $\rightsquigarrow$ *in der Tat*; the inverse direction is illustrated by *before the* $\rightsquigarrow$ *dem*. Figure 2.2 also shows an instance of a many-to-many alignment, where no single word on either side can be said to be a complete translation of the corresponding phrase: *am going* $\rightsquigarrow$ *nun will*.

Missing word alignment links are almost exclusively the result of structural differences between the two sentences, which can be observed on at least three different levels:

Lexical selection. The first group is formed by cases in which the two languages choose different wordings in an otherwise similar structural context. Recall the case of the example introduced above, *clearly* $\rightsquigarrow$ *in der Tat*, where an adverb in English is translated into German as a prepositional adjunct phrase.

Structural choice. The second group consists of cases where the languages differ in their syntactic structures, or in the overt realisation

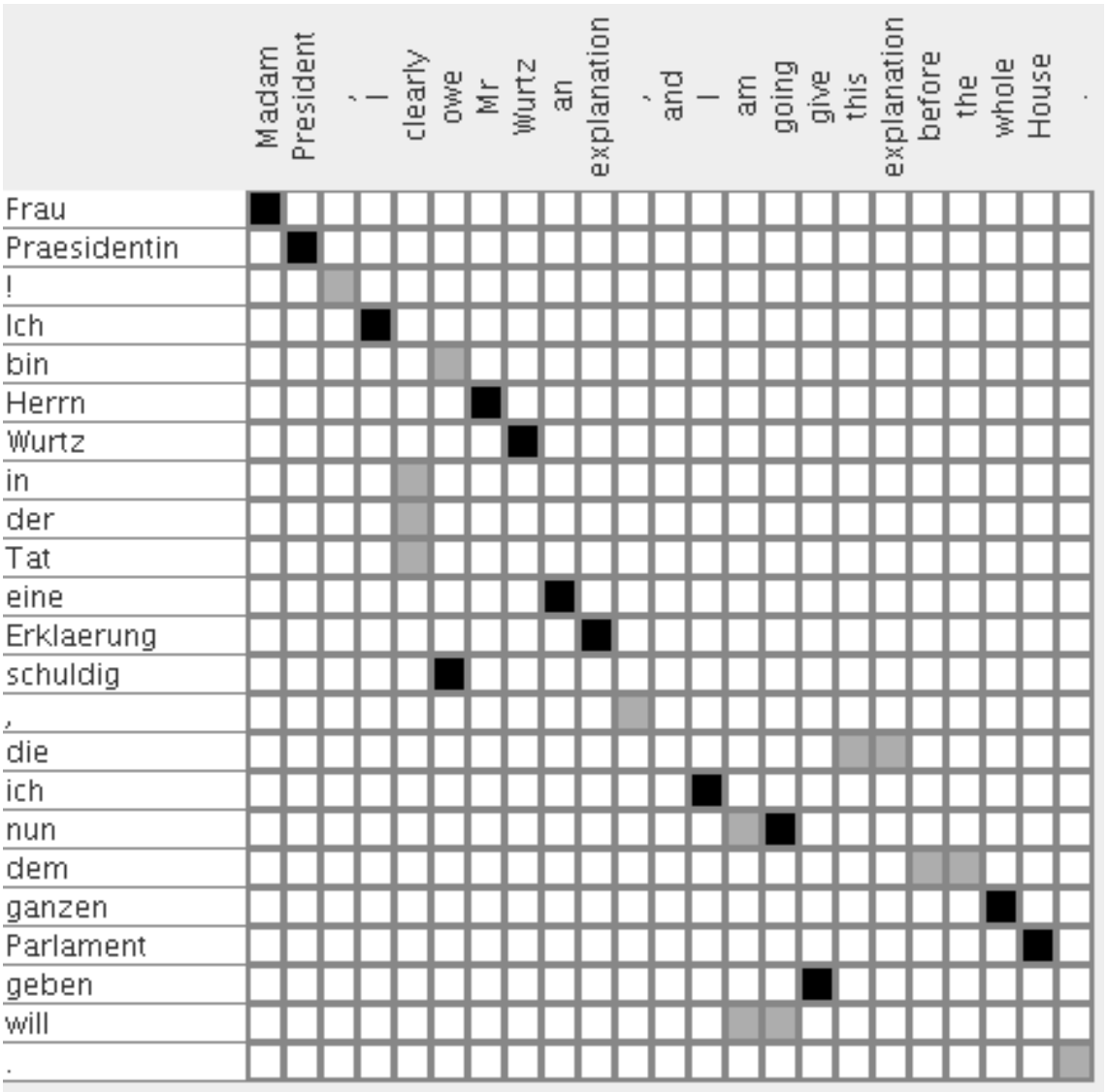


Figure 2.2.: Word alignment grid: Alignment links provided by the GIZA++ intersective alignment (black) and added by human annotator (grey).

of their syntactic structures. The alignment *before the* $\rightsquigarrow$ *dem* reflects a case where the recipient of the statement, the parliament, is referred to in English as a location, while it is expressed as a beneficiary in German, using simple Dative case. A different divergence is present in *owe* $\rightsquigarrow$ *bin schuldig*, where an English active construction is rendered in German with a passive.

Discourse organisation. The third group exhibits differences in text structure such as the manner in which discourse entities are referred to. For example, in the translation *this explanation* $\rightsquigarrow$ *die*, the English speaker chose to repeat a previously mentioned noun phrase a second time, while the German speaker replaced it by a pronoun.

The order of these levels also corresponds roughly to the difficulty of modelling them in a fundamentally word co-occurrence based approach. While it appears that variation in lexical selection can fall naturally out of state-of-the-art statistical alignment models, given enough training data, modelling structural choice appears to presuppose at least some structural knowledge, and the handling of discourse phenomena on a co-occurrence level is very ambitious.

In sum, the evaluation results indicate that automatically obtained word alignments exhibit high precision, even if low recall, and can thus be used as reliable indicators for translational equivalence. Consequently, we use these alignments as data source our automatic projection models, without employing further correction methods.

2.3. Corpus Preprocessing

In this thesis, we concentrate on two bitexts, namely English–German and English–French. The EUROPARL bitexts are tokenised corpora, i.e. they do not contain any information beyond the word forms with the exception of sentence alignments.

Figure 2.3 presents the structure of the preprocessing process on the example of the English–French (E–F) bitext. We first enriched the bitexts with cross-lingual information in the form of statistical word alignments, which could be computed without any further processing. Next, we added monolingual information on different linguistic levels, both word-specific

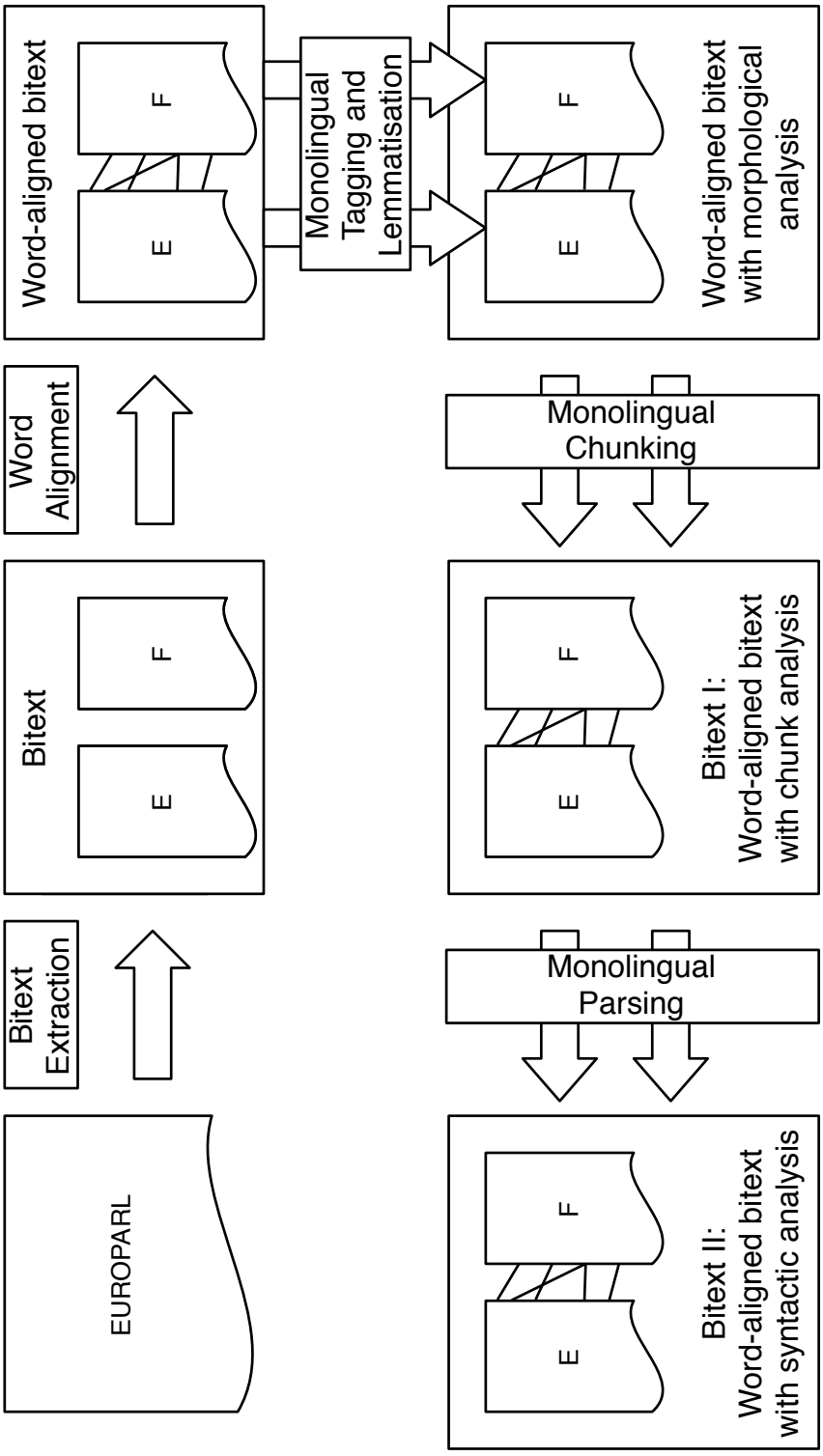


Figure 2.3.: Toolchain for the automatic linguistic analysis of bitexts

shallow information (lemmas and parts of speech) and structural information of increasing depth (chunks and full constituents). In addition, the English text was annotated with frames, using an automatic frame element assignment system (see Chapter 6 for details). For each of these linguistic levels, we chose the best-performing tools that were freely available. Our assessment of the tools' quality was based on the published evaluations of the individual tools; a detailed discussion on the different morphological and syntactic processing components is deferred to the next section (Section 2.3.1).

Of course, even choosing the best-performing tools does not provide a guarantee for error-free linguistic analysis. Considering the general scenario of this thesis, namely the induction of semantic annotation for arbitrary target languages, it may be the case that even comparatively simple preprocessing steps such as lemmatisation are difficult to obtain with high quality. For these reasons, the ability to perform robust projection even when faced with noise in the input data will form an important guiding principle in the development of projection models in the subsequent chapters.

It occasionally happens that processing breaks down completely. This can occur during all preprocessing steps, and we discarded all bi-sentences for which either word alignment failed or the preprocessing for at least one of the sentences failed. As a consequence, bitexts with increasingly deeper analysis become increasingly smaller. The complete EUROPARL bitexts have a size of approximately 1,029,400 sentences.

- We first part-of-speech tagged, lemmatised, and word-aligned the corpora, resulting in what we call *Bitext 1* (cf. Figure 2.3). Bitext 1 contains around 1,010,200 sentences (EN-FR), and 1,004,800 sentences (EN-DE), corresponding to a small loss of about 2%-2.5% of the complete data. This bitext forms the basis for the experiments described in Chapter 4.

- Next, we chunked and parsed Bitext 1. Since parsing is expensive, we limited the dataset by two considerations to keep parsing time manageable. First, we removed all bi-sentences containing a sentence with more than 40 words. In addition, we restricted our attention to "interesting" bi-sentences, i.e., bi-sentences whose English part contains at least one predicate that is listed in FrameNet, and has

a word alignment in the intersective GIZA++ alignment. This left us with *Bitext 2*, with 625,700 sentences (EN–FR) and 674,600 sentences (EN–DE), about 60-65% of the original corpus size. Bitext 2 forms the basis from which the parallel sample corpus described in Chapter 3 is drawn.

2.3.1. Morphological and Syntactic Preprocessing: Software Choice

This section gives details on the software tools used for the morphological and syntactic processing of the monolingual EUROPARL texts, and discuss why they were chosen.

English. We produced part-of-speech tags and lemmas for the English text with a widely used software package for these tasks, Schmid's (1994) Treetagger. One important criterion for choosing this system was its generality: It is both language-independent, providing pre-trained models for all three languages in question (English, German, and French), and covers two tasks at one time, which greatly simplified the preprocessing architecture. Treetagger uses a standard Markov model to determine the most likely part-of-speech sequence; however, instead of directly estimating the trigram tag probabilities from the corpus, it computes them by using a decision tree, thus avoiding sparse data issues. The pre-trained model for English uses the Penn Treebank tagset, and its tagging accuracy on a held-out portion of the Penn Treebank has been determined at around 96%, in the top range of state-of-the-art POS taggers. While there is no explicit evaluation of its lemmatisation component in the literature, the joint tagging and lemmatisation procedure means that correct tagging with a fine-grained tagset such as the Penn Tagset should imply correct lemmatisation, at least for regular forms.

For chunking, we used Abney's (1996) finite-state chunker CASS. CASS is a general parsing architecture which uses cascaded finite-state pattern grammar to (non-recursively) assign chunks and simplex clauses to input sentences. The parser takes part-of-speech tags as input, and groups elements iteratively into larger elements when they match a pattern. CASS is one of the best-documented chunkers for English; Abney (1996) estimates

the precision of the standard English CASS grammar on "a random sample of corpus positions" at 87.9% and the recall at 87.1%. Note that this evaluation is relative to manual gold standard chunk annotation. Since Abney's annotation guidelines these did not aim at producing a "complete" syntactic analysis, CASS's high recall values should have to be interpreted accordingly: In practice, base chunks are often recognised correctly in EUROPARL, but clauses are assigned only rarely.

Parsing was performed with Collins' (1997) probabilistic parser (Model 3). This parser is accurate and robust, and therefore one of the most widely used systems. It provides broad-coverage, relatively flat context-free analyses in the style of the Penn Treebank (Marcus et al., 1993). Its output is a tree with labelled nodes (i.e., syntactic categories). When evaluated on Section 23 of the Wall Street Journal Treebank (sentences with $\leqslant$ 40 words only), Model 3 showed a recall of 88.1% and precision of 88.6%.

German. German part-of-speech tagging and lemmatisation was performed with Treetagger, as outlined above. Treetagger uses the Stuttgart-Tübingen (STTS) tagset for German POS tagging. In the case of German, Treetagger is the only freely available, Unix-based morphological analysis system. Its German model was trained on the Stuttgarter Zeitung corpus and was evaluated on a held-out dataset. The resulting tagging accuracy was around 97%; again, the lemmatisation was not evaluated separately (see the discussion for English above).

Chunks were obtained from Schmid and Schulte im Walde's chunker (2000), which is to our knowledge the only freely available chunker for German. It is based on a hand-written probabilistic context-free grammar, which was extended with robustness rules to increase its coverage. The rule probabilities were acquired from a large corpus using a variant of the EM algorithm. Chunks (primarily noun phrases and some prepositional phrases) are extracted from parse forests. The system was evaluated on a sample corpus from the Frankfurter Allgemeine Zeitung, and showed an unlabelled precision of 93%, and an unlabelled recall of 92%.

Sleepy, Dubey's statistical parser (2005), was used for full syntactic analysis. It is the only available robust, broad-coverage parser for German optimised for accurate syntactic analysis. Its "Smoothed Tiger" grammar model provides a syntactic analysis in the style of the Tiger Tree-

bank (Brants, Dipper, Hansen, Lezius, and Smith, 2002), and performs at 76.3% labelled F-Score.

French. Resources for the computational analysis of French text are scarce, compared to German or English. As mentioned above, we used the French parameter files for Schmid's (1994) Treetagger to tag and lemmatise the French text. Treetagger is, to our knowledge, the only freely available complete morphological analysis system. It uses a set of 33 POS tags which was developed at Stuttgart University and is cited as obtaining a tagging accuracy of 92%.[8]

Syntactic analysis was performed by one of the few existing syntactic analysis tools available for French, the Syntex parser. Syntex is an incremental partial parsing architecture aimed at robust analysis of open-domain French text (Bourigault, Fabre, Frérot, Jacques, and Ozdowska, 2005; Bourigault and Frérot, 2005). It builds on the output of Treetagger and produces dependency parses using a hybrid approach. The backbone of the parser is formed by hand-written rules, but it integrates probabilistic information where appropriate (e.g., to resolve attachment ambiguities). The developers have performed a preliminary evaluation of the parser on a 500-sentence corpus, corresponding to roughly 5000 dependency links. In a comparison against a manually constructed gold standard, the parser showed a labelled recall of 83.4% and a labelled precision of 89.3%, i.e., an F-Score of 86.2% (Bourigault, p.c. 2006).

2.4. Summary

In this chapter, we have described the technical background of the thesis, which consists of three main elements. The first element (Section 2.1) is the EUROPARL corpus, a multilingual corpus in eleven parallel languages, which will form the basis for our experiments. EUROPARL is a body of transcriptions of debates from the European Parliament, with a size of around 30 million words per language. While the corpus covers only a specific genre, this is a common problem of corpus-based models; on the other hand, EUROPARL is reasonably domain-independent, thus meeting

[8] Source: `http://www.uni-stuttgart.de/lingrom/stein/forschung/resource.html`

the crucial desideratum for wide-coverage semantic modelling. We have also shown that systematic effects of translation are not to be expected, given the nature of the transcription and translation processes.

The second element (Section 2.2) is word alignment, i.e., the establishment of translational relations on the word level in a bitext. After discussing the two main families of word alignment models, we have argued for the choice of statistical word alignment models. These models require less linguistic resources, and are fundamentally adaptive, thus being better suited to new genres and language pairs. An evaluation of the automatically constructed word alignment for the language pair German–English has revealed that the obtained alignment links show a very high precision, but a comparatively low recall, resulting from structural divergences on a variety of linguistic levels between translated sentences.

The third element (Section 2.3) is the linguistic preprocessing steps that were applied to the EUROPARL bitexts which our experiments concentrate on, namely English–French and English–German. All bitexts are part-of-speech-tagged, lemmatised, and parsed. We use the best available state-of-the-art tools, but do not perform manual postprocessing to create a realistic experimental setup.

3. Cross-lingual Parallelism of Role-Semantic Annotation

In Chapter 1, we have touched on the topic of *cross-lingual parallelism* of linguistic annotation, stating that annotation projection will produce correct results only if the ideal annotations of two bi-sentences are parallel or identical. In this chapter, we investigate the topic of cross-lingual parallelism in more detail, finding that there are two levels of parallelism that can be distinguished, namely parallelism at the *concept level* and at the *instance level*. To our knowledge, this distinction has not been made explicit in the existing literature on annotation projection. It is however highly relevant, since parallelism on the concept level is not sufficient for annotation projection, which requires parallelism on the instance level.

We continue by defining the concepts of *frame instance parallelism* and *role instance parallelism*, which represent special cases of instance-level parallelism tailored to the needs of annotation projection for frame-semantic information. Since no empirical estimate for these classes of parallelism exist to date, we present the annotation of a trilingual *parallel sample corpus* extracted from EUROPARL with frame-semantic information to empirically assess the degree of frame and role parallelism. The encouraging results represent solid evidence for the practical applicability of the projection frameworks we will present in subsequent chapters.

3.1. Two Types of General Cross-lingual Parallelism

In Chapter 1, we referred to cross-lingual parallelism in a rather unspecific manner as "parallelism" between "annotation" in two languages. Upon closer consideration it becomes clear that this characterisation affords two interpretations: The first is as *concept-level parallelism*, the claim that

some linguistic theory, and the categories it introduces, can be used to describe some linguistic level in more than one language. The second is as *instance-level parallelism*, the claim that any pair of translationally equivalent linguistic entities in a parallel corpus should receive the same analysis on some linguistic level.

It turns out that the two levels of parallelism have traditionally been investigated in two different research areas: concept-level parallelism is discussed mostly in theoretical linguistics, while instance-level parallelism us a topic in translation science. It is an occasional topic in computational linguistics when parallel corpora are concerned, but is treated almost exclusively from an application perspective. However, in the context of the current thesis, it is crucial to highlight the interdependencies between the two levels: annotation projection requires instance-level parallelism, and instance-level parallelism, in turn, requires concept-level parallelism. To outline the general picture, we therefore give a general overview of the two types of parallelism in this section. A more detailed discussion of parallelism on the level of predicate-argument structure is deferred to Section 3.2.

3.1.1. Concept-level Parallelism

Concept-level parallelism is concerned with a general property of linguistic theories, namely the *language independence of their conceptual inventory*: Although most linguistic theories are intended to be applicable to more than one language, they are generally developed, or at least verified, on a restricted set of languages. It is therefore an empirical question whether the conceptual inventory can be used – as it is, without any modifications – to completely and appropriately analyse another language.[1]

Concept-level parallelism is a desirable property, both from a theoretical and a practical point of view. In the development of theories, strong cross-lingual parallelism indicates that generalisations to other languages are possible, and can thus act as a criterion in the decision between rival theories. With respect to practical language processing, the more parallelism between descriptions for different languages, the lesser overhead

[1]We use the term "theory" here in a very broad sense that encompasses all complete descriptions of some linguistic level. This includes, for example, part-of-speech tagsets or named entity classifications.

there is for the development of linguistic resources or the implementation of software.

It is not surprising that the development of theories with complete concept-level parallelism (i.e., complete language independence) was an early research goal in theoretical linguistics. Notable proposals in this respect were Chomsky's (1957) universal grammar for syntax, and Katz and Fodor's (1964) theory of semantic decomposition into universal primitives. Unfortunately, developing these outlines into detailed and usable theories proved impossible in practice. This is mainly due to the vast number of "design decisions" languages take, which frustrate attempts at arriving at reliable generalisations. In semantics, for example, languages differ widely in the lexicalisation of differences between related concepts and the criteria by which these differences are made; as a result, it has been impossible to construct a complete set of semantic primitives. In syntax, languages can choose different strategies to encode argument positions, for example, by word oder (as in English), by morphological case (as in German), or by particles (as in Japanese).

From the computational linguistics point of view, an important insight from these studies is that *granularity of description* is a crucial factor influencing the degree of cross-lingual parallelism: The coarser the categories are, the more they abstract away from idiosyncrasies of particular languages, and therefore the higher the degree of concept-level parallelism. On the downside, of course, coarse categories are unable to represent the more detailed information present in the individual languages and can therefore only yield an incomplete description.

This insight has guided some more recent research activities which have tried to develop practical, detailed linguistic theories at the "most informative practical" level of conceptual parallelism. One prominent project following this strategy is PARGRAM (Butt, Dyvik, King, Masuichi, and Rohrer, 2002), which aims at developing parallel grammars for initially six languages in the context of Lexical-Functional Grammar (LFG). PARGRAM proceeds phenomenon by phenomenon, constructing parallel analyses for all languages. Only if this is not possible, the partial analysis common to all languages is determined and extended with specific information in individual languages. A similar approach has been taken in the development of EuroWordNet (Peters, Vossen, Diez-Ortas, and Adriaens, 1998), a multilingual ontology based on the WordNet paradigm: There is a

set of *core concepts*, which are assumed to be language-independent, and which are used in all languages. In addition, each language may define more specific concepts (i.e., language-specific distinctions).

In conclusion, perfect concept-level parallelism is an idealisation whose linguistic reality is questionable, and whose realisation for detailed linguistic theories is in any case impractical. On the other hand, projected annotation results in interpretable annotations for the target language only if the conceptual inventory used for the analysis of the source language exhibits concept-level parallelism. A high degree of concept-level parallelism is therefore crucial for the feasibility of annotation projection. Fortunately, it appears that at least for some levels of linguistic analysis, a fairly high degree of parallelism can be achieved, provided that the description is situated at an appropriate level of granularity.

3.1.2. Instance-level Parallelism

While concept-level parallelism is an important property of linguistic theories, it is insufficient for any investigations which are concerned with actual multilingual data such as parallel corpora: Concept-level parallelism is only concerned with the inventory of categories, not with the analysis of concrete utterances. To be able to characterise the relationship between mutual translations, we have to consider parallelism at the level of individual instances.

Formally, *instance-level parallelism* is a property of a linguistic framework with respect to a particular bilingual (or multilingual) corpus. We define that instance-level parallelism holds if for all pairs of *translationally equivalent* linguistic units in the corpus, the pairs of analyses according to the framework in question are identical. Evidently, instance-level parallelism implies concept-level parallelism: Instances can receive parallel (i.e., identical) analyses only if categories are parallel across languages (or can be mapped). Of course, the inverse does not hold: Concept-level parallelism does not imply instance-level parallelism. Consider Figure 3.1, a repetition of Figure 1.3. It shows a short English–German bi-sentence, tagged with simplified "core" parts of speech which we assume to be concept-level parallel. We can now ask whether the POS analysis exhibits instance-level parallelism, and take word alignment as an indication of translational equivalence. We find that for the first pair of translated

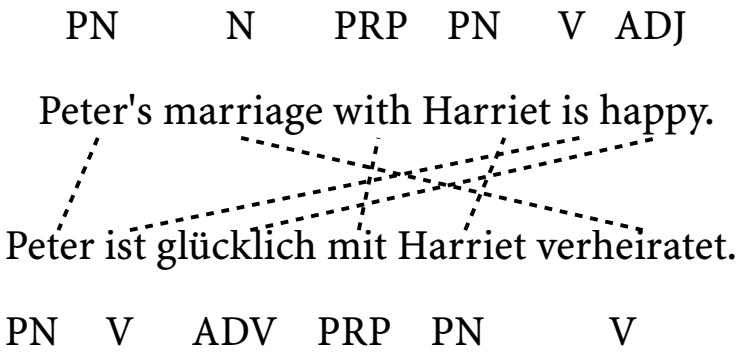


Figure 3.1.: A short English–German bi-sentence with part-of-speech analyses

words, *Peter's/Peter*, the analysis is in fact parallel: both words are analysed as proper nouns (*PN*). However, parallelism breaks down in the case of the second pair, *marriage/verheiratet (married)*: The English noun (*N*) *marriage* has been translated as the German participle (*V*) *verheiratet*: thus, the POS tag analyses are not parallel. Note that this non-parallelism implies a second non-parallelism, between the respective modifiers: English *happy* modifies a noun and is realised as an adjective (*ADJ*), while German *glücklich* is an adverb (*ADV*) modifying a verb. The analysis of the remaining words is again parallel.

Note that English and German are close enough that the English sentence could have been translated word by word, leading to a translation which would have been exactly parallel to the original English sentence in terms of part-of-speech tags, such as *Peters Ehe mit Harriet ist glücklich (Peter's marriage with Harriet is happy)*. However, the sentences are in fact not parallel on the instance level, the reason being that the translator did not choose this option.

This observation contains the key to the difference between concept-level parallelism and instance-level parallelism: while concept-level parallelism is only concerned with pairs of language *systems*, instance-level parallelism has to deal with pairs of concrete language *utterances* resulting from a *process of translation*. This places the translator at the centre of any serious account of instance-level parallelism, since his decisions during

translation determine the very shape of the target language text and all of its structural properties.

Current translation science emphasises the fact that translation is a semiotic process in which the different linguistic dimensions of a source utterance have to be reconciled with the often conflicting requirements of a target language (Matthiessen, 2001). Examples are surface-related properties (word order, lexical choice), semantics-related properties (predicate-argument structure), or even supra-sentential considerations (rhetorical structure, coherence). Which of the many possible translations will be chosen by the translator depends on a number of factors, such as the purpose of translation or language-specific preferences in formulation. In the example above, the translator might for example have considered the translation as a participle "more idiomatic".

From this perspective, it appears almost certain that on some linguistic level, the original and its translation will not be parallel. Still, instance-level parallelism is an idealisation that is methodologically very valuable: by assuming that perfect parallelism is the default case (the null hypothesis), deviations from this default can be identified and investigated. In fact, in translation science, such deviations are known as *translational shifts*, and are a topic of substantial long-term interest (Hawkins, 1986; van Leuven-Zwart, 1989; Cyrus, 2006). For typologically more distant languages, and for languages with culture differences, appropriate translation tends to demand more radical departures from the linguistic structure of the original (Truffaut, 1997). We will resume the discussion of translational shifts in Section 5.7 for our specific task of the induction of frame-semantic predicate classifications.

In computational linguistics, it is exactly the idealisation of instance-based parallelism that allows annotation projection to formulate cross-lingual transfer in the simple manner described in Section 1.3. The idealisation has been formulated most succinctly in Hwa et al.'s (2002) *direct correspondence assumption* which posits that all syntactic relations between mutually aligned nodes are instance-parallel.

In order to obtain a realistic picture of the chances of annotation projection, the actual degree of instance-level parallelism has to be gauged for each linguistic framework used to analyse a particular linguistic level, and each parallel corpus. Obtaining this figure is a important task, since the degree of instance-level parallelism for a given framework and a given cor-

pus constitutes an approximate *upper bound for the accuracy* of directly projected annotation information (without further preprocessing steps) within that setting.[2]

In this sense, the evaluation results for "direct projection" conditions for the existing studies on annotation projection, which were briefly mentioned in Chapter 1, can be interpreted as rough estimates for the instance-level parallelism in particular settings. In the following, we give a brief overview of these results.

Part-of-speech tags. Yarowsky et al. (2001) projected part-of-speech tags from English to French. Using a "core tagset" that distinguished only major morphosyntactic categories, they found instance-level parallelism of 76% for automatic word alignment, which could be improved to 85% by using manual alignment.

Word sense tags. Bentivogli and Pianta (2005) transferred EuroWordNet word sense tags from English to Italian. Instead of using existing parallel texts, they translated an English corpus with word sense annotation, which made is possible to produce both a free and a controlled translation. Automatic word alignment was used for all experiments. On the controlled translation, they obtained a precision of 88% and a recall of 71%. On the free translation, these numbers are somewhat lower at 85% precision and 63% recall.

Coreference. Postolache et al. (2006) projected coreference chains, i.e. binary relations between noun phrases which refer to the same discourse entity from English to Romanian. They used automatic word alignment and report projection accuracy on three texts. All results were very close, clustering around 87% precision and 62% recall.

NP chunks. Yarowsky et al. (2001) also transferred NP chunks information from English to two other languages. For French, they obtained

[2] In practice, an additional consideration with respect to the upper bound of annotation projection is the *complexity* of the annotation task: A comparatively simple task such as part-of-speech tagging will obtain high inter-annotator agreement in annotation, and does not require pre-processing. In contrast, more complex tasks like parsing or shallow semantic parsing are more prone to coder disagreement in annotation and preprocessing noise in practical implementations.

a precision of 43% and a recall of 48% with automatic word alignment, and 56% precision and 51% recall with manual word alignment. For Chinese, the numbers were 26% precision and 58% recall for automatic word alignment, and 47% precision and 61% recall for manual word alignment.

Dependency relations. Hwa et al. (2002; 2005) projected dependency relations from English to Spanish and Chinese. They found that direct projection resulted only in a comparatively small fraction of correct dependency links, namely 37% for Spanish and 38% for Chinese.

These results demonstrate large differences in instance-level parallelism between different levels of analysis. The general pattern that appears to emerge is that lexical-semantic properties in the widest sense (part-of-speech tags, word sense, and coreference) show a substantial degree of parallelism. When precision and recall are computed, precision is higher than recall. On the other hand, structural properties diverge considerably between translations. We observe the tendency that local structure (NP chunks) is still more parallel than global structure (dependency relations).[3] However, note that instance-level parallelism can only assessed on particular corpora and assuming particular linguistic theories (or annotation schemes, respectively); therefore, these numbers can only be generalised with due precaution. More systematic studies are clearly necessary before any stronger claims can be made.

3.2. Parallelism of Role-Semantic Analyses

We have established above that instance-level parallelism is the type of parallelism necessary for successful annotation projection, and we have also found that concept-level parallelism is a prerequisite for instance-level parallelism. In this section, we review the existing research on cross-lingual parallelism on the role-semantic level.

[3] Consequently, projection efforts for syntactic information such as Hwa et al. (2005) have often employed post-projection rewriting steps which transformed the source information according to linguistic knowledge about the source-target language pair.

The main part of this section will concentrate on FrameNet, our role-semantic framework of choice; the final subsection will discuss the case of alternative role-semantic frameworks, and will argue why they are more problematic in a cross-lingual setting.

3.2.1. Concept-level Parallelism of FrameNet

Almost all the evidence we cited in Chapter 1 on the cross-lingual parallelism of FrameNet annotation refers to concept-level parallelism. It concerns either the annotation of monolingual data for new languages with FrameNet frames (Burchardt et al., 2006b; Pallotta, 2005; Ohara et al., 2004; Subirats and Petruck, 2003) or more abstract lexicographic considerations (Boas, 2005). In both cases, frames developed by the Berkeley FrameNet project for English are re-used for the semantic analysis of other languages.

The reason why FrameNet frames show a large degree of concept-level parallelism is a result of the principle underlying its design. This basic principle is that frames, and their roles, are defined on a *conceptual* level by reference to the properties they exhibit and the inferences they allow. In this way, FrameNet abstracts away from many details of the lexical items it describes. While this granularity of description may not be fine enough for all purposes, it is crucial for the cross-lingual interpretability, as we have argued in Section 3.1.1, and FrameNet can be seen as a "most informative practical" level of description of predicate-argument structure.

However FrameNet does not rely exclusively on conceptual considerations, as ontologies usually do; membership of a predicate in a frame has to be "grounded" by the predicate's syntactic ability to realise the *core* frame elements.[4] As a consequence, non-parallelism on the frame level can occur when the subcategorisation patterns of predicates in new languages differ vastly from their English counterparts.

To date, no comprehensive quantitative studies of the concept-level parallelism of frame-semantic classes across languages exist. Evidence can however be drawn from the FrameNet annotation projects in other

[4]The *core* frame elements are those "that instantiate a conceptually necessary component of a frame". For example, the Speaker, Message and Addressee roles of the Commitment frame are all core roles, while Time, Place and Reason are not. See Ruppenhofer, Ellsworth, Petruck, and Johnson (2005) for a discussion.

languages mentioned above. The impression is that concept-level parallelism holds to a high degree, but is inversely correlated with typological distance from English. The recent establishment of a Global FrameNet group has led to the joint discussion of linguistic phenomena which can create problems for the concept-level parallelism of frames, and bears the perspective of an incremental modification of FrameNet towards more perfect concept-level parallelism, similar to the PARGRAM project (cf. Section 3.1.1).

An example for a problematic linguistic phenomenon is the general difference between "satellite-framed" languages such as English or German, where prepositions and adverbial particles tend to determine the meaning to a significant extent, and "verb-framed" languages where this information is encoded lexically (Talmy, 2000). The case of Motion verbs has been analysed in FrameNet terms for Spanish contrastively to English (Subirats and Petruck, 2003), and the latest FrameNet release (1.3) attempts to provide a representation of Motion verbs which abstracts over the language type.

Conflicts for particular predicates also arise occasionally between languages that are similar with respect to predicate-argument structure, such as English and German. A case in point is German *fahren*, which is the translation of both English *drive* and *ride*, which introduce two different frames in English: (*drive*: frame OPERATE_VEHICLE, subject is driver; *ride*: frame RIDE_VEHICLE, subject is passenger). In German, the context often does not disambiguate between the two frames, which makes it impossible for many instances to make the decision reliably. The appropriate level of description is a more abstract, using a frame that does not specify the role of the subject in detail. As a consequence, FrameNet has introduced the frame USE_VEHICLE, which subsumes both OPERATE_VEHICLE and RIDE_VEHICLE. While the frame is unlexicalised for English, it is the right level to describe the meaning of German *fahren*.

3.2.2. Instance-level Parallelism of FrameNet

In contrast to concept-level parallelism, almost no evidence exists with respect to the instance-level parallelism of FrameNet frames. The simple reason for this is that manual frame-semantic analysis of parallel, multilingual texts has so far only been proposed in the context of the Romance

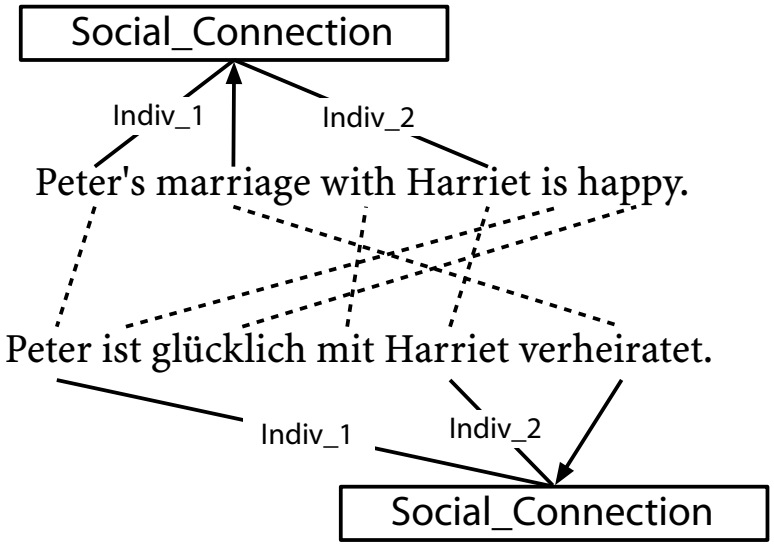


Figure 3.2.: A short English–German bi-sentence with frame-semantic analysis

FrameNet project (Pallotta, 2005), but has not materialised at the time of writing. However, since instance-level parallelism is central for annotation projection, we crucially need to obtain such evidence.

Intuitively, we would hope that role semantics in general, and frame semantics in its FrameNet instantiation in particular, stands a good chance of exhibiting a substantial degree of instance-level parallelism: In Section 3.1.2, we saw that "lexical" annotation generally exhibits more instance-level parallelism than "structural" annotation, and frame semantics is closer to the former than to the latter. We can bolster this intuition by extending the argument given in the last section for the concept-level parallelism of frame-semantic analysis: We argue that the largely conceptual definition of FrameNet frames means that a substantial fraction of possible translational shifts (primarily those related to the morphological and syntactic levels) does not lead to cross-lingual differences between the respective frame instances. As an example, consider Figure 3.2, which shows the frame-semantic analysis of the bi-sentence in Figure 3.1 (page 47). We find that the analyses are indeed parallel on the instance level; the improvements over the POS tag analysis, which was partially non-parallel, is a result of FrameNet's ability to abstract over the

part of speech of predicates.

If we try to spell out exactly which elements of the analysis correspond to one another across languages, we need to recall from Section 1.2 that frame-semantic analysis consists of two steps: (1), frame assignment, and (2), role assignment. To obtain a complete parallel frame-semantic analysis, clearly instance parallelism has to hold for both levels. Instance-level parallelism is therefore best decomposed into two components that we define below and illustrate in Figure 3.2.

Frame instance parallelism. We define frame instance parallelism to holds if for all pairs of source and target predicates that are translationally equivalent in a parallel corpus, both predicates evoke the same FrameNet frame.

In Figure 3.2, the translationally equivalent predicates *marriage* and *verheiratet* evoke the same frame, Social_Connection; this translation pair thus shows perfect frame instance parallelism.

Role instance parallelism. We define role instance parallelism to hold if for all parallel frame instances, each role that is realised overtly by one of the frames is also realised by the other frame.

In Figure 3.2, the parallel Social_Connection frames both realise the two roles Individual_1 and Individual_2; thus they also show role instance parallelism.

Role instance parallelism is conditional on frame instance parallelism: unless the frame introduced by two predicates is the same, the roles cannot be parallel. This is a result of the design of FrameNet, which defines roles at the level of individual frames. Even if two roles of different frames bear the same name, they do not necessarily afford the same interpretation (Fillmore, Wooters, and Baker, 2001). For example, the semantic role Source occurs in a large number of frames, but can stand for various source-like concepts in different prototypical situations.[5]

[5]Of course, the frame hierarchy might provide mappings between roles of different frames; however, such cases require careful consideration in a projection scenario and are best treated as cases of frame instance non-parallelism, at least in the first instance.

These two aspects of frame-semantic parallelism can now be tested in an *annotation experiment*: We manually create a gold standard corpus with independent frame-semantic annotation for three different languages (English, German, and French). Simple counting of the ratio of matching and mismatching frames and roles allows us to quantitatively assess the amount of cross-lingual parallelism. The details and results of this experiment will be described in Section 3.3 below.

In the preceding operationalisation of instance-level parallelism for frame semantics, we ignore the supra-frame structure (i.e., the relation between different frames in a sentence). This is in line with the general FrameNet approach to role-semantic analysis, which concentrates on the analysis of local predicate-argument structure. In fact, the FrameNet project until recently only annotated one predicate per sentence. Full-text annotation is now also available, but the mechanisms of interaction between different frames are only beginning to emerge (Burchardt, Frank, and Pinkal, 2005b; Padó and Erk, 2005). As a consequence, computational models of shallow semantic parsing in the FrameNet paradigm have always treated role-semantic analysis on a frame by frame basis, and so does annotation projection according to our definition from Chapter 1.

The projection of individual frames confers an important benefit with respect to a certain class of translational shifts, namely those which change relative position of two or more predicate-argument structures in a sentence. One well-known phenomenon of this class is *head switching*, where a first predicate governs a second predicate in one language, but is governed by it in the other language. As an example, consider the English sentence *Peter likes to sleep*, where the predicate *likes* has *sleep* as a complement. In the German translation *Peter schläft gerne (Peter sleeps willingly)*, however, it is *sleep* which is modified by *willingly*. In transfer-based machine translation, which assumes structural parallelism between source and target, head switching breaks this parallelism and therefore requires explicit treatment (Dorr, 1995). The projection of individual frames avoids this problem by concentrating on the level of individual predicate-argument structures, which are not affected by head switching: In both sentences, for example, *Peter* is analysed as the Sleeper of a Sleeping frame introduced by *sleep*. The fact that the syntactic embedding relations differ has no impact on the projection process, and does not have to be

taken into special consideration.[6]

We will further pursue the question of interactions between frames in Chapter 8, where we will use global frame structure to explore cases where frame instance parallelism breaks down and annotation projection according to the simple definition from Chapter 1 cannot be realised.

3.2.3. Parallelism in Other Role-Semantic Paradigms

In Chapter 1, we mentioned two other frameworks that, apart from FrameNet, figure prominently in current work on semantic roles in Computational Linguistics. These frameworks are PropBank (Palmer et al., 2005), and the Tectogrammatical Layer of the Prague Dependency Treebank (Hajičová, 1998). In this section, we discuss why these alternatives are more problematic than FrameNet for cross-lingual annotation projection.

PropBank. The aim of the PropBank project was to develop a theory-neutral, syntax-oriented annotation scheme for semantic roles, and to annotate a large corpus, the Wall Street Journal portion of the Penn Treebank, with these roles. The corpus has been completed in 2004 and is available through the Linguistic Data Consortium.

In PropBank, predicates are assigned coarse-grained sense tags. These senses are comparable in granularity to FrameNet frames within individual predicates, but they are not grouped into semantic classes across predicates. Semantic roles for arguments are defined for each verb sense separately, and are identified by indices (ARG-0 through ARG-n). In addition, there are universal argument roles which are written as ARGM with an affixed function tag. With respect to their semantics, the PropBank argument roles fall into two distinct classes: ARG-0 and ARG-1 correspond to Dowty's (1991) Proto-Agent and Proto-Patient roles, respectively. In contrast, the roles from ARG-2 onward are defined by syntactic criteria: Generally, higher indices correspond to increasing obliqueness (Keenan and Comrie, 1977). An effort was made to keep indices

[6] This is provided that no other considerations interfere; the example cited above is problematic insofar as the verb *likes* is translated as an adverb (*gerne*), whose status as a full-fledged predicate with an own argument structure is open to debate; however, this topic is outside the scope of the present thesis.

consistent within Levin's (1993) classes. Each role was also given a natural language mnemonic, but these have "no theoretical standing" (Palmer et al., 2005).[7]

It is the combination of a predominantly structural role definition with the absence of semantic classes which generalise roles across predicates that raises potential problems for using PropBank roles in an annotation projection scenario. More specifically, PropBank roles are defined by reference to structures *of the source language*, unlike FrameNet roles, which are defined by reference to participants of *(mostly) language-independent* conceptual situations. As a result, projected roles do not a priori have a well-defined semantics within the target language, without recourse to the source utterance. What is crucially necessary is to provide *grounding* for these roles, i.e. construct a definition in terms of the target language analogously to the definition for English provided by the PropBank dictionary. Without such a dictionary, concept-level parallelism cannot be guaranteed.

In an annotation projection scenario, the only source for this dictionary is the actual projection instances in the parallel corpus themselves. Note that this is problematic since it makes concept-level parallelism dependent on instance-level parallelism: only if the translation of *particular instances* in the corpus retains the structure of the role annotation can a consistent *general definition* be constructed for the roles in the target language. More specifically, the following well-formedness condition has to hold so that unique definition for roles in the target language can be provided: only one role may be projected onto a particular grammatical function of a target predicate.[8]

This well-formedness condition is problematic in practice, since each predicate in the target language can naturally be the translation of more than one source verb. In consequence, individual argument positions of these target predicate correspond to multiple argument positions of different source predicates. Due to the definition of semantic roles at the sense level in PropBank, these source argument positions are not guaranteed to be analysed consistently, and thus the target argument

[7] PropBank roles and example annotations are available online at the URL `http://www.cs.rochester.edu/~gildea/PropBank/Sort/`.

[8] More specifically, this has to hold at least within one subcategorisation frame. Of course, mappings may vary across diathesis alternations.

Language	Subject	Verb	Complement
English (I)	$[\text{Blair}]_{\text{ARG-0}}$	**begins**	$[\text{negotiations}]_{\text{ARG-3}}$
English (II)	$[\text{Blair}]_{\text{ARG-0}}$	**starts**	$[\text{negotiations}]_{\text{ARG-1}}$
English (III)	$[\text{Blair}]_{\text{ARG-0}}$	**launches**	$[\text{into negotiations}]_{\text{ARGM-DIR}}$
German	Blair	**beginnt**	mit Verhandlungen

Table 3.1.: Three synonymous English sentences with PropBank-style analyses.

position may be assigned more than one semantic role.

Table 3.1 illustrates this problem with a concrete example that consists of three synonymous English sentences and a German sentence that is a plausible translation of the English sentences. We assume for this argument that the German sentence occurs three times in a parallel corpus, once as translation of each English sentence. The German predicate *beginnen* has two arguments, a subject and a *mit*-PP. We now have to determine the set of ARG roles available for the predicate, and to define them in terms of their syntactic realisation. The subject, *Blair*, is unproblematic: all English source sentences analyse the corresponding argument, *Blair*, as ARG-0. Consequently, we can analyse the subject of the German *beginnen* as ARG-0 as well, which provides German syntactic grounding for the ARG-0 role. The situation is different for the German PP, *mit Verhandlungen*. The corresponding English phrase, *(into) negotiations* is variously analysed as ARG-1, ARG-3, or as an adjunct role (ARGM-DIR), and thus the three German instances will be annotated inconsistently. On which grounds should we choose one role over the others and include it in the definition of *beginnen*'s predicate as analysis for the PP?[9]

It is unclear how much of an impediment this kind of problem is for role projection; this question will have to be investigated empirically. It can however be speculated, based on the experiences from FrameNet annotation projects (cf. Section 3.2.2) that obtaining consistent projection results becomes increasingly difficult for typologically more distant languages. The use of the more general theta roles from VerbNet (Kipper, Palmer,

[9] In FrameNet, all three English predicates (*begin*, *start*, and *launch*) evoke the ACTIVITY_START frame, which guarantees consistent analysis.

and Rambow, 2002), which generalise the sense-specific PropBank roles, could potentially alleviate the consistency issue; but this claim must also be assessed empirically.

A more practical obstacle of using PropBank for annotation projection which should be mentioned nevertheless is that the PropBank annotation covers only verbal predicates, which rules out translation pairs whose English part is not a verb. This shortcoming is currently being remedied by the NomBank project which adds annotations for nouns to the PropBank corpus (Meyers, Reeves, Macleod, Szekely, Zielinska, Young, and Grishman, 2004).

Prague Tectogrammatical Structure. Functional Generative Description (FGD, Sgall, Hajičová, and Panevová (1986)) is a comprehensive theory of the linguistic system developed in Prague. It assumes three layers of representation: morphology, surface syntax, and deep syntax. The layer of deep syntax is called Tectogrammatical Structure.

In FGD, tectogrammatical structure is assumed to reflect the "literal meaning of the sentence" and, at the same time, to provide the interface between linguistic theory (the domain of FGD) on one hand and semantic interpretation or discourse analysis on the other hand, which are assumed to be interdisciplinary endeavours (Hajičová, 2000). Formally, the tectogrammatical structure is a labelled dependency tree whose nodes correspond to content words of the sentence, and whose edge labels describe the relation between these content words. Since the literal meaning is supposed to be identical with deep syntax, tectogrammatical structure expresses not only predicate-argument relations like semantic roles in the sense of PropBank or FrameNet, but provides a complete account of the structure of the sentence, including also head-modifier and other relations such as coordination or apposition.

In an annotation projection scenario, complete instance-level parallelism of the tectogrammatical structure thus presupposes parallelism not only on the semantic, but also on the syntactic level, which is difficult (compare the results on dependency parallelism in Section 3.1.2). It is therefore clear that annotation projection can only be successful for parts of the tectogrammatical structure, and presumably in particular for the predicate-argument relations, which are closest in spirit to semantic roles. For the predicate-argument relations, a problem similar to PropBank

arises: the labels (called *functors*) used by FGD to describe tectogrammatical relations are defined predominantly structurally (see Lopatková and Panevová (2005) for details), and it is unclear how well these structural considerations transfer to new languages.

A more general consideration is that tectogrammatical structure is defined by the application of certain transformation rules on the surface syntactic structure. It is a matter of future work to determine the implications of a close coupling between syntax and semantics for instance-level parallelism. A practical problem in resolving this issue is that, to our knowledge, no worked-account of surface syntax in the FGD framework exists for other languages such as English (Sgall, 2000).

3.3. Assessing the Cross-lingual Parallelism of Frame-Semantic Annotation

As mentioned in the preceding discussion, there are currently no studies on the actual degree of instance-level parallelism of frame-semantic annotations, neither on the frame level nor on the role level. This section describes an annotation study aimed at addressing this shortcoming. In the first phase of the experiment, we created a gold standard parallel sample corpus of 1,000 English–German sentence pairs sampled from EUROPARL. Both sides of this corpus were annotated manually with FrameNet frames and roles by two independent annotators. This corpus allows us to empirically estimate the *degree of instance-level parallelism*, thereby gauging the feasibility of automatic annotation projection for frame-semantic annotation. In the second phase, the French sentences corresponding to this sample were annotated, to validate our findings on a second language pair involving a language pair.

A second, equally important consideration was to create a *gold standard corpus* that could be used for the evaluation of our role projection models and of shallow semantic parsers in general. To our knowledge, our corpus is currently the only parallel corpus with manual role-semantic analysis. The annotations are freely available for research purposes and can be downloaded from the URL `http://www.coli.uni-saarland.de/~pado/projection/`.

3.3.1. Creation of an English–German Bitext

We first describe the extraction and annotation of the English–German bitext.[10]

Sample Selection. The least biased strategy for drawing a sample from a corpus is by random sampling. However, in the present setting, random sampling raises (at least) two problems. First, FrameNet is as yet incomplete. As a result, randomly selected sentences may contain readings of predicates which are simply not covered in with the current FrameNet frame inventory; an example is the "greeting" sense of *hail* which is not represented by any frame. Second, EUROPARL is a professionally translated corpus and likely to contain a large fraction of free translations that do not preserve the frame across languages. However, since role instance parallelism is conditional on frame instance parallelism (cf. Section 6.2), role instance parallelism can only be assessed on sentences with parallel frames.

Note that both of these problems cannot be detected automatically, but only after annotation has taken place. As a result, sampling runs the risk of yielding a dataset of which only a small fraction can be analysed with FrameNet in the first place (since many instances are outside FrameNet's coverage), and which contains an even smaller number of bi-sentences suitable for estimating the degree of role instance parallelism (since they do not show frame instance parallelism).

To avoid this issue, we decided to use a more informed sampling strategy for obtaining an English–German sample. Our selection was driven by two existing resources, the English FrameNet and SALSA, a FrameNet-compatible predicate classification for German currently under development (Erk et al., 2003). We used the intersective word alignment obtained from GIZA++ (see Section 2.2.2 for details) to identify one-to-one aligned English–German predicate pairs. We further required these predicates to be both listed in FrameNet and SALSA for at least one common frame. The resulting corpus contains 83 frame types and 696 predicate pairs (on the type level), exemplifying 265 unique English and 178 unique German lemmas. Sentence pairs were grouped into three bands according to their

[10] This bitext was prepared at Saarland University.

frame frequency (High, Medium, and Low). We randomly selected 380 instances of predicate pairs from each band, resulting in a sample of 1,140 bi-sentences.

This sampling procedure produces a corpus sample which is not entirely random, but which we consider a realistic input for the role projection task. However, this does not mean that all sentences which do not show frame instance parallelism are outside the scope of automatic processing in an annotation projection framework. We discuss how such cases can be handled in Chapter 8.

Annotation and Inter-Annotator Agreement. The English–German bitext was annotated manually by two annotators with native-level proficiency in German and English. For every predicate, the annotation task involved two steps: (a) selecting the appropriate frame and (b) assigning the semantic roles it instantiates to the constituents of the sentence under consideration. Annotators were provided with detailed guidelines, which are given as Appendix B. During annotation, they had access to parsed versions of the sentences in question (see Section 2.3.1 for details), and to the FrameNet and SALSA resources for English and German.

The annotation proceeded in three phases: a training phase (40 bi-sentences), a calibration phase (100 bi-sentences), and the annotation of the main dataset (1,000 bi-sentences). During training, annotators were acquainted with the annotation style. In the calibration phase, each bi-sentence was doubly annotated to assess the inter-annotator agreement and revise guidelines in case of low agreement (this turned out not to be necessary). Finally, in the main annotation phase, each of the 1,000 bi-sentences in the main dataset was split and each half randomly assigned to one of the coders for single annotation. We ensured that no annotator saw both parts of any bi-sentence in order to guarantee independent annotation of the two halves of each bi-sentence. Annotation proceeded predicate-wise; each coder annotated approximately the same amount of data in English and German.

The first two columns in Table 3.2 show inter-annotator agreement on the calibration set for English and German. In addition to the widely used Kappa statistic, we computed a number of different agreement measures: the ratio of frames common between two sentences (Frame Match), the ratio of common roles (Role Match), and the ratio of roles with identical

Measure	English	German	French
Frame Match	0.90	0.87	0.87
Role Match	0.95	0.95	0.89
Span Match	0.85	0.83	0.72
Kappa	0.86	0.90	0.75

Table 3.2.: Monolingual inter-annotator agreement on the calibration set (English and German) and on the complete dataset (French)

spans (Span Match). As can be seen from the table, annotators tend to agree in frame assignment; disagreements are mainly due to difficult distinctions between closely related frames (e.g., between AWARENESS and CERTAINTY). Annotators also agree well on what roles to assign (Role Match is 0.95 for both English and German). Agreeing on the exact role spans is a harder problem, since it involves additional syntactic decisions.

Our results show strong agreement, thus demonstrating that the task is well-defined. Unfortunately, no published agreement figures for English FrameNet annotation are available for comparison. Our numbers are comparable to published figures for the German FrameNet annotation in the SALSA project: Burchardt, Erk, Frank, Kowalski, Padó, and Pinkal (2006a) report 85% raw inter-annotator agreement for Frame Match and 86% inter-annotator agreement for Span Match. The figures we obtained are 2% higher for Frames and 2% lower for Spans. Given the small size of our calibration set, these differences are not statistically significant.

3.3.2. Creation of an English–French bitext

Any study restricted to a single language pair runs the risk of obtaining results that are biased by specific properties of that language pair. To avoid this problem, we repeated the annotation effort for a second language pair, in order to obtain a more general picture of the applicability of frame-semantic annotation projection. We chose French as our second target language. French belongs the Romance language group, while English and German are both Germanic languages. Thus, the comparison of the results for the two target languages, German and French, can reveal interesting

differences across languages and language groups.[11]

Sample Selection. Due to the absence of manually produced frame-semantic resources for French comparable to FrameNet, we could not follow the same sampling strategy as for German. Instead, we chose to annotate the French translations of the 1000 sentences making up the English–German bitext. This strategy has two advantages. First, it results in a parallel corpus of three languages, allowing for controlled comparisons between the two target languages (see above). Second, the English–French bitext is more representative of the entire EUROPARL corpus than the English–German bitext: The predicate pairs have not been selected on the basis of (potential) frame parallelism, as was the case for English–German. In this sense, the English–French bitext provides a more realistic estimate of the degree of frame instance parallelism in an unfiltered parallel corpus.

While extracting the French text, we found that sixty sentences of the 1000 English sentences were sentence-aligned to empty French sentences. These sentences were removed; the French corpus therefore consists of 940 French sentences. The automatically produced word alignments (cf. Section 2.2) were used to determine which French predicates corresponded to the relevant English frame-evoking elements.

Annotation and Inter-Annotator Agreement. The 940 French sentences were annotated with frames and frame elements independently by two annotators, both of which were native speakers of French. They followed the same annotation schemes used for the annotation of the English–German bitext (see Section 3.3.1 for details). During annotation, the annotators did not have access to the English–German bitext nor its frame-semantic annotations, but were provided with parsed versions of the sentences in question (see Section 2.3.1), and with the FrameNet resource for English.

The double annotation makes it possible to compute inter-annotator agreement on the entire dataset. The results are listed in the rightmost column of Table 3.2. Overall, we find that the strong inter-annotator

[11] The annotation was provided at INRIA Lorraine in Nancy in the context of the French FrameNet project. We would like to acknowledge their help and cooperation.

agreement results carried over. Frames can be assigned equally well for French as for German, leading to an equal Frame Match agreement of 87%. Role Match agreement (which only takes into account whether roles were assigned at all) is also high at 89%, albeit around 5% worse than for English and German. We attribute the difference to the fact that the annotation guidelines used for French annotation were originally developed for German and English annotation (cf. Section 3.3.1) and probably were not sufficient to resolve all problematic cases in the new language.

The main difference to the first language pair is the Span Match agreement that is clearly lower for French (72%) than for German (83%). The fact that Role Match does not show the same drop (see above) indicates that this is mainly a problem of assigning semantic roles to constituents obtained from the French automatic syntactic analysis. Further analysis of the data confirmed that frequent cases of incomplete and erroneous constituents compelled annotators to assign many semantic roles to more than one syntactic node, which adversely affects Span Match (Erk et al., 2003). In the following example, the frame element ADDRESSEE of the frame COMMITMENT, evoked by *menace (threat)*, spans a prepositional phrase including a coordination. Since no overarching constituent was found for the phrase, the role has to be assigned to a total of four smaller constituents, indicated by round brackets.[12]

(3.1) Peu m'importe que les menaces ou les crimes soient
I don't mind whether the threats or the crimes are
commis [(contre un Serbe)$_{PP}$, (un Rom)$_{NP}$, (un Bosniaque ou
committed [(against a Serb)$_{PP}$, (a Roma)$_{NP}$, (a Bosnian or
un Albanais)$_{NP}$]$_{Addressee}$.
an Albanian)$_{NP}$]$_{Addressee}$.

3.3.3. Evaluation

The annotated corpora also allows us to assess the degree of semantic cross-lingual parallelism. We computed both frame (instance) parallelism

[12] We provide glosses for all foreign language material in examples. Glosses can be distinguished from original English EUROPARL material by their position *underneath* the foreign language material.

Measure	Precision	Recall	F-score
Frame Parallelism	0.72	0.72	0.72
Role Parallelism	0.91	0.92	0.91
Frame Parallelism	0.65	0.74	0.69
Role Parallelism	0.88	0.87	0.88

Table 3.3.: A quantification of the cross-lingual instance-level parallelism for the language pairs English–German (above) and English–French (below)

and role (instance) parallelism. Frames were counted as matching when the same frame was annotated for a parallel predicate pair. Roles were counted as matching when they occurred in both halves of a bi-sentence regardless of the role spans, provided that the frames matched (recall that role instance parallelism presupposes frame instance parallelism). Since role spans are not easily comparable across languages, Span Parallelism was not applicable in this cross-lingual setting.

The results (shown in Table 3.3) were computed on the complete English–German (1,000 bi-sentences) and English–French (940 bi-sentences) bitexts. To facilitate comparisons with the output of our automatic projection methods, we present agreement in terms of precision, recall and F-score (see e.g. Baeza-Yates and Ribeiro-Neto (1999)), treating the annotations of the target language (i.e., German and French) as gold standard against which we compare the English annotations. This evaluation scheme directly gauges the usability of English as source language for annotation projection: A recall of less than 100% means that frame or role instances exist in the gold standard for the target language which are not present in English, i.e., which cannot be retrieved by annotation projection from English annotation. Conversely, imperfect precision indicates additional English frame or role instances, whose projection results in erroneous annotations of the target language.

Frame instance parallelism. We find a surprisingly similar degree of frame instance parallelism between the two language pairs. For English–German, we obtain F=0.72; the result for English–French is only slightly

lower for English–French at F=0.69. In other words, around 70% of all relevant bilingual predicate pairs reliably evoke the same frame in both bitexts. This result demonstrates substantial cross-lingual frame instance parallelism that is not restricted to closely related language pairs (e.g., English–German). The degree of instance-level parallelism that we found also corresponds well to results for other lexical-semantic annotation (cf. Section 3.1.2), confirming our intuitions about the nature of the frame assignment task.

Our interpretation of these numbers must however take *inter-annotator agreement* into account: Cross-lingual parallelism is estimated on sentence pairs which were annotated by two different annotators to guarantee independent annotation for each language. As a result, these numbers do not only incorporate the cross-lingual variance caused by the translation proper, but also variance due to disagreements among coders which naturally arise in any annotation study. A realistic upper bound for cross-lingual parallelism is therefore less than 100%, and the monolingual inter-annotator agreement for Frame Match can be considered as a more realistic upper bound (around 90% in our annotation setting). The degree of cross-lingual parallelism should therefore not be estimated as 30% below the upper bound, but only 15–20%.

A marked difference between the language pairs which we find for frame instance parallelism is the relation between precision and recall. The English–German bitext shows very similar precision and recall values, since almost the same number of frames and roles was annotated for both languages. In contrast, the precision in the English–French bitext is almost 10% lower than the recall, indicating that not all English frames had counterparts in the French gold standard. In fact, the difference in recall corresponds to around 120 French sentences which the annotators decided to leave unannotated, due to the lack of suitable FrameNet frames.

This difference can be traced back to the differences in the corpus creation procedure for the two bitexts: the English–French bitext included a considerable number of French sentences in the corpus whose predicates could not be covered by FrameNet. However, the sample corpora for the two language pairs are still crucially similar in that the *overall* frame instance parallelism is almost equal, even though we should expect a lower result for the English–French both from the perspective of language similarity and corpus creation. In the following, we will assume that

the two sampling strategies we have employed can be on the whole be disregarded in the interpretation of quantitative experimental results.

In sum, we conclude that frame instance parallelism holds for the majority of cases, largely independent of the language pairs and sample corpora we considered. We will therefore assume frame instance parallelism as a working hypothesis for the purposes of cross-lingual induction of predicate classifications (Part II) and the cross-lingual induction of semantic roles (Part III). We will discuss translational shifts, the mechanisms leading to non-parallel frames, in Section 5.7. In addition, Chapter 8 provides a generalisation of the principle of frame instance parallelism, which allows us to treat certain cases of individual non-parallel frames as translationally equivalent.

Role instance parallelism. We find a still higher, and more consistent, degree of parallelism on the level of semantic roles. In the English–German bitext, 91% of the semantic roles of matching frames are preserved in the translation process. The English–French bitext shows almost the same level of parallelism with 88% matching roles.[13] The influence of the language difference is even smaller when we again consider monolingual inter-annotator agreement (Role Match) as a more realistic upper bound: for both language pairs, agreement on the cross-lingual dataset is only marginally lower than the agreement on the monolingual dataset.

For roles, there is virtually no difference between precision and recall for either language pair, which indicates that role mismatches are not linked to general lexicalisation differences between languages requiring different frame-semantic analyses. Rather, further analysis revealed that the remaining mismatches are frequently cases of passivisation or infinitival constructions leading to role elision in one of the languages, like the following:

[13]Note that in the case of role parallelism, the sampling strategy does not have significant bearing on the dataset, since role parallelism is only computed on instances with parallel frames.

(3.2) So I ask that [Ireland]$_{\text{Content}}$ be **remembered** in this particular case.
Ich möchte deshalb darum bitten, dass [man]$_{\text{Cognizer}}$ in diesem
I would like therefore to ask, that [one]$_{\text{Cognizer}}$ in this
speziellen Fall auch [an Irland]$_{\text{Content}}$ **denkt**.
particular case also [of Ireland]$_{\text{Content}}$ **thinks**.

In this example, English uses a passive construction which leaves the deep subject position unfilled. In contrast, German uses an active construction where the deep subject position has to be filled; however, the filler is a semantically light pronoun.

In sum, we find that semantic roles exhibit almost perfect instance-level parallelism. We consequently adopt role instance parallelism as a working hypothesis as well. This assumption will be instrumental for the role projection methods developed in Part III, the cross-lingual induction of semantic roles.

3.4. Summary

In the present chapter, we have discussed the cross-lingual parallelism of linguistic analyses. In Section 3.1, we have introduced the more detailed categories of *concept-level* parallelism (the applicability of linguistic frameworks for the analysis of more than one language) and *instance-level* parallelism (the identical analysis of corresponding text in a parallel sentence). Importantly, we have established that successful annotation projection on any linguistic level presupposes a substantial degree of instance-level parallelism.

Next, Section 3.2 has analysed the design principles of frame semantics, finding that frames and semantic roles, which are characterised on the conceptual level, exhibit a high degree of concept-level parallelism. This has been verified by showing the usability of the English FrameNet resource for the annotation of corpora in other languages.

To also assess the degree of instance-level parallelism empirically, we have described in Section 3.3 how a 1000-sentence corpus sample from EUROPARL was annotated with frame-semantic analyses, i.e., frames and semantic roles. The annotation was carried out in parallel for three languages (English, French, and German). We found substantial parallelism

for both frames and semantic roles (roughly 70% for frames and 90% for semantic roles) and both language pairs. This result yields support for the applicability of annotation projection for frame-semantic analyses.

Part II.

Cross-lingual Induction of Frame-Semantic Predicate Classes

4. A Framework for the Projection of Frame-Semantic Predicate Classes

A frame-semantic predicate classification lists the possible frames for each predicate in a language. This chapter discusses the induction of such a classification for languages in which it does not yet exist. We first discuss the usefulness of frame-semantic predicate classifications (Section 4.1) and possible strategies for their construction (Section 4.2). Next, Section 4.3 presents a data-driven architecture for the cross-lingual induction of frame-semantic predicate classes which combines cross-lingual annotation projection with a subsequent filtering step. The architecture specifies two filter types which are aimed at generality and language-independence by relying mostly on shallow distributional information. We finally develop concrete filters by analysing the major error sources in the data (Section 4.4).

4.1. Motivation

A *frame-semantic predicate classification* is a lexicon resource which lists available frames for predicates (cf. Section 1.4). Frame-semantic predicate classifications form a subset of *semantic predicate classifications*, namely those that use FrameNet frames as semantic classes. Other prominent semantic predicate classifications are, for example, WordNet (Miller, Beckwith, Fellbaum, Gross, and Miller, 1990), whose classes consist of synonymous words, and Levin's (1993) verb classification which groups

verbs by their alternation behaviour.[1] Generally speaking, the information contained in a semantic predicate classification can be interpreted from two complementary angles:

- *Within* predicates, these semantic classes model the set of readings (or senses) of a predicate. This list forms the basis for almost all current computational approaches to word sense disambiguation (WSD, see Agirre and Edmonds (2006) for an overview): WSD is usually modelled as a classification task, in which each reading corresponds to one target class.

- *Across* predicates, semantic classes encode semantic properties which all predicates in the class have in common. This generalisation is often crucial to alleviate the sparse data issue that is ubiquitous in empirical lexical semantics, especially when it comes to lexical relationships. For example, selectional preferences are almost universally represented by reference to WordNet classes (Resnik, 1996); this builds on the hypothesis that all words in a WordNet class are equally well (or badly) suited to fill one particular argument position. Similarly, Lapata, Keller, and McDonald (2001) use class-based smoothing to estimate the plausibility of unseen adjective-noun pairs.

Frame-semantic frames, as semantic classes, express their own particular set of semantic generalisations, which have been discussed in detail in Section 1.1.1. At this point, we only reiterate and discuss the two most important roles of a frame-semantic predicate classification for computational modelling.

The first role is as a resource for shallow semantic parsing (cf. Section 1.2), and in particular for frame assignment. Recall that there is an interdependence between frame assignment and role assignment: Different frames introduce different role sets, and the frame assignment for a particular instance therefore has direct implications on the appropriate

[1] We use the term frame-semantic *predicate classification* rather than *lexicon* to indicate that at this stage, we limit our focus to the (binary) mapping between lexical units and semantic classes. We ignore, for example, frequency information or information about the syntax-semantics mapping that could be expected from a full-fledged frame-semantic *lexicon*.

Example	Frame	
He **asks** [about her health]$_{\text{Topic}}$.	QUESTIONING	sich erkundigen
He **asks** [for the jam]$_{\text{Message}}$.	REQUEST	bitten
He **asks** himself [if it is true]$_{\text{Content}}$.	COGITATION	sich fragen

Table 4.1.: Senses, corresponding frames, and German translations of the verb *ask*

role set for the instance. This is illustrated in Table 4.1, which summarises the FrameNet analysis for the English verb *ask*. According to FrameNet, *ask* can introduce three different frames; for each sense of *ask*, the semantic arguments in the example sentences receive a different semantic analysis, i.e. semantic role label. As a consequence, a comprehensive frame-semantic predicate classification is a prerequisite for shallow semantic parsing of a new language: unless it is known which frames can be evoked by a predicate, no analysis is possible.

The second role concerns annotation projection, and thus is especially relevant in the context of the current thesis. Recall from Section 3.2 that there is no perfect frame instance parallelism – in other words, not every pair of aligned predicates in a parallel corpus evokes the same frame. Consider the following example, where the English expression *take place* introduces the frame EVENT, which depicts the enlargement as an EVENT that is happening without the presence of a causally involved agent. In contrast, the German translation *vornehmen* evokes the frame INTENTIONALLY_ACT, which implies that the enlargement is an ACT performed by some (not overtly realised) AGENT:

(4.1) [An enlargement of the EU]$_{\text{Event}}$ is **taking place**.

[Eine Erweiterung der Union]$_{\text{Act}}$ wird **vorgenommen**.
[An enlargement of the union]$_{\text{Act}}$ is being **enacted**.

Such *translational shifts* will be discussed in detail in Section 5.7; at this point, it should suffice to point out that role annotations which are projected between non-parallel frames are bound to be wrong, or at least nonsensical.

This observation suggests a two-step approach for the projection of frame-semantic information: first, a frame-semantic predicate classification for the target language is being induced. This classification can then serve as a *filter* on the projection of roles by allowing projection only for pairs of predicate instances evoking the same frame. Note that for this application, high precision appears to be more crucial than high recall, a point we will return to several times in the course of this part.

4.2. Constructing Frame-Semantic Predicate Classifications

In the preceding section, we have shown that frame-semantic predicate classifications are a central resource for the frame-semantic analysis of a language. English is currently the only language for which a comprehensive frame-semantic predicate classification exists, in the form of the definitional part of the FrameNet database, i.e. the frames, their definitions and corresponding predicates. Almost all other languages, in contrast, suffer from *resource scarcity* with respect to frame-semantic predicate classifications.

Fortunately, in Section 3.2.1 we have come to the result that the frames themselves (i.e., the predicate classes) exhibit a large degree of concept-level parallelism, and can therefore, by and large, serve as semantic classes for other languages as well.[2] What is necessary to extend the frame-semantic paradigm to a new language, therefore, is to assign lexical units of the new language to the given set of frames. There are, in principle, three approaches to this task.

Manual Lexicography. This is the approach followed by the Spanish and Japanese FrameNet projects as well as the manual track of the German SALSA project (cf. Section 1.1.1). This strategy shares the usual benefits and problems of manual resource building: On the one hand, its results are of a very high quality that cannot usually be obtained by automatic means. On the other hand, the work has to be repeated for each new language, which is a long and laborious process.

[2] Of course, this does not preclude problems with individual frames; cf. Section 3.2.1.

Automatic Induction. There has been considerable research on learning semantic classifications "from scratch". Since semantic properties generally show a weaker correlation with surface phenomena than e.g., syntactic properties, this has turned out to be a difficult problem. It is compounded by the vast space of possible semantic generalisations that might plausibly underlie semantic classes and which are difficult to encode in the form of priors. Good results have been obtained when the desired semantic properties could be tied to clear morphosyntactic realisation regularities (Merlo and Stevenson, 2001; Schulte im Walde, 2006). In the case of FrameNet, there is one study which has attempted to "re-induce" FrameNet frames for English (Green, Dorr, and Resnik, 2004). See Section 5.5 for a discussion; for the moment, we only note that successful induction required a rather sophisticated combination of several resources.

Cross-lingual Transfer. The most promising strategy at the present time is to bootstrap the resource for a new language by *transferring* the information contained in the FrameNet resource across languages, that is, from English to the new language. More specifically, the list of English frame-evoking elements given in FrameNet for each frame is used as input to some procedure that produces the list of frame-evoking elements of the other language for that frame. This strategy effectively combines the advantages of both above approaches: we can produce a resource with high quality, to the extent that the quality can be retained during transfer. The procedure can also be repeated for several languages without additional manual cost, provided that it is sufficiently language-independent. We will thus pursue this strategy.

Cross-lingual transfer of FrameNet information can be effected in a number of ways, which can be grouped into two broad classes: *Data-driven approaches* use an unprocessed parallel corpus or one with shallow linguistic analysis to transfer the information, while *resource-driven approaches* use multilingual linguistic resources for the task, such as dictionaries or ontologies.

The only example of a resource-driven approach, to our knowledge, is the study by Fung and Chen (2004), who construct a Chinese frame-semantic predicate classification by mapping English FrameNet entries to

concepts listed in HowNet, an on-line ontology for Chinese[3]. The cross-lingual transfer is performed by using two bilingual dictionaries, with a subsequent monolingual disambiguation step. Fung and Chen obtain an F-Score of 82% as the final quality of their induced resource; however, their evaluation was only exemplary, restricted to six predicates from six frames.

We do not follow this approach for a number of reasons. First, resource-driven approaches strongly rely on the existence of large-coverage, high-quality language resources. Unfortunately, large translation dictionaries are not readily available for all languages, and multilingual ontologies even less so. Second, such resources are usually designed for human readers. This means that even if they are manually created (and presumably clean), they are prone to inconsistencies and omissions. A third, related problem is translational ambiguity: bilingual dictionaries often contain one-to-many correspondences, of which only some translations are general, while others may only be felicitous in specific contexts. All of these issues are difficult to resolve automatically, since translational equivalence is rarely quantified in manually constructed dictionaries.

On the other hand, quantitative information is readily available from parallel corpora, for example in the form of word alignment probabilities. Therefore, we pursue a data-driven approach. In particular, we adopt an instance of *cross-lingual annotation projection* (as introduced in Section 1.3): For each known English frame-evoking element, we observe its translations in a parallel corpus, according to an automatically obtained word alignment. These translations represent *candidates for frame-evoking elements* of the same frame in the other language. The assumption behind this strategy is that word alignment can serve as a proxy for frame instance parallelism (cf. Section 3.2.2): if a bilingual pair of expressions is found to be word-aligned in a corpus, it can evoke the same frame.

This assumption is clearly an oversimplification: Due to the polysemy of frame-evoking elements in the source language, as shown in Table 4.1, not all translations of a source lemma will be good frame-evoking elements for the same frame. Nevertheless, we retain this assumption for the moment: It allows us to formulate a baseline model, which can be subsequently

[3]See `http://www.keenage.com/zhiwang/e_zhiwang.html`.

refined. In addition to resulting in a "division of labour", this strategy allows us to gauge the actual influence of polysemy-related errors in the FEE projection task.

Projection of frame-evoking elements is illustrated in Figure 4.1. Consider the English frame STATEMENT, and assume that it has just one frame-evoking element, SAY. The comprehensive set of German FEE candidates is given by the set of German lemmas of which at least one instance is word-aligned to an instance of *say*. In the figure, the German FEE candidates for STATEMENT are consequently *sagen (say), meinen (say)* and *beispielsweise (for example)*.

The major benefit of this data-driven strategy is its ability to deal with low-density languages. The only kind of data required are parallel corpora, which are much easier to obtain than wide-coverage resources with a high quality. For example, Resnik and Smith (2003) have developed a method for automatically harvesting parallel corpora off the web. We will also show that high-quality frame-semantic classifications can be obtained solely with shallow linguistic analysis (in particular, lemmatisation and part-of-speech tags), which are usually among the first analysis tools developed for any language. A second benefit of the data-driven strategy is the wealth of linguistic patterns in the corpus: A corpus can provide a frame-semantic lexicon with a wider range of translations than can be expected from a dictionary. In addition, domain-specific corpora can be used to obtain frame-evoking elements for these domains, analogously to the work on word sense by Koeling, McCarthy, and Carroll (2005).

To verify that a data-driven strategy in fact confers advantages over a resource-driven one, we will compare our data-driven strategy against cross-lingual transfer based on wide-coverage dictionary resource with high quality in Section 5.4.

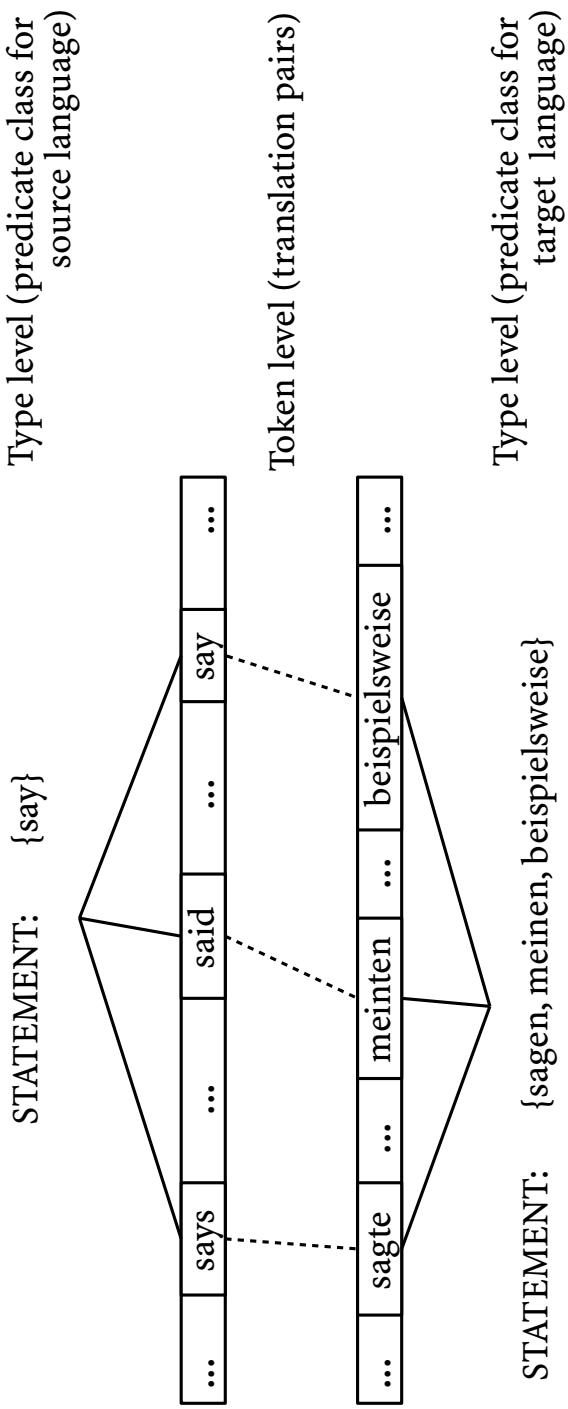


Figure 4.1.: Noisy induction of German FEEs for the frame STATEMENT, using translation pairs in a parallel corpus for an English FEE of the same frame.

4.3. Projection with a Generate-and-filter Strategy

In this section, we describe and define the general architecture which we use for projecting frame-evoking elements across languages, and introduce the notation we will use to formalise the projection regime and which is summarised in Table 4.2.

One central observation which can already be made in Figure 4.1 is that projection based on word alignment links is evidently imperfect: not all of the German FEE candidates which we have obtained for STATEMENT can actually evoke the frame. In particular, *beispielsweise (for example)* is not a STATEMENT expression. In the general case, such errors can arise from three main sources: noise in word alignments, polysemy of frame-evoking elements (briefly mentioned above), and translational shifts leading to frame instance non-parallelism. These error sources will be discussed in detail in Section 4.4.

To address this problem, we used a *generate-and-filter* approach: In a first *generation step*, we transfer frame annotations across word alignment links in a parallel corpus to generate a noisy, but comprehensive list of target language FEE candidates for each frame. More specifically, each word in the target language which has been seen at least once as translation of a frame-evoking element for a frame is considered as FEE candidate for this frame. In a second *filtering step*, we then compute different statistics over the corpus which allow us to sort out unreliable candidates to obtain a high-precision lexicon.

Such a generate-and-filter approach makes optimal use of the parallel corpus as a source of linguistic information: In the generation step, we exploit the large amount of lexical variation present in the different translations of the parallel corpus. In the filtering step, we take advantage of the local and distributional patterns of translation pairs in the corpus.

The result we obtain is a very clean, but possibly small, *seed lexicon* for the target language. Our rationale for optimising precision over recall is that high precision seed lexicons can be extended with the help of *monolingual* resources, which are usually more plentiful than multilingual ones; on the other hand, improving the quality of a large, noisy semantic lexicon is a difficult task. In addition, we will show that even this precision-

$l_s \in L_s, l_t \in L_t$	lemmas (word types) of source and target language
$i_s \in I_s, i_t \in I_t$	instances (word tokens) of source and target language
$f \in F$	frame in FrameNet
$i_s \rightsquigarrow_1 i_t$	i_s and i_t are one-to-one aligned
$tp(i_s, i_t, f)$	i_s and i_t form a translation pair for f
$fee(f, l_s)$	l_s is a known FEE for f
$wsd(f, i_s)$	f is the disambiguated frame for instance i_s
$cont(l)$	l is a content word (adjective, noun, or verb)
$cand(l_t, f)$	l_t is a FEE candidate for f
$lemma : (I_s \cup I_t) \to (L_s \cup L_t)$	Function assigning a lemma to an instance
$insts : (L_s \cup L_t) \to 2^{(I_s \cup I_t)}$	Function assigning the set of instances to a lemma
$p : L_s \times L_t \times F \to [0; 1]$	translation pair probability
$f^i : L_s \times L_t \times F \to \mathbb{B}$	instance-level filter
$f^l : L_t \times F \times (L_s \times L_t \times F \to [0; 1]) \to \mathbb{B}$	lemma-level filter

Table 4.2.: Notation overview for Chapter 4

oriented strategy will retain a usable recall level, even without further generalisation.

4.3.1. The Generation Step

The goal of the generation step is to produce a list of FEE candidates for the target language. In the strategy we outlined above, this actually involves two different levels, namely *lemmas* and *instances*. This can be illustrated on Figure 4.1: Both predicate classifications (for the source and target languages) are located on the *lemma* level in that they map

semantic classes to lemmas. The annotation projection, on the other hand, is driven by word alignment links on the level of individual instances. On the target side, then, the instances have to be mapped onto lemmas again.[4]

As a consequence, the central unit of observation in the parallel corpus will is located at the instance level, namely the *translation pair*. A translation pair is a triple of a source instance, a target instance, and a frame, where the source instance is an instance of a word which can evoke the frame. For example, Figure 4.1 contains three translation pairs, all for the frame STATEMENT, the first one being (*says*,*sagte*,STATEMENT). We define translation pairs formally as follows (the notation is summarised in Table 4.2):

$$\mathrm{tp}(i_s, i_t, f) \equiv ((i_s \rightsquigarrow_1 i_t) \wedge \mathrm{fee}(f, \mathrm{lemma}(i_s))) \tag{4.2}$$

On the level of lemmas, on the other hand, what we are interested in is the joint probability that some target *lemma* l_t is a translation of a source lemma l_s for a given frame f. We also call l_t the candidate, and l_s the support. For example, in Figure 4.1 the lemma *say* is a support for the candidate *meinen* for the frame STATEMENT. We obtain this *translation pair probability* by simple maximum likelihood estimation (see e.g., Manning and Schütze (1999)) as the relative frequency of this specific translation pairs compared to all observed translation pairs:

$$p(l_s, l_t, f) = \frac{|\{\mathrm{tp}(i_s, i_t, f) \mid i_s \in \mathrm{insts}(l_s) \wedge i_t \in \mathrm{insts}(l_t)\}|}{|\{\mathrm{tp}(i_s, i_t, f)\}|} \tag{4.3}$$

The formulation of the translation pair probability as a joint probability makes it possible to describe many interesting properties of the translation process by marginalisation. For example, we can express the *candidate*

[4]This is not the only possibility to exploit word alignment information: we could also have used the weighted correspondences present in the model's translation table, i.e. on the lemma level. However, by taking local context into account, instance alignments are a richer source of information and do not involve setting thresholds to separate genuine alignments from spurious ones as would be necessary for using the translation table. In addition, the presence of both the instance and the lemma levels will allow us to formulate powerful filtering mechanisms in Section 4.4 (see there).

probability that a target lemma l_t is an FEE of frame f simply as follows:

$$p(f|l_t) = \frac{p(f, l_t)}{p(l_t)} = \frac{\sum_{l_s} p(l_s, l_t, f)}{\sum_{l_s} \sum_{f'} p(l_s, l_t, f')} \tag{4.4}$$

Since the example in Figure 4.1 is simplified in that it uses only one frame, the candidate probability of STATEMENT for all target-language candidates is one. Generally, we can specify our initial, unfiltered list of FEE candidates for a frame f as the set of all candidates with non-zero candidate probability for f (i.e., candidates for which we have observed at least one translation pair for f):

$$cand(t, l_t) \equiv (p(f|l_t) > 0) \tag{4.5}$$

In Figure 4.1, all three German candidates *sagen, meinen,* and *beispielsweise* are therefore candidates for STATEMENT.

4.3.2. The Filtering Step

This section provides a modular framework for the filtering procedures which are suited to prune the FEE candidate list obtained from the generation step. The description of the actual filters we have used is deferred to Section 4.4 and is preceded by a detailed review of the error sources we observed in the parallel corpus, and which motivate the filters.

In order to reap the benefits of the *data-driven* strategy discussed above, our filtering framework has to be as language-independent as possible; in particular, it has to be applicable to low-density languages for which language resources are scarce. This requirement precludes the use of "deep" linguistic analyses. Instead, we will base our filters shallow lexical information, namely part-of-speech tags and lemmatisation (recall also the discussion in Section 4.2). This knowledge-poor setting has clear implications on the architecture of our filtering procedures: In the absence of detailed linguistic cues, filtering has to proceed mainly on the basis of *patterns* in the corpus which are likely to correspond to *frame-preserving translations*. Our general approach can thus be characterised as *knowledge-lean*.

Naturally, both the instance level and the lemma level introduced above can yield such patterns, and we will therefore introduce two kinds of

filters, one for each level of information. Instance-level filters perform *local* filtering, accepting or rejecting translation pairs based on properties of the aligned tokens, or the frame involved. Lemma-level filters exploit *distributional* properties of candidates. We define both kinds of filters as additional conjunctive constraints on definitions from the generation step, which makes it possible to combine filters in a modular manner.

An instance-level filter is a binary predicate which takes a translation pair as input (a support, a candidate, and a frame). It evaluates to true if the translation pair is likely to be frame-preserving, according to its filtering criterion. This filtering step is integrated with the creation of translation pairs: We extend Equation (4.2) with a set of instance-level filters F^i, marked in boldface in the equation below. Instance-level filtering thus results in a smaller set of translation pairs, namely those for which all filters evaluate to true.

$$tp(i_s, i_t, f, F^i) \equiv i_s \rightsquigarrow_1 i_t \wedge fee(f, lemma(i_s)) \wedge \bigwedge_{\mathbf{f^i} \in \mathbf{F^i}} \mathbf{f^i}(\mathbf{s_i}, \mathbf{t_j}, \mathbf{f}) \quad (4.6)$$

On the lemma level we extend the definition of the binary candidate predicate $cand$ (Equation (4.5)). The definition now involves a set of lemma-level filters F^l, again marked in boldface. The input to the lemma-level filters is a candidate, a frame, and the translation pair probability distribution, which can be seen as a function assigning probabilities to translation pairs. As argued above, this function contains all relevant information about distributional characteristics of candidates and supports in the corpus. Lemma-level filtering thus reduces the set of final FEE candidates to those whose distribution in the corpus meets the requirements imposed by the lemma-level filters.

$$cand(l_t, f, F^l) \equiv (p(f|l_t) > 0) \wedge \bigwedge_{\mathbf{f^l} \in \mathbf{F^l}} \mathbf{f^l}(\mathbf{l_t}, \mathbf{f}, \mathbf{p}) \quad (4.7)$$

4.4. Error Sources and Filtering Procedures

In this section, we first discuss the different error sources resulting from the generation step which can be found in the FEE candidate list. This allows us to develop filtering procedures which specifically address these

error sources by identifying the statistical regularities by which errors from these sources can be recognised.

4.4.1. Sources of Errors

The architecture for frame projection we have developed in this chapter contains three main sources of erroneous FEE candidates: polysemy in the source data, noise in the word alignment, and frame instance non-parallelism in the translation.[5]

Polysemy of frame-evoking elements. The first source of errors is the polysemy of source language lemmas with respect to frames, which has been briefly mentioned in Section 4.3. Recall the case of *ask*, whose three frames are listed in Table 4.1 (page 75). Since Equation (4.2) does not include a disambiguation step for individual instances, all instances of *ask* together with their translations form translation pairs for *each* of these frames. Clearly, this leads to erroneous FEE candidates, since different senses of source language FEEs are likely to have different translations. For example, the German verb *bitten* can only be used in the REQUEST sense of *ask*; however, in the unfiltered list produced in the generation step, it will be listed as FEE candidate for QUESTIONING and COGITATION as well.

Polysemy is especially problematic for high-frequency predicates, which tend to be more polysemous than low-frequency ones (Miller et al., 1990) and often participate in idiosyncratic or only partly compositional constructions such as support constructions or metaphors (Burchardt et al., 2006b). The erroneous candidate *beispielsweise* from Fig. 4.1 is actually such a case, stemming from an infrequent corpus occurrence of *say* as apposition:

(4.8) Taking account of, **say**, economic, geographical and social criteria will lead to a blurring of these clear criteria.

Die Berücksichtigung von **beispielsweise** wirtschaftlichen,
The consideration of **for example** economic,

[5] We identified these error sources on the basis of the English-German and English-French EUROPARL bitexts. On the qualitative level, there was little difference between the two language pairs. Detailed quantitative comparisons follow in Chapter 5.

geographischen und sozialen Kriterien wird zu einer Verwischung
geographic, and social criteria will to a blurring
dieser nachvollziehbaren Kriterien führen.
of these comprehensible criteria lead.

Errors in statistical word alignment. Spurious and missing links in the automatically induced word alignment are another source of errors. This category is somewhat difficult to characterise, since a large number of linguistic phenomena can result in errors in word alignments. We will concentrate on the two main systematic shortcomings of the statistical word alignment models which we assume in this thesis (cf. Section 2.2). Recall that these models establish word alignments primarily by observing the co-occurrence frequency of all cross-lingual word pairs in a large corpus. They then establish links between the most frequently co-occurring words, taking a limited context into account.

The first problem of this strategy arises from frequent monolingual co-occurrences, so-called *indirect correspondences* (Melamed, 1996), which often lead to misalignments, especially for infrequent words. Consider the following example: the English verb *increase* often cross-lingually co-occurs with the German noun *Anstieg (increase)*. However, German *Anstieg* forms a collocation with *plötzlicher (sudden)*; therefore, *plötzlicher* also tends to co-occur with English *increase*. As a result, the alignment algorithm cannot decide well whether to align *increase* to *plötzlicher* or *Anstieg*. Similar patterns arise from collocations between verbs and prepositions, e.g., the frequent German phrase *darum bitten, dass (to ask that)*.

The second shortcoming is that statistical word alignment models are best suited to account for alignments between individual words, i.e. one-to-one alignments. State-of-the-art models have the expressive power to model one-to-many relations as well; but one-to-many alignments are only chosen if the corpus evidence is strong enough to overcome the bias towards one-to-one alignments, and, at least according to our experience, the resulting one-to-many alignments are generally of mixed quality. On the other hand, the translations of a large number of linguistic phenomena requires one-to-many or even many-to-many alignments to be modelled correctly. This comprises cases like noun compounds in German (e.g., *Reisekosten*) which correspond to two English words (*travel*

expenses), but also the complete area of multi-word expressions. To name just two subclasses, support verb constructions such as *eine Entscheidung treffen (to make a decision)* can be translated either as single verbs (*to decide*), or again as light verb expressions (*to make a decision*). Similarly, idioms usually require many-to-many alignments: *kick the bucket* can only be translated as a whole into *den Löffel abgeben (to hand over the spoon)*; no translational correspondences can be established on the level of individual words. On the technical level, the computation of one-to-many word alignments is made more difficult by the fact that the individual parts of multi-word expressions usually have very high monolingual co-occurrence frequencies, which gives rise to the indirect correspondence problem introduced above. Thus, the resulting word alignment is often inconsistent across instances: for example, English *decide* can be found to be aligned to either *Entscheidung* and *treffen* individually, or even to both.

Frame instance non-parallelism. The third error source are cases of frame-instance non-parallelism. These are cases of translation pairs where the support is used in the right sense, evoking the frame in question, and where the word alignment link is correct; however, the FEE candidate in the target language cannot evoke the frame. We have already discussed such a case (Example (4.1)) in Section 4.1 (Page 75)). Such instances arise through *translational shifts* on the semantic level, i.e., translations which do not preserve the frame-semantic class of the source predicate. Such "non-literal" translations will be discussed in detail in Section 5.7. The most relevant points for the present discussion are that a clean semantic lexicon for the target language needs to filter out instances of semantic translational shifts as far as possible. This is a difficult task, though, since these cases do not form a uniform pattern, but range from the systematic to the highly idiosyncratic and productive.

4.4.2. Filtering Mechanisms

In Section 4.3, we have stated the general framework for the filtering step: In order to be as language-independent as possible, filters should be informed primarily by shallow linguistic and distributional cues. In this section, we exploit the analysis of error sources presented in the last

section to define filters within this framework. Not surprisingly, some error sources can be addressed best by filtering on the instance level, while others are amenable to filtering on the lemma level.

Instance-level filters

We start with instance-level filters, which rely on properties of individual translation pairs.[6]

Instance-level filter 1: Restriction to content words (POS). This filter simply constrains the grammatical category of FEE candidates by discarding translation pairs whose target token is not a content word (a verb, noun, or adjective):

$$f^{i}_{POS}(i_s, i_t, f) \equiv cont(lemma(i_t)) \tag{4.9}$$

In this manner, the filter addresses target language collocations of a content word and a function word, an important subset of cases of indirect correspondence. Examples are noun-preposition (*fragen nach (ask for)*) and noun-adverb (*plötzlicher Anstieg (sudden increase)*) collocations. Since the English FrameNet lists almost exclusively content words as frame-evoking elements, we assume that this coverage is sufficient at least for related target languages. Infrequently, English FEEs are translated felicitously as adverbials or prepositions. However, it is currently an open research question whether prepositions or adverbials can be said to have an argument structure (similar to most content words), and whether they can therefore be analysed in frame-semantic terms (see Saint-Dizier (2005) for an affirmative account).

Instance-level filter 2: Bidirectional alignment (BD). The second filter targets the level of word alignment links themselves. As we have discussed in Section 2.2, state-of-the-art word alignment models are directed, i.e., asymmetric. It is therefore standard practice in statistical machine translation to obtain a bidirectional alignment by taking the intersection of the target–source and source–target alignments. Such a bidirectional

[6] Each filter is given a two- or three-letter shorthand which will be used in Chapter 5 to refer to its definition.

alignment contains only reliable one-to-one correspondences that occur in both word alignment. Since the the standard definition of translation pairs (Equation (4.2)) already requires the two words to be aligned in the source–target direction, we can enforce bidirectional alignment simply by requiring the link to hold in the inverse alignment as well:

$$f^{i}_{BD}(i_s, i_t, f) \equiv i_t \rightsquigarrow_1 i_s \tag{4.10}$$

Bidirectional alignment can be expected to address the problem of multi-word expressions, at least to some extent. The fact that multi-word expressions can generally only be aligned in part, often results in an empty bidirectional alignment for the participating words. This effectively removes incomplete multi-word expressions from the set of FEE candidates.

Instance-level filter 3: Frame disambiguation (FrD). The third filter addresses the problem of frame polysemy by performing frame disambiguation for each instance:

$$f^{i}_{FrD}(i_s, i_t, f) \equiv wsd(f, i_s) \tag{4.11}$$

This reduces the set of translation pairs for each frame to those instances which actually evoke the frame, according to the disambiguation model.

Lemma-level filters

Now we turn to lemma-level filters, whose input is the translation pair probability distribution p over candidate lemmas and frames. Recall from above that other probability distributions can be computed from the translation pair probability by marginalisation (e.g. the candidate probability, see Equation (4.4)).

Lemma-level filter 1: Predominant frame (PrF). This filter is a heuristic approximation of Instance filter 3 (FrD): It assigns each instance of a candidate the same frame, namely the most frequent one. On the technical level, for each candidate the filter evaluates to true only for the frame with the highest candidate probability, thereby removing the candidate for all other frames.

$$f^{l}_{PrF}(l_t, f, p) \equiv (f = \operatorname*{argmax}_{f'} p(f'|l_t)) \tag{4.12}$$

This strategy is known as the *predominant sense* heuristic (McCarthy et al., 2004). Its results are clearly an approximation, since target language words can also be polysemous, and should therefore be listed for more than one frame. However, work on word sense disambiguation has shown that this heuristic is surprisingly hard, and sometimes impossible, to beat by more informed methods. Therefore, it is worthwhile to compare the results of using the heuristic to solving the complete task.

Lemma-level filter 2: Translational consistency (Ent). Melamed (1997) has proposed a measure of *translational (in-)consistency* based on information theory. Translational consistency is defined for every target word as the entropy of the conditional translation probability distribution $p(s|t)$ over all aligned source words, given the target word.[7] The entropy of a probability distribution ranges between zero and infinity, with zero signifying that the complete probability mass is concentrated on a single event, and higher values that the mass is distributed across more events. The use of entropy captures the intuition that consistent translations will lead to a concentration of the probability mass on a few source words, resulting in low entropy values. Reliable translations can thus be enforced by keeping only FEE candidates whose entropy is below some threshold n:

$$f^{l}_{Ent}(l_t, f, p) \equiv (-\sum_{l_s} [p(l_s|l_t) \log p(l_s|l_t)] < n) \qquad (4.13)$$

Translational consistency is a general-purpose filter which attempts to remove chance errors by detecting rare patterns, which are assumed to be unreliable. This filter should have an impact on word alignment errors, and particularly multi-word errors, by systematically dispreferring target predicates with inconsistent alignment links. More importantly, though, it should have a bearing on frame instance non-parallelism, which is the most difficult error source to detect, and which has not been addressed by the other filters so far. Our hypothesis here is that translational divergence, a result of contextual factors forcing the translator to choose a different surface realisation, will lead to increased variance in translation, which can be detected by the consistency filter.

[7] The conditional probability distribution can also be computed from the basic translation pair probability, analogously to the candidate probability.

Lemma-level filter 3: Frame-based translational consistency (FrEnt). One shortcoming of Melamed's definition in our setting is that it does not take frame information into account by measuring consistency on the word level, and not on the frame level. As an example, consider the translation of German *versprechen (promise)*. If it is translated as *vow* or *promise*, both of which evoke the COMMITMENT frame, we still want to assign *versprechen* a high consistency (low entropy) for the COMMITMENT frame. If, on the other hand, *versprechen* is translated as *vow*, *say*, or *answer*, this indicates that *versprechen* does not always evoke a COMMITMENT situation.

To better capture this intuition, we propose an alternative, frame-based definition of translational consistency, which does not measure the *lexical* variation of a target lemma (i.e., how many source words it corresponds to), but the *semantic class* variation (i.e., how many frames it corresponds to). To do so, we collapse all known FEEs of the current frame into one "meta-word" (e.g., for COMMITMENT, all FEEs in Table 1.1 on Page 10 are treated as a single word). The results in the intended behaviour, namely that within-frame variance does not increase the entropy, but that across-frame variance does. In Equation (4.14), the term in the first line is the entropy of the event that the candidate t corresponds to one of the source language FEE for the frame f; the term in the second line accounts for all other translations, treated as single events. As before, we retain only FEE candidates if the entropy lies below a threshold n:

$$\begin{aligned} f^{l}_{\mathrm{FrEnt}}(l_t, f, p) \equiv (\; - \; & [\sum_{\mathrm{fee}(f,l_s)} p(l_s|l_t) \log \sum_{\mathrm{fee}(f,l_s)} p(l_s|l_t)] \\ & - \sum_{\neg \mathrm{fee}(f,l_s)} [p(l_s|l_t) \log p(l_s|l_t)] < n) \end{aligned} \tag{4.14}$$

The impact of frame-based entropy filtering on error types should be similar to basic entropy filtering, but more effective due to the improved ranking criterion.

4.5. Summary

In this chapter, we have presented the problem of inducing frame-semantic predicate classifications for new languages (Sections 4.1 and 4.2). When

comparable resources are available for other languages, the induction can be phrased as a cross-lingual transfer task, for which two main strategies are available: resource-based transfer (using e.g. bilingual dictionaries) and data-driven transfer (using bilingual corpora). We have adopted the data-driven approach, whose main advantages are lower prerequisites for new language pairs (only bilingual corpora are necessary), the presence of variation in translation which leads to a high-recall classification for the target language, and the availability of rich quantitative information as a basis for judging the confidence of translation pairs.

The second part of the chapter (Sections 4.3 and 4.4) has introduced and discussed our data-driven framework for the cross-lingual transfer of frame-semantic predicate classifications. The framework consists of two steps: in the first step, generation, automatically constructed word alignments are interpreted as indicators of frame-semantic parallelism to produce a comprehensive, but noisy, set of candidates for frame-evoking elements in the new language. The second step, filtering, removes spurious candidates by general, largely language-independent filtering mechanisms that rely mostly on shallow distributional information. Concrete filters were developed by analysing the major error types in the data.

5. Experimental Evaluation

In this chapter, we provide an experimental evaluation of the framework developed in Chapter 4 by inducing frame-semantic predicate classes for two different target languages, French and German, on the basis of the EUROPARL corpus. We establish that our filtering framework can be used to construct such classes with a high degree of precision.

We begin by detailing our general experimental setup (Section 5.1), followed by the actual evaluation experiments for the language pairs English–German (Section 5.2) and English–French (Section 5.3). After presenting a short comparative study on resource-based projection (Section 5.4) and an overview of related work (Section 5.5), we give a general discussion of our experimental results (Section 5.6). The chapter concludes with an analysis of the most important remaining conceptual problem, namely translational shifts that break frame instance parallelism.

5.1. Experimental Setup

The logic of our experiments followed the two steps of the cross-lingual induction architecture developed in Chapter 4. We performed two experiments whose only difference is the language pair: Experiment 1 considered the language pair English–German, and Experiment 2 the language pair English–French. Otherwise, both experiments followed the same design:

- We first ran the generation step to acquire an initial, unfiltered list of FEE candidates for each target language. Since the generation step did not involve any parameters, we did not consider it as part of the experiment itself; it can rather be seen as a preprocessing step. From the resulting unfiltered list, we drew a sample of frames for evaluation. The FEE candidates for the sample frames were annotated manually as correct or erroneous, and formed our gold standard.

- The subsequent filtering step formed our experiment proper. It consisted of applying different filter combinations to the unfiltered candidate list and evaluating the filters' output for the sample frames against the gold standard.

5.1.1. The Generation Step

As datasets, we used the English-French and English-German EUROPARL bitexts whose preprocessing was described in Chapter 2. Since the present task does not require "deep" processing, we use the larger Bitext 1 for this experiment. Bitext 1 contains word alignments, part-of-speech and lemma information, and consists of roughly one million bi-sentences for each language pair (see Section 2.3 for details).[1]

The generation step then consisted in constructing lists of *FEE candidates* for German and French as follows: We extracted all translation pairs which involved a frame-evoking element described in the English FrameNet resource (release 1.1) from the parallel corpus. Their one-to-one aligned translations in the target language formed our initial candidate list. We did not use any filtering at this stage apart from applying a simple frequency threshold: We disregarded all translation pairs for which we observed less than five instances in the corpus. This is standard practice in empirical modelling to rule out sparse, and therefore unreliable, observations.

5.1.2. Construction of Sample Gold Standard Classifications

Drawing a sample. The unfiltered lists of FEE candidates that resulted from the generation step contained over 44,000 candidates for each language, clearly too much for a complete manual annotation. Therefore, we drew a representative sample from this population as follows: To investigate how our approach performed across a range of frames with varying

[1]The information included in Bitext 1 is sufficient to apply all filters except Token filter 3, the Frame disambiguation filter. This filter requires frame disambiguation for each English predicate. We employed Erk's (2005) frame disambiguation system for this task. It is, to our knowledge, the only available standalone frame disambiguation system, and is described in Section 1.2.

DE Band	TP	FNr	AvgC
High	> 7836	159	199.7
Medium	< 7836	159	64.7
Low	< 959	158	12.6

FR Band	TP	FNr	AvgC
High	> 9528	163	203.0
Medium	< 9528	163	56.6
Low	< 1277	162	11.3

Table 5.1.: Frame frequency bands (TP: translation pair instances; FNr: number of frames; AvgC: average number of candidate FEEs per frame)

frequencies, we split the set of frames into three equal-sized bands, based on the number of translation pair instances per frame in the corpus. For each of the resulting bands for German (DE) and French (FR), Table 5.1 shows the number of translation pairs (TP), the number of frames in the band (FNr), and the average number of FEE candidates per frame (AvgC) in each band. We randomly selected five frames from each band while ensuring that frames fell into the same bands in French and German. This made our samples from the two language pairs as comparable as possible, and allowed us to draw contrastive conclusions. The selected frames are shown in Table 5.2, together with the total number of initial (unfiltered) FEE candidates for each band.

A number of additional observations can be made in Table 5.1. First, the EN–FR data covers 12 frames more than the EN–DE data. Closer analysis revealed that these are all "marginal" frames with just one English frame-evoking element, which did not happen to be involved in any translation pair in the EN–DE corpus. Note also that the EN–FR dataset generally contains a higher number of translation pairs than the EN–DE dataset, while the average number of candidates is about the same. The main reason for this pattern is the slightly larger size of the English-French bitext. The second important factor is word order, which differs more between English and German than between English and French, especially with regard to the position of the predicate: Declarative sentences, which make up the vast majority, are (almost exclusively) verb-second in English and French, while German makes a difference between main clauses, which are verb-second as well, and embedded clauses, which are verb-final. Current statistical word alignment models, which reward similar

Band	Frames	# Cands (DE)	# Cands (FR)
High	Cause_change_of_scalar_position, Communication_response, Deciding, Giving, Preventing	828	799
Medium	Employing, Evaluative_comparison, Judgment_communication, Sensation, Travel	366	347
Low	Adding_up, Congregating, Escaping, Recovery, Suspiciousness	84	68
Sum		1278	1214

Table 5.2.: Frames from different frequency bands selected for evaluation, with total number of unfiltered candidates for German and French

positions of translation candidates within their respective sentences, tend to run into problems aligning predicates at differing positions. This results in a sparser word alignment for the EN–DE dataset compared with the EN–FR dataset, which notably leaves more English predicates, potential FEEs, unaligned.

Annotation scheme. All FEE candidates from the sample (a total of 1278 candidates for German and 1214 for French) were annotated for correctness by two annotators. Annotators tagged one FEE candidate at a time, but were able to see and revisit earlier annotation decisions. They based their decisions on the candidate itself, and on its list of English supports (cf. Section 4.3). We also provided bilingual concordances for candidate-support translation pairs from EUROPARL to furnish some typical context.

The annotation scheme was defined by detailed English annotation guidelines. These operationalised the main annotation criteria, gave

illustrative examples, and discussed borderline cases. They are provided in Appendix A.

In addition to the fundamental classification of the candidates as either correct or spurious, we asked annotators to subdivide spurious cases into three errors classes: *polysemy errors (P)*, *multi-word errors (M)*, and *remaining errors (noise) (N)*. Polysemy errors are cases where the candidate is a valid translation of the support, but for a different sense (i.e., frame). Multi-word errors are cases where either candidate or support form part of a multi-word expression which breaks the translational equivalence between the individual words. The Noise category covers all remaining errors.

The aim of the error classification was to obtain a more detailed picture of the success of our different filtering procedures in removing the errors they were designed for. Note, however, the error annotation is necessarily based on distinctions which could be made easily and reliably by the annotators. It therefore differs somewhat from the error classification given in Section 4.4.1, which categorises errors by error *sources*. The error class P corresponds directly to the polysemy errors in Section 4.4.1; error class M is a subset of the errors in statistical word alignment introduced there; errors from class N, however, are either alignment errors or cases of frame instance non-parallelism. The distinction between these two classes is a very difficult one to draw, touching on the general problem of the appropriateness of claiming that two words are translationally equivalent, and can often only be made in particular contexts.

Annotation process and reliability. The FEE candidates for each language pair were annotated independently by two graduate students. Both annotators were native in one language and had very good competence in the other. They were also familiar with FrameNet. Annotation speed was between 100 and 200 instances per hour, using a simple text editor-based setup. After annotation was finished, we computed inter-annotator agreement by a pairwise comparison of the annotated categories. We computed both raw agreement, and κ, a chance-corrected measure of agreement (Carletta, 1996). We found an agreement of 85% (κ =0.79) for English–German and 84% (κ =0.78) for English–French. The remaining disagreements were discussed, and a gold standard was created by consensus.

These agreement numbers demonstrate good reliability, reflecting that the task is well-defined and the output can be given a meaningful interpretation. More generally, this suggests that a generate-and-filter approach can also be reasonably integrated into a more elaborate workflow, e.g., by adding a subsequent manual verification step.

5.1.3. Experimental Method

Evaluation strategy and the precision-recall tradeoff. As indicated above, the experiment itself consists in filtering the initial set of FEE candidates. By applying different filter combinations, we explore the *precision-recall tradeoff*: With more filtering, the resulting resulting list of FEE candidates is bound to be cleaner, but also smaller. This is not necessarily a problem, since our aim is to induce a high-precision seed lexicon (cf. Section 4.3). However, the seed lexicon must retain a substantial size to be amenable to monolingual extension methods; we will discuss this point below.

On the technical level, combining filters is trivial, since all filters are defined as conjunctive constraints either on translation pairs (Eq. (4.6)) or on candidates (Eq. (4.7)). Therefore, the result of any filter combination will be a subset of the initial candidate list. This means that any filtered lists can be evaluated against the manually annotated gold standard.

Recall from Chapter 4 that all filters are parameter-free binary filters, with the exception of the entropy-based translational consistency filters. These impose a linear order on FEE candidates and thus use a threshold that determines which candidates to accept. To minimise the number of times the threshold parameter has to be set, we first consider all binary filters to identify the optimal combination of binary filters in terms of F-Score. We then apply the consistency filter in the following form: we rank the candidates according to their consistency and keep only the top n candidates. These n-*best lists* can be evaluated as before.

Type- vs. Token-based evaluation. There are two possible counting regimes for the evaluation: by type or by token. Evaluation by type is based on the number of correct candidates, and is therefore mainly a quality measure in terms of the vocabulary, while evaluation by token

measures the ability to cover words in running text. While type-based evaluation is arguably a more useful statistic for predicate classifications, token-based evaluation can complement the picture by including the effect of word frequency. Therefore, we provide the basic evaluation (binary filters) additionally on the token level. However, if no level is explicitly mentioned, evaluation figures always refer to the type level.

Evaluation measures. We use the standard quality measures from Information Retrieval, namely precision, recall, and F-Score (see for example Baeza-Yates and Ribeiro-Neto (1999)). We assess whether differences in F-Score are statistically significant using stratified shuffling (Noreen, 1989), an instance of assumption-free approximative randomisation testing (see Yeh (2000) for a discussion). Note that we define recall with respect to the unfiltered list: a recall of 100% corresponds to the retrieval of all true positives in the gold standard.[2]

As stated above, the translational consistency filter uses a threshold parameter which allows different settings. We also provide an overall assessment of the filter's quality by computing *mean average precision (map)*. This statistic is applicable to any retrieval model that ranks its returned candidates so that only the top N candidates can be considered. For each possible value of N, it computes the precision of the top-n list, and returns the average:

$$\mathrm{map} = \frac{\sum_{r=1}^{n} \mathrm{Precision}(r)}{n}$$

By considering top-n lists for increasing n, mean average precision rewards assigning good candidates high ranks.

In addition to these quality measures, we break down the false positive FEEs in terms of polysemy (P), multiword (M), and noise (N) errors. For each filter combination, the error percentages plus the precision (i.e., the ratio of true positives) sum to 100%, i.e. all candidates of that condition.

[2] This manner of computing recall provides an accurate comparison of the different filters within the generate-and-filter paradigm. A more comprehensive evaluation of recall against the complete target language would be more informative, but would require the exhaustive annotation of a sample of free text from the target language. Such an evaluation was outside the scope of the present study.

5.2. Experiment 1: Language pair English–German

The quality of the induced lexicon for German prior to and following binary filtering is summarised in Table 5.3 in terms of types, and Table 5.4 in terms of tokens.

The unfiltered list (NoF) has a recall of 1.0 (by definition), and a precision of 0.35 (0.69 on the token level, which corresponds well to the token-based frame parallelism numbers we identified in Section 3.3). These already substantial precision numbers indicate that the general approach of inducing frame-semantic predicate classifications from parallel corpora is in fact promising: The performance of the NoF baseline is 52% F-Score (82% F-Score for tokens). However, filtering is still clearly necessary: In the unfiltered list, two thirds of the unfiltered FEE candidate types, or one third of the unfiltered FEE tokens, are still false positives and cannot evoke the frame they are listed under.

5.2.1. Binary Filters

We first assessed the impact of the individual filters on the list of FEE candidates. The part-of-speech filter (POS) affects recall only marginally, but filters out a number of false positives. This leads to an improvement of 5% F-Score (3.6% F-Score for tokens). This filter was designed to remove spurious links resulting from indirect alignments, and it fulfills its task: The percentage of N-type errors is reduced from 24% to 14% (8% to 3% for tokens). The slight increase in P-type and M-type errors for tokens does not signify the introduction additional errors. It reflects the stronger decrease of true positives compared to the remaining errors.

The bidirectional filter (BD) leads to a stronger decrease in recall, but also a corresponding increase in precision, with a net gain of 4% F-Score (2% on the token level). Our design hypothesis about the BD filter was that it should retain only reliable one-to-one word alignment correspondences. This is confirmed in the data: Both M-type errors, which result from one-to-many correspondences and N-type errors, are reduced – in the case of types, to less than half their original number. A comparison of the type-based and token-based evaluations reveals that the filter is

Model	Recall	Precision	F-Score	%P	%M	%N
NoF (baseline)	100	35.2	52.0	30	9	24
POS	98.0	40.2	57.0	33	10	14
BD	70.2	47.1	56.4	36	4	11
FrD	79.7	34.8	48.4	31	10	21
PrF	49.7	56.7	53.0	21	16	3
POS FrD	77.7	39.5	52.4	34	11	13
BD POS	**68.4**	**50.0**	**57.8**	**37**	**4**	**7**
BD POS FrD	54.0	49.5	51.7	37	4	7
BD POS PrF	36.0	68.3	47.1	24	4	2

Table 5.3.: English–German: Type-based evaluation of binary filters and relative frequency of error classes (100% = all candidate types)

Model	Recall	Precision	F-Score	%P	%M	%N
NoF (baseline)	100	69.0	81.6	19	4	8
POS	99.0	74.8	85.2	18	4	3
BD	97.8	72.7	83.4	19	3	5
FrD	96.5	70.4	81.4	19	4	7
PrF	67.5	84.9	75.2	11	3	1
POS FrD	95.5	75.9	84.6	17	4	3
BD POS	**96.8**	**77.7**	**86.2**	**18**	**2**	**2**
BD POS FrD	90.3	78.0	83.7	17	3	2
BD POS PrF	65.5	89.3	75.6	8	2	1

Table 5.4.: English–German: Token-based evaluation of binary filters and relative frequency of error types (100% = all candidate tokens)

even somewhat overzealous: The recall loss is much more pronounced in the type-based evaluation (30%, vs. 2% for tokens), which indicates the existence of a group of low-frequency true positives that are erroneously removed after being aligned differently in the two directions. Problems with low-frequency items are a typical shortcoming of statistical word alignment models.

The next filter, frame disambiguation (FrD), was created to address the polysemy problem (P-type errors). Unfortunately, its results are disappointing: the precision can hardly be improved, and for types, recall drops significantly by 20%. The result is an even slightly lower F-Score. The hardly changed relative frequencies of the errors types show that FrD fails to address the task at hand, namely deciding between frame-preserving and frame-changing FEE candidates.

The predominant frame filter (PrF) should address the same problem, and does so to a certain extent, reducing the number of both polysemy and noise errors. The precision is thus at its highest level for all individual filters at 57% (85% for tokens), however at the expense of a sharp drop in recall. For types, the overall result is still positive – the F-Score increases by 1% – however, for tokens, the F-Score drops by 6%.

The lower portions of the tables show the most informative combinations of binary filters. The first row (POS FrD) shows that combining FrD with other filters does not yield better results: The loss in recall of POS FrD, compared with POS alone, leads to a worse overall result. The second row shows the best filter combination both for type-based and token-based evaluation, namely BD POS. Evidently, BD and POS apply somewhat orthogonal filtering criteria to the N-type errors and are able to reduce their relative frequency from 24% to 7% (from 8% to 2% for tokens). Together with BD's effective filtering of M-type errors, this results in a candidate list with a relatively small amount of alignment-related (M- and N-type) errors; however, polysemy as an error source remains an essentially unsolved problem. The difference between the overall F-Scores of the baseline (NoF) and BD POS is significant ($p < 0.001$). BD POS also significantly outperforms the best single filter, POS ($p < 0.01$).

The bottom two rows specifically address the polysemy problem by combining BD POS with frame-based filters. BD POS FrD again obtains bad results, probably due to the use of FrD. BD POS PrF shows the best precision we have obtained with binary filters, namely 68% (89% for

tokens). This is mainly the result of filtering out a substantial number of P-type errors; some N-type errors were removed as well. However, at the same time BD POS PrF shows the lowest recall we obtained (36% for types, 66% for tokens), which leads to a net decrease in F-Score.

In sum, we found that the relative performance of all binary filter combinations was almost completely identical between type-based and token-based evaluation, with the absolute performance around 25% to 30% higher in the token-based scheme. Not surprisingly, this indicates a general frequency effect, as generally observed in empirical language modelling: More frequently observed events (in this case, more frequent FEE candidates) tend to be more reliable.

5.2.2. Consistency Filters

We consider four filter combinations for consistency filtering. Recall that the consistency filters provide a *ranking* of a given candidate list. Therefore, we combined each of our two two consistency filters, the vanilla entropy filter (Ent) and the frame entropy filter (FrEnt), with two candidate lists, the unfiltered list (NoF) and the best binary filtered list (BD POS).

As described in Section 5.1.3, we extract the n best percent of candidates for different settings of n, starting with 100% and decrementing in 3-percent steps. The resulting datapoints are shown in Figure 5.1 in the form of a precision-recall plot. Each condition is visualised by thirty-three datapoints, with low n (strict filtering) to the left, and high n (lenient filtering) to the right. Note that for n=100%, no consistency filtering takes place, and the result corresponds exactly to the previously computed results for NoF (recall=100%, precision=35%) and BD POS (recall=68%, precision=50%).

For orientation, we have added in grey the recall level at which the resulting resource contains on average as many true frame-evoking elements for the new language as FrameNet does for English, namely an average of 10 FEEs per frame. For German, this recall level is 30%, since the initial, unfiltered list of 1278 candidates (cf. Table 5.2) has a precision of 35%, which corresponds to an average of 29 true FEEs per frame.

Two main trends emerge: First, consistency filtering can hardly improve the precision of the unfiltered candidate list (starting point: recall 100%, precision 35%), while the list pre-filtered with BD POS (starting

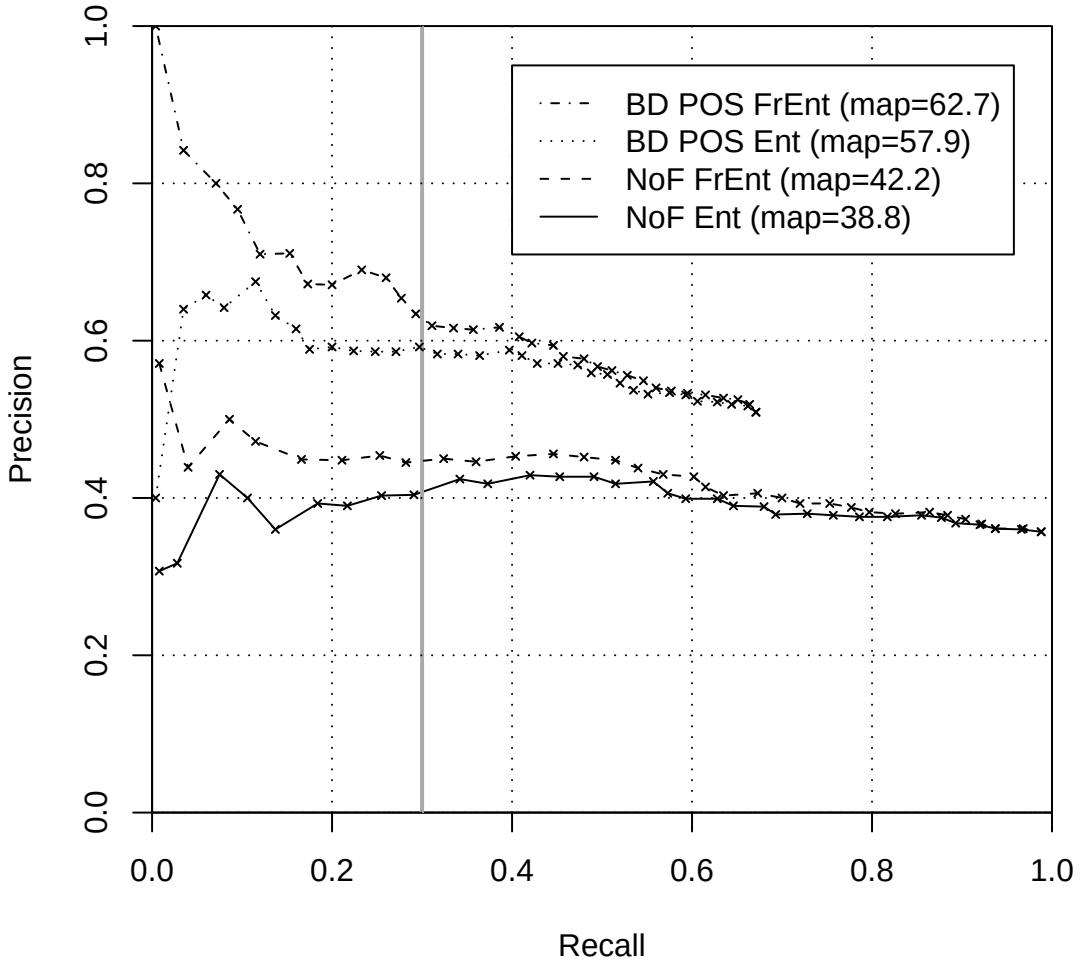


Figure 5.1.: English–German: Precision–Recall tradeoff and mean average precision (map) for consistency filters (type-based evaluation). The grey horizontal line corresponds to the size of FrameNet (Recall level $\approx$ 30%).

point: recall 68%, precision 50%) can be ranked more successfully. Second, in ranking the BD POS-filtered list, frame entropy systematically outperforms vanilla entropy. While frame entropy consistently balances lower recall with higher precision, while vanilla entropy cannot exceed a precision level of approximately 65%. These observations are confirmed by the mean average precision scores shown in the legend of Figure 5.1. They show that the presence (or absence) of binary filtering is the more

Band	High	Medium	Low
NoF	51.2	32.0	6.8
BD POS	36.6	20.0	5.0
BD POS PrF	17.2	11.2	4.0
BD POS FrEnt (recall $\approx$ 0.3)	19.6	8.6	2.0

Table 5.5.: English–German: Average number of true positives per frequency band for filter combinations (type-based evaluation)

important factor, but that the type of entropy filter is also important.

In sum, a combination of binary filtering (BD POS) with frame-based entropy filtering (FrEnt) makes it possible to obtain very high precision lists for German, albeit at low recall levels.

5.2.3. Frequency Bands

Table 5.5 shows how the induced lexicon varies in size (average number of true positives per frame) across frequency bands (High, Medium, Low) before filtering, and with different filter combinations. Unsurprisingly, in the unfiltered list (NoF) more true FEE candidates are found for high-frequency frames than for low-frequency frames; since the frequency bands were selected based on the number of translation pairs, the frames in this band tend to have many, and frequent, English FEEs. Simple filtering with BD POS expectedly reduces the number of true positives in each band due to the decrease of recall. Overall, the impact of BD POS is rather similar across frequency bands.

As described above, additional filtering to improve precision cuts down drastically on recall. The PrF filter removes more high-frequency items than low-frequency items: The true positives in the High band are reduced by a factor of 2.1 and the low band by 1.25. For the consistency filter, we optimised the threshold to obtain a recall of about 30%, which corresponds to a resource about the size of English FrameNet. It shows exactly the opposite behaviour, reducing low-frequency candidates much more (by a factor of 2.5) than high-frequency candidates (by a factor of 1.9).

5.3. Experiment 2: Language pair English–French

The results of binary filtering for French are summed up in Table 5.6 (type-based evaluation) and Table 5.7 (token-based evaluation). As for German, the recall of the unfiltered list (NoF) is 100%. The precision is in the same region, but (at least on the type level) slightly lower for French (30%) than for German (35%). Correspondingly, the F-Score is also about 5% lower at 46%. While polysemy errors appear to be of about the same importance, we detect significantly more N-type and slightly less M-type errors in French than in German.

5.3.1. Binary Filters

As for German, we first apply the filters individually. The relative performance of the different filtering combinations is very similar to what we find for German, albeit at a lower level of precision.

The POS filter improves precision by 3% to 32.6% (72% for tokens). As expected, this is mainly a result of the elimination of N-type errors (32% to 26%). Analogously, the BD filter successfully reduces the number of N-type errors (from 32% to 22%) by removing asymmetric translation pairs. This improves the F-Score by 4% for types, though only by 1% for tokens. As for German, high-frequency candidates appear to be less susceptible to alignment-related errors. Interestingly, BD has only a very small impact on the number of M-type errors. Also in parallel with the results for German, the application of FrD results mainly in a drop in recall for types; the precision decreases (on the type level) or increases (on the token level) slightly.

Compared to German, the three first filters have less impact on the candidate list: the recall remains higher, and the precision lower. The inverse is true for the PrF filter: it manages to double precision from 30% to 60% (token level: increase from 70% to 87%), but at the expense of almost 70% recall (token level: 40%). As a result, the overall F-Score is lower than for the baseline, NoF. This filter reduces error types across the board, while polysemy still remains the major source of errors, reduced only from 31% to 26% (tokens: 20% to 9%).

Model	Recall	Precision	F-Score	%P	%M	%N
NoF (baseline)	100	29.8	45.9	31	6	32
POS	98.6	32.6	49.0	34	6	26
BD	80.3	35.9	49.7	37	6	22
FrD	79.2	29.4	42.9	32	7	31
PrF	31.4	60.0	41.3	26	3	11
POS FrD	77.9	32.1	45.5	35	7	26
BD POS	**79.8**	**37.9**	**51.4**	**37**	**6**	**19**
BD POS FrD	64.3	37.4	47.3	38	6	19
BD POS PrF	28.1	65.3	39.3	24	1	10

Table 5.6.: English–French: Type-based evaluation of binary filters and relative frequency of error types (100% = all candidate types)

Model	Recall	Precision	F-Score	%P	%M	%N
NoF (baseline)	100	70.2	82.5	20	3	7
POS	99.9	72.0	83.7	20	3	5
BD	99.4	71.9	83.5	20	3	5
FrD	98.6	72.0	83.3	19	3	6
PrF	60.1	87.3	71.2	9	0	4
POS FrD	98.5	73.5	84.2	19	3	5
BD POS	**99.4**	**73.3**	**84.4**	**20**	**2**	**4**
BD POS FrD	97.9	74.7	84.7	19	2	4
BD POS PrF	60.0	88.3	71.5	8	0	4

Table 5.7.: English–French: Token-based evaluation of binary filters and relative frequency of error types (100% = all candidate tokens)

Regarding filter combinations, in the lower portions of the tables we observe that combinations with the FrD filter are still always worse than without it, at least on the type level. The best filter combination remains BD POS with 38% precision and 80% recall at the type level. The differences between BD POS and both the baseline list (NoF) and the best single filter (POS) are highly significant ($p < 0.001$).[3] As for German, BD POS manages to remove more N-type errors than either filter alone; however, more than 60% of the false positives remain, most of which are polysemy errors. Further combining BD POS with the PrF filter results in the best precision we have obtained for the French data, namely 65% (88% for tokens). However, due to the very low recall of 28%, the F-Score falls below baseline.

Overall, the type-based and token-based evaluations yield parallel results for French as well. The difference in overall performance is even greater for French than for German (30-35% F-Score), which indicates that low-frequency candidates are more difficult to treat for French than for German. The only major difference between types and tokens is observed in the case of FrD, which leads to actual slight improvements in the token-based evaluation, indicating that it performs above chance for very high-frequency candidates.

5.3.2. Consistency Filters

Figure 5.2 shows the impact of the consistency filters on French data. We use the same two-factor setup as for German, varying both the dataset (NoF vs. BD POS) and the consistency filter (Ent vs. FrEnt). Again, we consider top-n% lists, starting with 100%. In the case of French, considering all candidates from the NoF list yields a recall of 100% and precision of 30%; all candidates from the BD POS exhibit a recall of 80% and a precision of 38%. Again, we indicate the recall level of FrameNet (10 true FEEs per frame); in the case of French, this is 40%, since the unfiltered candidate list contains on average 24 true positives per frame.

Overall, we find that the precision increases when a consistency filter is applied, i.e., when only the n% highest-ranked candidates are considered.

[3] In the token-based evaluation, BD POS FrD is actually numerically better than BD POS by 0.3% F-Score. This difference is not statistically significant.

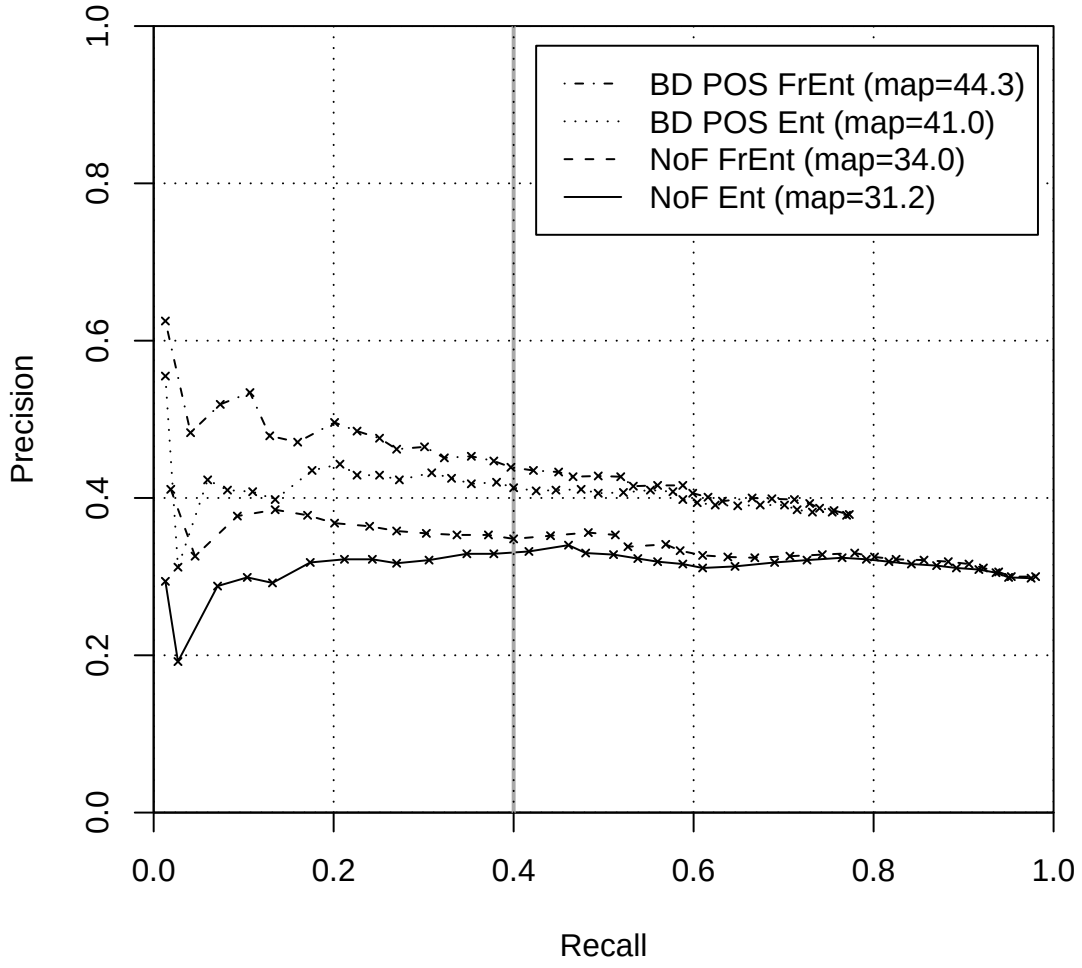


Figure 5.2.: English–French: Precision/recall tradeoff and mean average precision (map) for consistency filters (type-based evaluation). The grey horizontal line indicates the size of FrameNet (Recall level $\approx$ 40%).

Qualitatively, the two main results for German carry over to French: frame entropy (FrEnt) yields better performance than vanilla entropy (Ent), and filtered data (BD POS) fares better than unfiltered data (NoF); the second factor is more important than the first. This is confirmed by the mean average precision values, also shown in Figure 5.2.

However, the precision margin that can be gained from consistency

Band	High	Medium	Low
NoF	37.0	29.6	5.8
BD POS	31.0	23.2	3.6
BD POS PrF	6.8	11.8	1.8
BD POS FrEnt (recall≈ 0.4)	13.8	8.8	0.8

Table 5.8.: English–French: Average numbers of true positives per frequency band for filter combinations (type-based evaluation)

filtering is disappointingly small for French. For the NoF data, we can at most increase the precision by 10% to 40% (at 10% recall). For the binary filtered data (BD POS), we can improve the precision by 17% to 55% (also at 10% recall). We will discuss this problem below in Section 5.6.

5.3.3. Frequency Bands

Table 5.8 shows the impact of filtering on different frequency bands for French. The result for NoF corresponds well to German. Note, however, that the number of true positives in the High band is much higher for German than for French; this indicates that in EUROPARL, the German translations of English high-frequency items show more lexical variation than the French translations. Medium and Low bands show comparable numbers, though.

The side-effect of BD POS to remove true positives (i.e., to decrease recall) is less strong for French than for German, especially in the High and Medium bands, which explains the overall lower German recall (68% vs. 80%).

Contrastively, the decrease in recall for the predominant frame filter and the consistency filter are both much more pronounced for French than for German. Still, the effect on different frequency bands carries over: A comparison of BD POS with BD POS Prf shows that PrF cuts down mostly on high-frequency items, leading to a reduction by a factor of 4.6 for the High band and 2 for the Low band. FrEnt, on the other hand, is more severe on the low band (4.5) than on the high band (2.2).

5.4. Dictionary-based Predicate Class Induction

Recall from Section 4.2 that two basic strategies are available for the cross-lingual transfer of predicate classes: data-driven transfer, and resource-based transfer. While we have advocated a data-driven approach, in this section we repeated our experiment using dictionaries instead of parallel corpora to gauge the potential of the alternative approach. As dictionary resources, we used the machine-readable versions of two professional, comprehensive dictionaries, the Oxford–Duden English–German and Oxford–Hachette English–French bilingual dictionaries[4].

We paired each English lemma in the dictionary with all of its translations. In the process, we removed nonliteral translations, translations in context (i.e., senses given in the form of complete phrases) and translations for specific collocations, as far as they were marked consistently. For each frame, we then computed the set of FEE candidates as the union of all translations for the English FEEs of this frame. We are aware that this procedure is bound to return translations of wrong senses; however, there is no a priori way of disambiguating dictionary entries. We annotated the same sample used in Experiments 1 and 2 for correctness, also using the same annotation scheme (see Section 5.1.2).

The results of the evaluation are shown in Table 5.9. The precision of the FEE candidate sets is comparable to the precision of the unfiltered FEE candidate sets we extracted from parallel corpora in Experiments 1 and 2 (35% for German and 30% for French). There are two main differences to the results of the earlier experiments. The first is the size of the FEE candidate set: while we were able to obtain between 1200 and 1300 candidates for each language from the corpus(cf. Table 5.2), the dictionaries is considerably smaller, containing only two thirds the number of candidates for German, and half the number for French. The second difference is the distribution over error types, which is as would be expected from a clean dictionary from human usage: almost all errors in the dictionary-induced candidate set are polysemy errors, that is, translations of inappropriate senses. There are virtually no noise errors, and just a small number of multi-word errors, which results from inconsistently marked multi-word

[4]We are grateful to Oxford University Press for allowing us to use these resources.

Language	Precision	%P	%M	%N	#Cands
German	34	59	6	1	901
French	38	59	3	0	623

Table 5.9.: Evaluation of dictionary-based induction of semantic predicate classes.

expressions. The incidence of multi-word translations is somewhat more pronounced in the English–German dictionary and also appears to be the main reason for the difference in precision between the language pairs.

In sum, the first impression of this strategy is favourable: the dictionary-induced candidate lists are roughly comparable to the unfiltered candidate lists obtained from parallel corpora. Of course, given the large number of P-type errors, the observation we have made in Section 4.3 applies here as well: filtering is crucial to obtain usable predicate classes with a high precision. This is, however, where we encounter the fundamental problem of dictionary-based transfer: being targeted at a human audience, the data encoded in a dictionary is much more difficult to use as basis for automatic filtering procedures. A first aspect of this is that the linguistic descriptions in entries are not normalised (as for collocates) or refers to the vague notions of traditional grammar (as for valency). Another second aspect is that no quantitative information is present in the entries, which makes it difficult to judge the confidence in particular translation pairs. In combination with the smaller initial size of the candidate sets, and the absence of comparable dictionaries for many language pairs, we conclude that a dictionary-based strategy for cross-lingual transfer that relies exclusively on bilingual dictionaries is both less generally applicable and presents more obstacles to the development of suitable filtering schemes than a data-driven strategy.

5.5. Related Work

The cross-lingual transfer of frame-semantic predicate classifications has been the topic of two other studies. The first one, Fung and Chen (2004), follows an approach different from the one taken in this thesis, namely

cross-lingual transfer on the basis of a bilingual ontology, and has been already discussed in Section 4.2. The second study, Kanamaru, Murata, Kuroda, and Isahara (2005), is more similar to ours, exploiting translation pairs in a parallel corpus to identify Japanese lexical units for FrameNet frames. Its motivation, however, was substantially different: Instead of automatically deriving a comprehensive semantic predicate classification, the researchers were interested in a lexicographic resource comparison across languages. The study concentrated on the FrameNet analyses for one particular, highly polysemous Japanese predicate, *osou*, which were compared to classifications drawn from monolingual Japanese dictionary resources. The study avoided the need for automatic filtering methods by manually validating all translation pairs.

The task of inducing frame-semantic predicate classes completely "from scratch" has been addressed by Green et al. (2004) in an English monolingual setting. Using the information from two English dictionary resources (WordNet and LDOCE), they first clustered verbs into frame-semantic classes, and then determined names and sets of semantic roles for these frames. Their results corresponded well to FrameNet frames, where available, and were generally found to be adequate by human judges (Green and Dorr, 2004). However, the application of this approach to induce classifications for new languages would meet two difficulties. Firstly, it relies on high-quality dictionary resources, which may not be available, especially for low-density languages. The second, and more general, problem indicated by the study is the huge parameter space imminent in the task: Recreating exactly the particular meaning distinctions made by FrameNet is a very difficult problem. Green et al.'s solution for English builds on specific properties of the resources involved, and cannot be expected to generalise easily to other resources or languages.

A more modest aim has been to extend the FrameNet resource, typically also with the help of other resources. For example, Burchardt, Erk, and Frank (2005a) identify probable frames for unknown English predicates by identifying the most similar FrameNet entry according to WordNet distance. Giuglea and Moschitti (2006) combine FrameNet with VerbNet by defining a mapping between the semantic roles of the resources. This results in an increased coverage both of previously unknown verbs, and of linking patterns for known verbs. The success of these methods indicates the feasibility of extending a small predicate classifications, such as a seed

lexicon produced by our methods.

Another similar task is the induction of clean, broad-coverage translation lexicons. This task is almost universally approached in a data-driven fashion, with the help of large parallel corpora in which distributional properties of translation pairs can be observed. However, in contrast to our generate-and-filter strategy, it is usually assumed that it suffices to define an appropriate *association score* between source and target lemmas and to apply a simple n-best cutoff. A number of different association scores have been proposed (see e.g., Melamed (1996) or Tufiş (2002) and the references therein). The crucial difference between the two tasks is the linguistic relation that is being modelled, namely translational equivalence and frame-semantic equivalence, respectively. These two relations are not identical: While frame-semantic equivalence is largely lexically determined, translational equivalence is to a certain degree dependent on the context. As a result, translation pairs which are equivalent in certain contexts such as *answer (a question) / addresser (une question)* are acceptable entries for a translational lexicon (see also the discussion in Melamed (1996)). On the other hand, the two words evoke different frames (COMMUNICATION_RESPONSE and TOPIC, respectively), and thus cannot be used for transferring frame information. In comparison, we have developed a more flexible filtering paradigm, whose usefulness has been verified by the experimental results – recall that purely association-based filtering, at least with our Ent and FrEnt filters, is outperformed by a combination of association-based filtering with other mechanisms (see Figures 5.1 and 5.2).

5.6. General Discussion

In Part II of this thesis, we have discussed the task of inducing a frame-semantic predicate classification for new languages, arguing that the cross-lingual transfer of an existing classification in a data-driven fashion holds the greatest promise. We have presented a general framework that decomposes the cross-lingual transfer into a generation and a filtering step: In the generation step, word alignment links from a large parallel corpus serve as a proxy for frame-semantic equivalence to produce a list of *FEE candidates* in the new language. In the filtering step, intuitions about

properties of frame-preserving translations can be exploited to remove erroneous candidates.

We have evaluated our approach by inducing frame-semantic classifications for German and French, using the English FrameNet resource and EUROPARL bitexts. We find that for both languages, only about one third of the unfiltered FEE candidate lists resulting from the generation step consist of true positives, which confirms general expectations about the necessity of filtering in annotation projection. At the same time, due to the variety of translations, these unfiltered lists contain about three times as many true positives for the new languages as FrameNet does for English. This high recall constitutes a promising basis for filtering, since it allows us to apply restrictive filters to the list and still retain sufficient candidates to form a high-quality seed lexicon. In fact, we find that the application of a small number of shallow filters, which are predominantly based on distributional, and shallow linguistic, properties of translations, results in a significant increase of the precision of the induced classifications.

Finally, we have repeated our experiments, using bilingual dictionaries in place of bilingual corpora. The results are comparable in precision to the unfiltered corpus-induced candidate lists, but somewhat smaller in size. In addition, the absence of quantitative information in the dictionary makes it difficult to devise automatic filtering procedures for the prevalent polysemy errors. We therefore consider our data-driven strategy as more powerful as well as more easily applicable.

5.6.1. Measuring the Quality of a Predicate Classification

We have compared the performance of our binary filters and filter combinations mainly on the basis of F-Scores. This is usual practice, since F-Scores combine precision- and recall-based evaluation aspects in a single overall figure of quality. We have been able to show that almost all binary filters lead to significant increases in F-Score; this means that their gain in precision outweighs their loss in recall.

However, an F-Score based evaluation does not do justice to at least one binary filter, PrF, which produces the highest precision values, but whose F-Score is near or even below baseline due to the recall penalty it incurs.

A similar point can be made for the consistency filters, which give rise to a continuum of predicate-recall combinations whose usefulness is not easily compared on the level of F-Scores.

An alternative evaluation strategy could start out from the observation that the recall used in F-Score is computed with respect to the unfiltered candidate list (cf. Section 5.1.3), which is somewhat arbitrary; in practice, the absolute size of the resulting predicate classification will be more relevant. As a result, it might be instructive to compare precision values at a fixed level of recall. For comparatively low recall values, this corresponds to the "seed lexicon" idea we have referred to throughout this chapter. A natural choice for this recall level might be the size of the original FrameNet resource for English. The complete English FrameNet resource (release 1.2) lists on average 10 FEEs per frame; our unfiltered candidate lists contain on average 30 (German) and 24 (French) true positives. Thus, a lexicon whose size roughly corresponds to FrameNet must have a recall level of 30% for German and 40% for French. These levels are shown as grey lines in Figures 5.1 and 5.2.

For German, the best consistency filter combination at this recall level (BD POS FrEnt) yields around 63% precision; interestingly, the binary filter combination BD POS PrF yields a higher precision, 68%, at a higher recall of 36%. For a resource of approximately this size, therefore, BD POS PrF appears to be the better choice. For French, the same consistency filter obtains (BD POS FrEnt) 44% precision. The precision of BD POS PrF is considerably higher at 65%, but at a much lower recall (28%). Whether the high precision of BD POS PrF can be exploited therefore depends on external considerations, namely whether a lower bound for the size of the induced resource exists.

An additional consideration, which is especially important for comparatively small induced classifications (i.e., at low recall levels), is the effect of filtering procedures across the different frequency bands. We found that PrF and the consistency filters differ substantially in their behaviour. As can be seen in Tables 5.5 and 5.8, PrF filters out candidates across the board, while the consistency filter tends to remove low-frequency items. Since frames with a single correct FEE are probably to be avoided, PrF is arguably more appropriate.

5.6.2. Comparison between French and German

The majority of our observations generalises well across the two target languages, in particular the relative performance of different filtering mechanisms. This indicates that the results we obtained are not idiosyncratic for particular language pairs. The most important difference is that all filtering procedures (with the exception of PrF, the predominant frame filter) were less aggressive for French, i.e., less candidates were removed. This resulted in lower recall losses, but simultaneously lower precision gains than for German.

A possible explanation for this observation is suggested by results from contrastive linguistics, where English was found to prefer more abstract verbs, while German tends to use specific or concrete verbs (Hawkins, 1986); French appears to side with English in this respect. Indeed, an inspection of our bitexts revealed roughly 3400 English verb types, 3100 French verb types, and 4200 German verb types. This indicates that the distinction between meanings is organised differently: in German, it is to a high degree a matter of lexicalisation, whereas in French other linguistic means are used. This observation is supported by the lower numbers both of unfiltered candidates and true positives we find for French, compared to German. Since our filtering architecture is primarily based on distributional properties of translation *pairs*, and largely ignores monolingual context, it is not surprising that it is better able to distinguish between German candidates and pick out the most consistent ones, which results in the low combination of low recall and high precision we observe. French candidates, on the other hand, are more difficult to distinguish in our framework, resulting in a smaller impact of our filters.

Judging from these observations, we would expect our framework to work well for target languages that distinguish meanings by lexicalisation and less well for languages that use other means. Performance for such languages can presumably be improved by extending our framework with a component that models monolingual context as a richer basis to determine (frame)-semantic equivalence. Supporting evidence for the role of context as a source of information can be drawn from studies which induce semantic predicate classifications exclusively from unannotated monolingual corpora (Schulte im Walde, 2006).

5.6.3. Error Types

Recall that our annotators classified false positives into three classes: polysemy-related errors (P), multiword-related errors (M), and all other errors (N), in particular cases of alignment noise and frame non-parallelism. We found that P and N were the dominating error types in the unfiltered data for both languages, both of which accounted for around 30% of candidates. M-type errors occurred only in smaller quantities.

The improvement we obtain from filtering is almost exclusively a result of a significant reduction in N-type errors, both for French and German. The most successful filters are BD (bidirectional alignment) and POS (part-of-speech filtering), which are able to remove largely complementary sets of N-type errors, so that the combination BD POS consistently obtains the highest F-Score. While this strategy is able to remove the majority of N-type errors for German (from 24% down to 7%), more than half of the original number remains for French (19%, compared to an original 32%). These are almost exclusively cases of frame instance non-parallelism, which will be considered in detail in Section 5.7.

Disappointingly, we find that a considerable ratio of over 20% of polysemy (P-type) errors remains in the filtered lists for both languages. The frame disambiguation filter (FrD) has hardly any impact. The predominant frame (PrF) filter addresses the polysemy problem to some extent, since it can combine evidence from translation pairs for different supports, and thus increase precision significantly (from 35% to 57% for German, and from 30% to 60% for French); however, this comes at the cost of a large loss in recall. This error class therefore merits some closer inspection.

In turns out that polysemy errors are not artefacts of the projection process. They are already present in the source language, independent of any cross-lingual considerations. Their main cause is the current incomplete coverage of FrameNet in terms of senses for predicates: FrameNet, being primarily a lexicographic project, extends its semantic lexicon one frame at a time, without necessarily covering all senses of a given lemma. This happens frequently, as a comparison of FrameNet with the SALSA lexicon, a FrameNet-compliant semantic lexicon for German which lists all senses for predicates, shows: The average polysemy is 2.3 in FrameNet and 4.1 in the SALSA lexicon (Erk, 2005). As an example, consider the

	WordNet gloss	FrameNet frame	German translation
1	declare to be true	STATEMENT	zugeben
2	allow to enter	–	einlassen
3	allow participation	–	zulassen
4	admit into a group	–	einlassen, zulassen

Table 5.10.: Incomplete coverage of FrameNet: the case of *admit*

FEE *admit.v*. Table 5.10 shows the four WordNet senses of *admit* which are documented in the SemCor corpus. Only the most frequent sense, sense 1, is covered by FrameNet. In the absence of evidence to the contrary in the training data, the supervised frame disambiguation system underlying the frame disambiguation (FrD) filter therefore wrongly classifies all tokens of *admit.v* as cases of STATEMENT. This error percolates through projection and results in the listing of German *einlassen* and *zulassen* as STATEMENT verbs. These are clear polysemy errors, since *einlassen* and *zulassen* cannot be used to express the STATEMENT sense of *admit* at all. Since the translations of "wrong" senses are often quite reliable, they are difficult to remove with the predominant frame filter.

A promising direction for future work is provided by recent advances in *outlier detection*, also known as *single-class classification* (Markou and Singh, 2003). This is a general strategy to identify outliers in datasets; in our application, these outliers correspond to instances of predicate senses not yet covered in FrameNet. Erk (2006) has presented a first encouraging application of outlier detection to FrameNet example sentences, where she identified outliers by their distance to nearest neighbours. An application of this technique to our corpus which removes all instances of senses not covered by FrameNet can be hoped to result in a substantial decrease of polysemy errors.

5.7. A Closer Look at Translational Shifts

In our evaluation, we have found two major error sources which were difficult to address by our filtering procedures, namely P-type (polysemy) errors (for both languages) and N-type (noise) errors (for French). In the last section, have sketched a proposal for addressing polysemy errors; this section considers N-type errors in more detail. Even though we were able to remove N-type errors almost completely for German in our sample, this sample was admittedly small, and these errors remain a significant factor for French. It is conceivable that these errors become more prominent for languages which are typologically further removed from English.

Recall from Section 5.1.3 that the N-type error category covers all erroneous candidates which do not directly evoke the frame for which they are listed in the induced lexicon, but which are caused neither by polysemy nor multiword-related problems. This leaves two main processes as sources for N-type errors, namely errors in the word alignment, and translational shifts proper. In an inspection of relevant FEE candidates, we found that alignment errors do not figure prominently: in almost all cases, the alignment links the source predicate to what can be considered its best translational equivalent in the target sentence. This ties in well with the results for German, where the N-type errors could be eliminated almost completely. Our conclusion is that the remaining "difficult" cases of N-type errors are cases of proper *frame instance non-parallelism*, where the translation crosses frame boundaries. In other words, these are the instances where the fundamental assumption of our projection approach, namely that word alignment can be interpreted as frame-semantic equivalence (cf. Section 4.2) fails.

The purpose of this section is to develop a better qualitative understanding of the complexity of the problem, and to assess to what extent it can be solved by automatic means. To this end, we investigate the linguistic phenomena that give rise to the N-type errors found in the induced French frame-semantic lexicon, giving consideration to the properties of FrameNet that are relevant to the errors. We also link these results to more general insights from translation science on translational shifts, that is, cases where a translation differs in its linguistic properties from the original.

5.7.1. Translational Shifts and Frame Instance Non-parallelism

In translation science, a number of classifications has been proposed for translational shifts (see Cyrus (2006) for a discussion). For our purposes, we adopt Cyrus' classification, a variant of van Leuven-Zwart (1989), since it is explicitly aimed at investigating the relationship between predicates (and arguments) and their translations, the same level of description that frame-semantic analysis addresses. Cyrus' classification distinguishes two main classes of translational shifts: *grammatical shifts* and *semantic shifts*, both of which are subdivided into finer categories which are based on the relation holding between original and translation. For predicates, the most relevant grammatical shifts are *(de-)passivisation* and *category change*; the most relevant semantic shift is *semantic modification*, broadly defined as "some type of semantic divergence", and its subcategories *explicitation/generalisation*.

Applied to our data, we find that the complete group of grammatical shifts is unproblematic with respect to frame instance parallelism, since frame-semantic analysis abstracts over grammatical features such as voice or part of speech. Thus, if only grammatical shifts occur, a translation evokes the same frame as its original, provided that the frame is applicable to the new language. In contrast, semantic shifts clearly have the potential to lead to frame instance non-parallelism. Note, though, that they do not necessarily do so, due to the fairly coarse granularity of frame-semantic classes. For example, the verbs *say* and *reiterate* are FEEs of the same frame, STATEMENT, although their relation is probably one of explicitation/generalisation.

Figure 5.3 illustrates the relationship between frame instance parallelism/non-parallelism and translational shifts, using a simple model of translation. The process is modelled as consisting of an interpretation step, which recovers an underlying state of affairs from a source language expression, and a generation step, which re-expresses this state of affairs in a new language. This leads to an upside-down version of the well-known Vauquois triangle (Vauquois, 1975), with the frame as an intermediate, partly language-independent layer. The figure on the left shows the unproblematic case: If the state of affairs is expressed very similarly in the source and target languages, the original predicate and its translation are

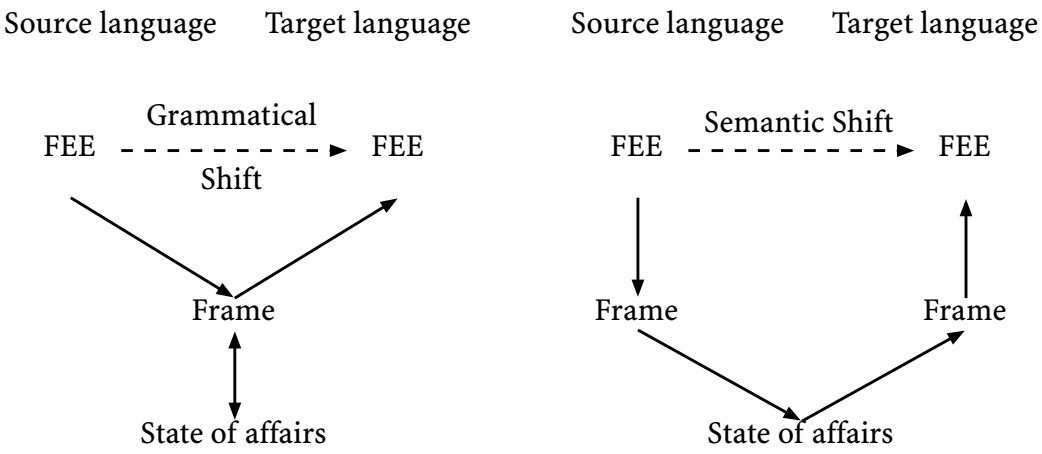


Figure 5.3.: Translation of predicates: The case of frame instance parallelism (left) and frame instance non-parallelism (right)

FEEs of the same frame, and no shifts, or only grammatical ones, occur on the surface level. In this case, frame instance parallelism is preserved, and our projection method will obtain correct FEE candidates for the target language. The right-hand side shows the problematic case that occurs when the translator decides to express the state of affairs in the target language using a frame different from the one used in the source language: On the surface level, a semantic shift is visible, and on the semantic level, frame instance parallelism breaks. In consequence, projection is prone to result in errors.[5]

5.7.2. Which Semantic Shifts Break Frame Parallelism?

The next step of our analysis is to investigate what the prominent causes are for those semantic shifts that can break frame instance parallelism. Cyrus' classification cannot provide an answer for two reasons. First, her scheme is framework-neutral and therefore cannot be expected to distinguish specifically between frame-breaking and frame-preserving semantic

[5] In the case of EUROPARL, where both languages of a bitext can be translations from an original third language, the two expressions can be thought of as created by parallel generation; this does not change the gist of our argument.

shifts. Second, Cyrus herself notes that "it is rather difficult to find objective criteria" for the subclassification of semantic shifts. This is due to the immense number of semantic, pragmatic and stylistic considerations which can lead to semantic shifts in translation (cf. Section 3.1.2).

To obtain a better understanding of the causes, we manually analysed the N-type errors which we found in the induced French frame-semantic lexicon. While our results cannot form a the basis of a comprehensive discussion of semantic shifts, we were able to identify two groups of phenomena which contributed very strongly towards cases of frame instance non-parallelism in our sample, namely the interdependence between arguments and predicates, and complex event structure.

Interdependence between arguments and predicates. In its current state, FrameNet makes the simplifying assumption that frames are introduced primarily by lexical items, i.e. the FEEs.[6] In general, no great influence is assigned to the specific semantics of the frame element fillers other than the general implications provided by the frame description. In our data, however, we find that languages differ in the degree to which the role fillers influence the choice of the frame and frame-evoking element. This is especially prominent for the CAUSE_CHANGE_OF_SCALAR_POSITION frame, whose FEEs are used in English to very generally express processes of change. In French, there appears to be a tendency to systematically use more specific frames, depending on the semantic type of the changing ITEM. For instance, in Example (5.1), the combination *increase [undesirable property]* is translated as the French predicate *aggravation*.

(5.1) An **increase** in social inequality

L' **aggravation** des inégalités sociales
The **aggravation** of the inequalities social

Arguably, *aggravation*, as a predicate specific to changes in quality and restricted to contexts with a strong evaluative component, should not be considered as French FEE for the frame CAUSE_CHANGE_OF_SCALAR_POSITION. Another instance is Example (5.2), where the combination *increase* + *[weight]* is translated with the more specific French *alourdir*

[6] It is acknowledged that frames can also be introduced by *constructions*, but this is currently not reflected in the resource.

(to make heavier), for which no corresponding English single-word expression exists, and which presumably evokes some more specific frame dealing with weight change.

(5.2) Extending the Community's legal competence within the framework of the third pillar has **increased** the burden.

Le fardeau s'est **alourdi** avec une extension de la
The burden has itself **made heavier** with an extension of the
compétence juridictionnelle communautaire dans le cadre
competence legal of the Community in the frame
du troisième pilier.
of the third pillar.

Complex event structure. The inventory provided by FrameNet to describe meaning is the set of frames, i.e. prototypical situations. Recall that FrameNet frames model a rather coarse level of semantic granularity by grouping predicates into frames, semantic classes defined by common linguistic realisation patterns and common characterisations of the participants and props of the event they describe. Frames do not claim to capture the complete meaning of the predicates they describe, but only its most salient meaning component. On the other hand, all but the most simple real-world states of affairs that are the subject of linguistic discourse combine more than one meaning aspect. As a result, there is almost always "more than one way of putting it", with different frames competing for the linguistic realisation of this state of affairs. In a translation setting, this can result in frame mismatches (see the right-hand side in Figure 5.3).

Clearly, this is not a problem which is specific to FrameNet; rather, it is one of the central problems of lexical semantics. The examples we found in our sample exhibit a continuum of the degree to which the relation between the candidate and its support can be characterised by a single lexical relation. This continuum consists of the complete range from (quasi)-synonymy, whose treatment appears possible, to pure association, whose treatment would involve a great deal of world knowledge. One of the simpler cases is Example (5.3), where the event is expressed in English as a process caused by an actor, using the transitive *raise*, while no actor is possible in French with the intransitive use of *monter*. This kind of shift corresponds to the causative/inchoative alternation (Levin, 1993),

and there exists a quasi-synonymy relation between the two predicates, even though they evoke different frames.

(5.3) [...] The employment rate within the EU can be **raised** to 70%.

Le niveau d' emploi pourrait **monter** à 70% dans l' UE.
The level of employment could **rise** to 70% in the EU

The next two cases, Examples (5.4) and (5.5), show the gradually increasing difficulty of characterising the lexical relationship which holds between candidate and support.

(5.4) Why, for example, was the proposal to **increase** Europe's active population to 75% of the total population removed?

Pourquoi a-t-on retiré par exemple la proposition
Why has one retracted for example the proposition
prévoyant que la population active devait **atteindre** 75% en
foreseeing that the population active had to **reach** 75% in
Europe?
Europe?

(5.5) [...] The legal issue should take second place to consumer protection and **preventing** the public from harm.

La question juridique doit venir après la protection des
The question legal must come after the protection of the
consommateurs et les **précautions** pour nos citoyens.
consumer and the **precautions** for our citizens.

In Example (5.4), English expresses the process itself (*increase*), while French expresses the resulting end state (*atteindre (reach)*). In this case, the process–result relation is still clear, whereas it becomes more elusive in Example (5.5): what exactly is the relation between *prevent* and *précaution*? A precaution that is being taken against some X does not imply that X is prevented, although this is typically the case. This appears to be an instance of a weaker process–result relation, namely process–(intended result) or process–(typical result). The last example, Example (5.6), is even further along this continuum:

(5.6) Questions that were not **answered** during Question Time shall be answered in writing, Mr Gahler.

Les questions qui ne sont pas **examinées** pendant l' heure des
The questions that not are not **examined** during the hour of
questions recevront une réponse écrite, Monsieur Gahler.
questions will receive an answer written, Monsieur Gahler.

Here the relation between the target language FEE candidate (*answer*) and its source language support (*examiner*) is one that is purely associative: answering a question typically involves examining it. Still, the fact that this kind of relation turns up as errors in our lexicon induction task shows that this relation holds systematically enough in the corpus to lead to the consistent translation pair *answer-examiner*.

5.7.3. Controlling Semantic Shifts

The final question relevant in the context of the projection of frame-semantic predicate classes is to what degree translational shifts such as the ones described above can be controlled.

The interdependence between arguments and predicates appears to be moderately amenable to automatic processing. One possibility is to capitalise on the fact that instances of this kind, by definition, involve arguments with specific semantic properties, such as the undesirability that warrants the translation of *increase* as *aggravation* in Example (5.1). Such properties can be captured by taking monolingual context into account in the projection model; note that this extension of our framework has been motivated independently in Section 5.6.2 to obtain a model of equivalence that is appropriate for a broader range of languages that our present framework. Another idea would be to extend our setup to the true multilingual case and consider translations in more than two languages. In this setting, only translational shifts which show the same systematicity in several target languages would remain in the induced set of candidates. Under the assumption that the interdependence between arguments and predicates is expressed differently across languages, this would lead to smaller, but highly precise lexicons. Importantly for practical application, the necessary multilingual data is readily available from EUROPARL. The problem of complex event structures is much more general and much

harder. Only the first part of the continuum appears to be amenable to automatic processing techniques at the present state of the art: For the limited case of alternations, which usually involve changes in syntactic valency, the shift can presumably be controlled by requiring the syntactic valency of the candidate to match the support's; however, this strategy already breaks down for translation pairs involving changes in the part of speech.

More generally, the automatic distinction between different kinds of lexical relations by distributional means is an actively researched issue in lexical semantic modelling. Unfortunately, current state-of-the-art methods are still largely unable to make the fine-grained semantic distinctions necessary to address translational shifts. As an indication, consider the VerbOcean resource (Chklovski and Pantel, 2004), which classifies predicate pairs into different lexical relations based on search engine results for specific query patterns. An analysis of VerbOcean shows that its *semantic similarity* relation cannot distinguish between (frame-)semantic equivalence proper and more associative relationships: it lists *increase* as related to (e.g.) *accelerate* as well as *widen* and *decline*.[7] Still, this approach appears to be more promising than its alternative, namely the use of conceptual knowledge, which re-introduces all the problems of resource-driven approaches discussed in Section 4.2.

5.8. Summary

This chapter has provided an experimental evaluation of the data-driven framework for the cross-lingual transfer of frame-semantic predicate classifications developed in Chapter 4. Using the English FrameNet database (release 1.1) as basis, we produced frame-semantic predicate classifications both for German and French, using the respective EUROPARL bitexts as parallel corpora.

The experimental results (Sections 5.2 and 5.3) have confirmed the intuitions we have used to develop our framework in Chapter 4 in at least two important aspects. First, the first step of the data-driven transfer results in comparatively large predicate classifications for the target languages,

[7] The confidence scores VerbOcean supplies cannot be used to make this distinction, either.

taking advantage of the lexical variation in translation. However, second, the induced classifications contain only about one third true positives, which highlights the need for powerful filtering. We have been able to show that our filtering schemes can indeed result in a significant improvement of the predicate classification, both in terms of F-Score, where recall was computed with respect to the unfiltered classification, and in terms of precision for a given recall level that corresponds to the size of the original English FrameNet resource. In the latter condition, the obtainable precision figures are 68% for German and 50% for French. In an analogous experiment with bilingual dictionaries (Section 5.4), we obtained comparable results to the unfiltered condition of the data-driven transfer; however, we found the information encoded in the dictionary not to be easily amenable to automatic filtering. Sections 5.5 and 5.6 have presented related work and discussed our results against this background.

The last part of the chapter, Section 5.7, has provided an extensive data analysis of the induced frame-semantic predicate classification. We have shown that the largest remaining problem is the lack of sense disambiguation for instances of English frame-evoking elements in the corpus and sketched a possible solution for this problem. We have attributed the inferior performance for French results to typological differences between English and French, which can also be addressed by a conservative extension of our framework. Lastly, we have analysed the group of "hard" error cases created by translational shifts on the semantic level, concluding that these are currently outside the scope of simple filtering methods.

Part III.

Cross-lingual Projection of Frame-Semantic Roles

6. A Framework for Cross-lingual Role Projection

In this chapter, we develop a framework for the cross-lingual projection of semantic role information. Section 6.1 motivates the task. In Section 6.2, we discuss the decomposition of annotation projection for semantic roles into two subtasks, namely alignment and transfer. Finally, Section 6.3 presents our general framework, which phrases alignment as an optimisation problem, and describes concrete instantiations.

6.1. Motivation

In Section 1.2, we have introduced the task of *shallow semantic parsing*, i.e., the assignment (a) of semantic classes to predicates, and (b) of semantic roles to surrounding constituents. We have argued that shallow semantic parsing is an important step towards a robust, but informative, representation of the meaning of the predicates. The resulting representation abstracts over different syntactic configurations and provides fine-grained information on the event type as well as on the relations that hold between the predicate and its arguments.

We have shown in Chapter 3 that role semantics in general, and Frame Semantics in particular, exhibits a substantial degree of cross-lingual parallelism; therefore, the frames and roles provided by FrameNet appear suitable for the analysis not only of English, but also of other languages. However, the development of shallow semantic parsers for new languages runs into a *resource scarcity problem*: As outlined in Section 1.2, all state-of-the-art models for shallow semantic parsing make use of supervised learning techniques, inducing models from annotated training data; however, sufficient training data for both frame assignment or role assignment are only available for English and a small number of other languages. In

Chapter 4, we have presented a method for constructing frame-semantic classifications for languages without such resources; the current chapter is concerned with inducing role-annotated corpora in a similar fashion.

The general framework we assume is annotation projection, as introduced in Section 1.3. Since role-semantic annotations are assigned to complete constituents (as opposed to single words), we concentrate on obtaining appropriate *semantic alignments* between source and target sentences. We formalise the search for the best semantic alignment as an optimisation problem in a bipartite graph. We argue that bipartite graphs offer a flexible and intuitive framework for modelling semantic alignments that is able to represent translational divergences arising from typologically distinct languages. We present different classes of models with varying assumptions regarding admissible alignments between the source and target language. In addition, we investigate the amount of linguistic knowledge required for effecting projection successfully. We thus compare and contrast word-based semantic alignment models against more syntactically-informed ones that employ constituent structure. Experimental results on the parallel corpus sample (cf. Section 3.3) demonstrate that semantic roles can be projected relatively accurately when syntactic knowledge is taken into account.

This chapter begins with a presentation of the structure of our modelling framework (Section 6.2), and continues with a detailed formalisation and discussion of different instantiations of the framework in Section 6.3. Chapter 7 provides an evaluation of the framework: After our experimental setup is discussed in Section 7.1, the experiments for the language pairs English–German and English–French are presented in Sections 7.2 and 7.2.2, respectively. After reporting on related work (Section 7.4), we conclude with a discussion of our results and future work (Section 7.5).

6.2. Decomposing Projection into Alignment and Transfer

We first give an informal description of the structure of the semantic role projection process using the parallel sentence in Figure 6.1 as an example. As usual, we assume an initial analysis of the source language text (Step 1).

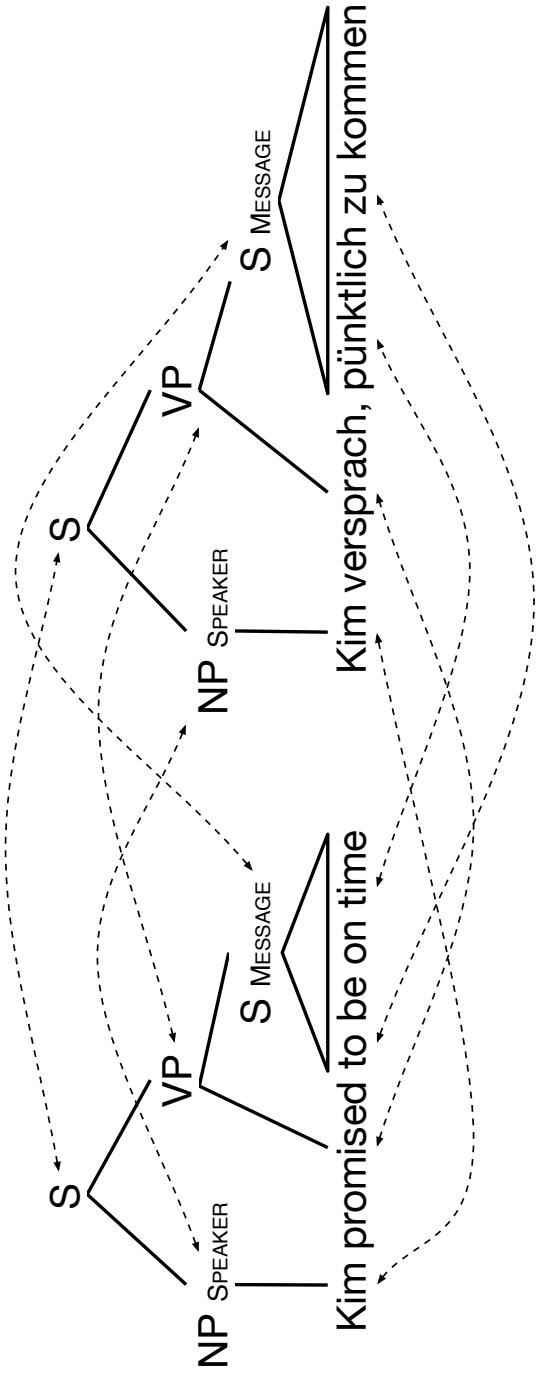


Figure 6.1.: Cross-lingual projection of semantic role information with optimal alignments.

However, we subdivide the subsequent cross-lingual annotation projection into two independent steps, namely *alignment* (Step 2) and *transfer* (Step 3):

1. **Semantic role annotation of the source (English) side.** Recall from Section 1.2 that this can be done automatically using a shallow semantic parser. The result are semantic role labels for English constituents. In Figure 6.1, the NP *Kim* is labelled with the role SPEAKER and the embedded sentence *to be on time* is labelled with the role MESSAGE.

2. **Induction of semantic alignments.** We next induce *semantic alignments*, indicated by dashed lines in Figure 6.1. Semantic alignments are generalisations of word alignments to arbitrary constituents. They connect linguistic units of the two languages which realise the same semantic content, independent of their surface form; phrased differently, they indicate *translational equivalence*. Semantic role information is completely ignored in this step. As an example of a semantic alignment, the English phrase *to be on time* is semantically aligned to German *pünklich zu kommen*.

3. **Transfer of role information.** Once the semantic alignments are in place, semantic role information is transferred along alignment links, thus creating role labels for the target constituents. In Figure 6.1, the semantic role SPEAKER will be transferred onto the German NP *Kim*. Similarly, the embedded sentence *pünklich zu kommen* will be assigned the role MESSAGE.

As far as we know, the decomposition of projection into alignment and transfer has not been proposed in the literature. We assume that this is the case because existing studies were mostly concerned with the projection of information for individual words (such as parts of speech (Hwa et al., 2005)) or short phrases (such as NP bracketing (Yarowsky and Ngai, 2001)), where correspondences could be established from word alignment alone. However, the situation is different when it comes to semantic roles, which can span arbitrarily long constituents, for example embedded sentences. As we will show later on, relying on word alignments for the projection of longer roles is bound to suffer from alignment errors.

Separating alignment and transfer is beneficial both with respect to methods and applications: First, on the level of methods, a formulation of the alignment task that is independent from the annotation projection scenario allows us to investigate solution strategies from different areas of research, in particular graph matching and phrase alignment. Second, with respect to applications, the modularisation enforces the development of general and reusable models. Even though we specifically address the task of projecting semantic roles, the models we develop do not use the semantic role annotation at all, and can thus serve for the projection of other types of linguistic annotations.

Note, though, that we specifically model *semantic* (as opposed to structural) alignments. Using semantic alignments for projection is appropriate for all types of annotation with a predominantly semantic definition, such as named entities, anaphoric links, and discourse relations. On the other hands, projection of, for example, grammatical function labels along semantic alignments cannot be guaranteed to succeed, since semantically equivalent units are not necessarily realised using the same grammatical function.

In what follows, we develop a model for the projection of semantic roles which capitalises on the idea of semantic alignments. We discuss the properties of the alignment process in detail, and show how models with varying degrees of linguistic knowledge can be derived. For ease of exposition, we assume that Step 1 (analysis of the source text) has already been performed. This issue will be discussed in detail in during evaluation (Section 7.2.2).

6.3. Framework Formalisation

We represent a bi-sentence (i.e., a pair of parallel source and target sentences) as two sets of linguistic units for the source ($u_s \in U_s$) and target ($u_t \in U_t$) languages. These units can be words, chunks, constituents, or other groupings. The semantic roles for the source sentence are modelled as a labelling function $a_s : R \rightarrow 2^{U_s}$ which maps roles to sets of source units. We view projection as the construction of a similar role labelling function for the target sentence, $a_t : R \rightarrow 2^{U_t}$. We thus restrict ourselves to projecting one frame at a time (cf. Section 3.2.2).

Following the decomposition introduced above, we first induce A, the set of semantic alignments between source U_s and target U_t units. We formally define A as a subset of the Cartesian product of the source and target units:

$$A \subseteq U_s \times U_t \tag{6.1}$$

The (informal) semantics of A is as follows: An alignment link between $u_s \in U_s$ and $u_t \in U_t$ implies that u_s and u_t are semantically equivalent. Provided with A and the source role assignment function a_s, the transfer step consists simply of assigning each source label r to the union of target units that are semantically aligned with the source units labelled with role r:

$$a_t(r) = \{u_t \,\|\, \exists\, u_s \in a_s(r) : (u_s, u_t) \in A\} \tag{6.2}$$

The decomposition of projection into alignment and transfer allows us to model the induction of semantic alignments explicitly as an *optimisation problem*. Given a cross-lingual similarity metric sim on links between source and target units, we search for the semantic alignment with a maximal product of link similarities. Ideally, the similarity function sim should measure the *semantic* similarity between source and target units. Several methods have been proposed in the literature for computing similarity within the same language (see Weeds (2003) and Budanitsky and Hirst (2001) for overviews) but not across languages. For the experiments reported in this thesis, we employ automatic word alignments as a proxy for semantic equivalence (cf. Section 2.2). Following general practice, we assume that sim is a function ranging from zero (minimal similarity) to one (maximal similarity).

In addition, we require that each semantic alignment A we consider is a member of a set of *admissible alignments* $\mathcal{A}$. The notion of admissible alignments is central in our approach; they impose constraints on the set of appropriate alignments, thus guiding the search towards linguistically meaningful correspondences (admissible alignments are discussed in detail in Section 6.3.2). Admissible alignments can be seen as counterparts of *priors* in statistical modelling, which are also used to induce models with certain characteristics. For example, we may want to avoid leaving linguistic units unaligned, or we may want to be able to model one-to-many alignments. Our optimisation problem to find the optimal

admissible alignment is thus:

$$\hat{A} = \underset{A \in \mathcal{A}}{\operatorname{argmax}} \prod_{(u_s, u_t) \in A} sim(u_s, u_t) \tag{6.3}$$

A wealth of algorithms can be used to solve this optimisation problems. In this thesis, we consider *bipartite graph optimisation* methods. Bipartite graphs provide a simple and intuitive framework for cross-lingual semantic alignment that allows to derive several model classes based on different assumptions with respect to the units of projection and the set of admissible alignments. Comparison of these models can shed new light on the properties of the projection task (see below). In addition, solution algorithms for graph optimisation are well-understood and computationally moderate.

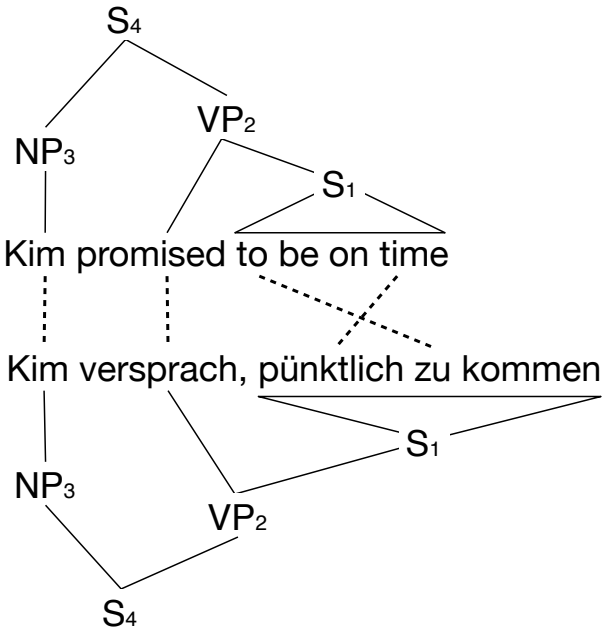
Formally, a weighted bipartite graph is a graph $G = (V, E)$ whose node set V is partitioned into two nonempty sets V_1 and V_2 in such a way that every edge E joins a node in V_1 to a node in V_2 and each edge has an associated weight. In our projection application, the two partitions are the sets of linguistic units U_s and U_t, in the source and target sentence, respectively. We assume that G is complete, that is, each source node is connected to all target nodes and each target node to all source nodes. Edge weights model the (dis-)similarity between a pair of source and target units.

The optimisation problem from Equation (6.3) identifies the alignment that maximises the product of link similarities. Similarity links between linguistic units in a bi-sentence correspond to edges in our bipartite graph, and thus possible semantic alignments to subgraphs. Optimisation problems over subgraphs in bipartite graphs are usually phrased as minimisation problems over the sum of edge weights (see e.g., Cormen, Leiserson, and Rivest (1990)):

$$\hat{A} = \underset{A \in \mathcal{A}}{\operatorname{argmin}} \sum_{(u_s, u_t) \in A} weight(u_s, u_t) \tag{6.4}$$

At first sight, Equation (6.4) seems to optimise a different figure of merit than Equation (6.3); the two problems are in fact identical when the following edge weight function is used:

$$weight(u_s, u_t) = -\log sim(u_s, u_t) \tag{6.5}$$



U_t / U_s	S_1	VP_2	NP_3	S_4
S_1	**0.58**	0.45	0	0.37
VP_2	0.45	**0.68**	0	0.55
NP_3	0	0	**1**	0.18
S_4	0.37	0.55	0.18	**0.73**

U_t / U_s	S_1	VP_2	NP_3	S_4
S_1	**0.54**	0.80	∞	1.00
VP_2	0.80	**0.39**	∞	0.60
NP_3	∞	∞	**0**	1.70
S_4	1.00	0.60	1.70	**0.31**

Figure 6.2.: Bi-sentence (left) with matrices of constituent similarities (top) and edge weights (bottom). Similarities computed according to Eq. (6.12).

In other words, minimising the sum of negative log similarities corresponds exactly to maximising the product of similarities. Consequently, we can use efficient algorithms developed for graph optimisation problems (Equation (6.4)) and directly interpret their output as a semantic alignment which is optimal according to Equation (6.3). To avoid confusion, we will continue making reference to Equation (6.3) and the semantic similarity measure (sim) which is more intuitive than the graph-related term edge weight.

A complete bipartite graph for the bi-sentence from Figure 6.1 is shown in Figure 6.2 using a matrix notation. The nodes in the graph denote constituents S_1–S_4. The numbers in the top matrix correspond to example similarity scores, the corresponding edge weights of which are shown in the bottom matrix. High similarity scores correspond to low edge weights. Edges with similarity zero are set to infinity (in practice a very large number). Finally, notice that alignments with high similarity scores (low edge weights) are in the diagonal of the matrix.

The projection framework sketched above is unsupervised, as it does not require training data explicitly labelled with semantic role alignments. Nevertheless, it has three parameters which have to be instantiated to obtain complete projection models, all of which can be identified in Equation (6.3):

1. The choice of linguistic units U_s and U_t, which determine the domain and range of A.

2. The definition of the similarity function $\mathtt{sim}$, and

3. The specification of a set of admissible alignments $\mathcal{A}$.

With respect to the linguistic units, we present two broad model families, one based solely on words (Section 6.3.1), and one that uses constituents, obtained either from the output of a chunker or a parser (Section 6.3.2).

6.3.1. Word-based Projection Model

In our first model family, the linguistic units are *word tokens*. Source and target sentences are represented by sets of words, $U_s = \{w_s^1, w_s^2, \dots\}$ and $U_t = \{w_t^1, w_t^2, \dots\}$, and semantic alignments are simply word alignments (e.g., Viterbi alignments produced by statistical word alignment models).

Recall from Section 6.3 that semantic alignments must be admissible, i.e., conform to some notion of well-formedness. It is of course possible to impose constraints on word alignments (e.g., by enforcing one-to-one alignments), and this is often the case for heuristic word alignment models (Melamed, 2000). Statistical word alignment models, on the other hand, already encode constraints as a byproduct of the translation model they are based on. For example, the IBM models introduced in Brown et al. (1993) require each target word to be aligned to exactly one source word (which may be the empty word), thus allowing one-to-many alignments in one direction. Our experiments use word alignments induced by the publicly available statistical word alignment software GIZA++ (Och and Ney, 2003) (see Section 2.2 for details).

Assuming that no other indicators of semantic equivalence are available, we would therefore like to adopt the constraints implicit in the word alignment and assume it *in toto* as an optimal semantic alignment. This

can be achieved by defining a binary similarity function that distinguishes "semantically related" from "unrelated" words as follows:

$$\mathrm{sim}(w_s, w_t) = \begin{cases} 1 & \text{if } w_s \text{ and } w_t \text{ are word-aligned} \\ 0 & \text{else} \end{cases} \tag{6.6}$$

Evidently, the optimisation of a binary similarity function is trivial: the product A of the similarities $\mathrm{sim}(w_s, w_t)$ will be 1 as long as all word pairs are aligned, and 0 otherwise (see Equation (6.3)). As a result, the word alignment is an optimal semantic alignment itself:[1]

$$\hat{A} = \{(w_s, w_t) \mid w_s \text{ and } w_t \text{ are word-aligned}\} \tag{6.7}$$

Given (6.7), the target labelling function for word-based models is:

$$a_t(r) = \{w_t \parallel \exists\, w_s \in a_s(r) : w_s \text{ and } w_t \text{ are word-aligned}\} \tag{6.8}$$

Word-based projection models can be easily derived for different language pairs without recourse to any corpus-external resources. Unfortunately, as discussed in Section 2.2.2, automatically induced word alignments are often noisy, and therefore lead to errors in annotation projection. Contributing phenomena are – as discussed in Section 4.4.1 – monolingual collocations, the inconsistent translation of function words, and the multi-word expressions which are difficult to model in a single word-based account.

As an example for the application of word-based projection, consider again Figure 6.1. Using Equation (6.7), the word alignment links can be interpreted as the following semantic alignment:

$$\begin{aligned} \hat{A} = \{&(\mathrm{Kim}, \mathrm{Kim}), (\mathrm{promised}, \mathrm{versprach}), \\ &(\mathrm{to}, \mathrm{zu}), (\mathrm{time}, \mathrm{p\ddot{u}nktlich})\} \end{aligned} \tag{6.9}$$

The resulting German role labelling function is as follows:

$$a_t = \{\textsc{Speaker} \rightarrow \{\mathrm{Kim}\}, \textsc{Message} \rightarrow \{\mathrm{p\ddot{u}nktlich}, \mathrm{zu}\}\} \tag{6.10}$$

The projection of the Speaker role produces the intended labelling for German; however, the Message role is projected onto an incomplete set

[1] Note that all subsets of a given word alignment are also optimal.

of German words, due to gaps in the word alignment (e.g., the English words *be* and *on* are not aligned with any German words).

The large number and varied nature of such word alignment-related errors precludes their correction by general heuristics, which furthermore would have to be re-developed for new language pairs. We will therefore address these errors on a fundamental level, by integrating structural information into the alignment process.

6.3.2. Constituent-based Projection Model

In our second model family the linguistic units are *constituents*. Consequently, role projection is based on semantic alignments between source ($U_s = \{c_s^1, c_s^2, \dots\}$) and target ($U_t = \{c_t^1, c_t^2, \dots\}$) constituents, as opposed to words.[2] This has two main effects:

Meaningful target spans. The target spans which can receive role annotations are guaranteed to be constituents, in contrast to word-based models, where annotations are projected on arbitrary word spans. This introduces a bias towards linguistically meaningful spans for role annotations.

Increased robustness. Constituents can be aligned correctly even if only a subset of its yield is correctly word-aligned, whereas in word-based models, missing alignments or alignment errors invariably result in wrong semantic alignments, and thus in projection errors. This renders projection more robust to alignment noise.

In general, the additional level of structure increases the expressive power of the framework; it is now possible to explore a wider range of classes of admissible alignments and to define a more plausible similarity function. Our intuition for the similarity function is that it should capture the extent to which two constituents c_1 and c_2 from the two languages express the same semantic content. We approximate this by measuring the *word overlap* between the constituents c_1 and c_2, using the Jaccard coefficient. Let $\mathtt{yield}(c)$ denote the set of words in the yield of a constituent c, and

[2] Constituents can be either recursive (the output of a parser) or non-recursive (the output of a chunker).

$al(c)$ the set of words in the other language aligned to the yield of c. Then:

$$o(c_1, c_2) = \frac{|al(c_1) \cap yield(c_2)|}{|al(c_1) \cup yield(c_2)|} \tag{6.11}$$

Note that this formula is asymmetric; it will consider how well the projection of some constituent $al(c_1)$ matches the constituent c_2 in the other language, but not vice versa. In order to take both target-source and source-target correspondences into account, we model the actual similarity between a pair of source and target constituents c_s, c_t by measuring word overlap in both directions and taking the mean:

$$sim(c_s, c_t) = (o(c_s, c_t) + o(c_t, c_s))/2 \tag{6.12}$$

As an example, consider the two nodes labelled S_1 in Figure 6.2. We first compute the overlap between the source node and target node ($o(c_s, c_t)$): The alignment of the source node's yield consists of two words (*pünktlich*, *zu*), which are both in the yield of the target node; therefore, the size of their intersection is 2. Their union is given by the yield of the target constituents (size 3), resulting in $o(c_s, c_t)$=2/3. Analogously, the target node's projection covers 2 of the 4 tokens of the source node, so that $o(c_t, c_s)$ =1/2. The final similarity score is therefore $sim(c_s, c_t) = 7/12 \approx 0.58$.

We next discuss the most important parameter in constituent-based models, namely the choice of the class of admissible alignments $\mathcal{A}$. The latter allow us to examine in more detail the tradeoff between *expressive* and *corrective* models. On the one hand, more permissive alignments (e.g., which allow one-to-many correspondences such as constituent mergings or splittings) have a higher expressive power, being able to model a wider range of translational divergences. Their downside is that they have a relatively small corrective power, since they rely exclusively on the information provided by the similarity measure. On the other hand, more restrictive alignment models (e.g. which allow only one-to-one alignments between units) cannot capture translational divergences – but due to the bias introduced by the constraints of the model, they can correct errors in the word alignment to a larger extent.

On an ideal corpus, with ideal word alignments, permissive models would presumably work best, being able to capture a great range of trans-

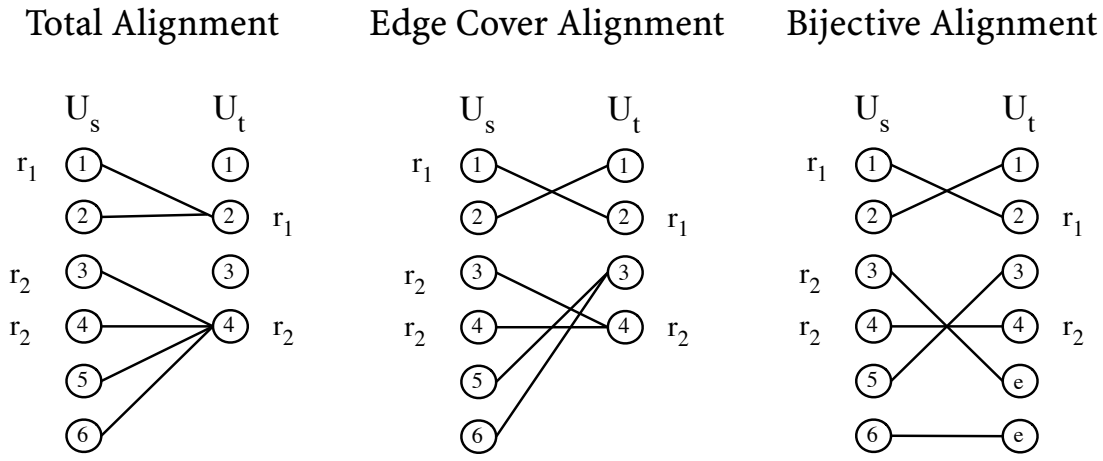


Figure 6.3.: Constituent alignments modelled as bipartite graphs and role projections resulting from different specifications of the class of admissible alignments (U_s, U_t: sets of source and target constituents; r_1, r_2: two semantic roles).

lational divergences. In practice, however, noise in the data cannot be disregarded. We will therefore consider three general increasingly restrictive classes of admissible alignments (total alignments, edge covers, and perfect matchings) by limiting the class of admissible alignments, thereby imposing constraints on the alignments. Examples of bipartite graphs corresponding to these three classes of admissible alignments are shown in Figure 6.3. For each class, we discuss efficient algorithms to solve the corresponding optimisation problem[3].

Total Alignments

We first consider the relatively permissive class of total alignments which only requires that each source constituent be linked to some target constituent. In graph-theoretic terms, this means that total alignments must contain at least one edge (c_s, c_t) for each source constituent c_s. We consider total alignments as an obvious first choice for semantic role projec-

[3]Of course, we may choose to impose no constraints on the alignment at all. In this case, however, the empty alignment will be optimal, since we do not require for any constituent to be aligned.

tion: they guarantee that every source role will be projected onto some target span (i.e., no source roles can be left unaligned). The search for an optimal total alignment corresponds to the following optimisation problem:

$$\hat{A} = \underset{A \text{ is total function}}{\operatorname{argmax}} \prod_{(c_s, c_t) \in A} sim(c_s, c_t) \tag{6.13}$$

Total alignments do not impose any constraints on the target nodes, which can therefore be linked to an arbitrary number of source nodes. Consequently, all alignment links are independent of each other, and we can assemble a globally optimal alignment by combining *locally optimal* alignment links. These in turn can be obtained by linking each source node to its maximally similar target node:

$$\hat{A} = \{(c_s, c_t) \mid c_s \in U_s \wedge c_t = \underset{c'_t \in U_t}{\operatorname{argmax}}\, sim(c_s, c'_t)\} \tag{6.14}$$

We argue that the resulting alignment is optimal by informally showing that all three possible structural modifications to A do not yield total alignments with higher scores. If we removed a link for a source constituent c_s, the alignment would no longer be total. Alternatively, we could replace an existing link (c_s, c_t) with another link (c_s, c'_t); however, $sim(c_s, c'_t)$ cannot be higher than the original $sim(c_s, c_t)$, else the link (c_s, c_t) would not have been created in the first place. Finally, adding a link is guaranteed to produce an alignment with a lower score given that the score is a product of similarity scores (Equation (6.13)). The time complexity of this optimisation procedure is quadratic in the number of constituents: $O(|U_s||U_t|) = O(\max(|U_s|, |U_t|)^2)$.

An example of a total alignment is shown in Figure 6.3 (left) where all source nodes have links to some target node. Total alignments can model one-to-many alignments, albeit only asymmetrically in the target-source direction. In Figure 6.3, target nodes (1) and (3) are inserted and target nodes (2) and (4) are the result of merging source nodes (1)–(2) and (3)–(6), respectively. The example highlights a potential shortcoming for total alignments; due to the nature of the our optimisation procedure, there is a tendency to form alignments primarily with target constituents whose similarity scores are high. In practice this means that potentially important, but idiosyncratic, target constituents with low similarity scores are often left unaligned.

Alignments as Edge Covers

A linguistically implausible feature of total alignments is their asymmetry: translational shifts can be modelled on the target side but not on the source side. To remedy this, we next consider *edge covers*, a symmetrical class of admissible alignments. An edge cover is a subgraph of a bipartite graph where each node is adjacent to at least one edge. Interpreted as a semantic alignment, an edge cover alignment is somewhat more restrictive than a total alignment. It forces *all* source and target constituents to participate in an alignment. This is illustrated in Figure 6.3 (middle), where all source and target nodes are adjacent to an edge (i.e., alignment link). Of course, this means that edge covers cannot model unaligned nodes on either side – each node will be aligned to something. On the other hand, one-to-many alignments can be modelled in both directions. In Figure 6.3 source nodes (3) and (4) are merged into target node (4). Optimal edge cover alignments can be obtained as solutions of the following optimisation problem:

$$\hat{A} = \underset{A \text{ is edge cover}}{\operatorname{argmax}} \prod_{(c_s, c_t) \in A} \operatorname{sim}(c_s, c_t) \tag{6.15}$$

Unlike total alignments, optimal edge cover alignments have to be obtained using *global optimisation*, since the alignment links are no longer independent of one another. Algorithms for computing optimal edge covers have been investigated by Eiter and Mannila (1997) in the context of distance metrics for point sets. They show that minimum-weight edge covers can be reduced to minimum weight perfect matchings (see below) of an auxiliary graph with two partitions of size $|U_s| + |U_t|$. Thus, minimum weight edge covers can be computed in time $O((|U_s| + |U_t|)^3) = O(\max(|U_s|, |U_t|)^3)$, roughly cubic in the number of constituents.

Alignments as Perfect Matchings

Finally, we consider an class of alignments which are even more constrained than edge covers, namely *perfect matchings*. A perfect matching is an edge cover in which each node has *exactly one* adjacent edge. Semantic alignments with this property can be thought of as bijective functions: each source constituent is mapped to one target constituent, and vice versa. An example of a perfect bipartite matching is given in Figure 6.3

(right). Note that the target side contains two nodes labelled (e), a shorthand for "empty" node. Since bi-sentences will often differ in size, the resulting graph partitions will have different sizes as well. In such cases, we introduce empty nodes in the smaller partition to enable perfect matching. Empty nodes are assigned a similarity of zero with all other nodes. Alignments to empty nodes (such as for source nodes (3) and (6)) are ignored for the purposes of projection.[4]

Being bijective, the resulting alignments cannot model one-to-many relations at all; on the other hand, perfect matchings introduce strong competition between nodes and thus have a high corrective power. Unaligned nodes can be modelled only indirectly by aligning nodes in the larger partition to dummy nodes on the other side (see the source side in Figure 6.3 where nodes (3) and (6) are aligned to (e)). Optimal perfect matching alignments are defined as:

$$\hat{A} = \underset{A \text{ is perfect matching}}{\operatorname{argmax}} \prod_{(c_s, c_t) \in A} \operatorname{sim}(c_s, c_t) \tag{6.16}$$

Perfect matchings can be computed efficiently using algorithms for network optimisation (Fredman and Tarjan, 1987; time $O(|U_s|^2 \log |U_s| + |U_s|^2 |U_t|)$). Furthermore, perfect matchings are equivalent to the well-known *linear assignment problem*, for which many solution algorithms have been developed (e.g., Jonker and Volgenant, 1987, with a time complexity of $O(\max(|U_s|, |U_t|)^3)$). The computation of perfect matchings is approximately cubic in the number of constituents, similar to edge cover alignments.

6.3.3. Noise Reduction

As discussed in Section 1.3, noise elimination techniques are generally considered important in annotation projection to obtain accurate projections. The models presented above differ with respect to their robustness to noise. Constituent-based models will be generally less sensitive to noise, since syntactic information will tend to compensate for alignment errors.

[4] Technically, empty nodes can also be introduced into Edge Cover models. However, this will not lead to a proper account of unaligned nodes: due to their low similarity with everything else, empty nodes will only be aligned as a "last resort".

Word-based models will be more error-prone since they rely solely on automatically obtained alignments for effecting the projection. Below we introduce several filtering techniques that either correct or discard erroneous alignments.

Convex complementing. According to our definition of projection (see Equation (6.2)), the span of a projected role r corresponds to the the union of all target units that are aligned to source units labelled with r. This definition is sufficient for constituent-based projection models, where roles rarely span more than one constituent but will yield many wrong alignments. For word-based models, where roles will typically span several source units (i.e., words), the target span of a role will often be a non-contiguous set of words due to errors and gaps in the word alignment. We can repair non-contiguous projections to a certain degree by applying a "convex complementing" heuristic to the output of the word alignment. Without recourse to syntactic information, the heuristic simply fills gaps in a target role span. Let pos return the index of a word token t in given sentence. We define convex complementing as:

$$a_t^{cc}(r) = \{u \mid \min(pos(a(r))) \leqslant pos(u) \leqslant \max(pos(a(r))\} \quad (6.17)$$

Convex complementing repairs missing alignments in a general fashion without being model specific. It is thus applied to all projection models developed in this chapter. This is not the case for the techniques which are limited to particular model classes.

Word filtering. This technique removes words from a bi-sentence prior to projection, according to certain linguistic or alignment-based criteria. We apply two intuitive instantiations of word filtering in our experiments. The first filter removes non-content words from the bi-sentence, i.e., all words which are not adjectives, adverbs, verbs, or nouns. The motivation is similar to the part of speech-based filters used in the context of the projection of frame-evoking elements (Section 4.4.2): content words show a higher translational consistency than non-content words, and their tokens thus have a higher probability of being correctly aligned. As a result, content-word links provide more reliable, but possibly incomplete, guidance for the alignment of constituents.

The second filter removes all words which remain unaligned in the output of the automatic word alignment. This filter follows up on the observation from Section 2.2.2 that the main problem of statistical word alignment models, especially of intersective models, is not incorrectness, but incompleteness. As a result of missing links, the similarity metric (Equation (6.12)) tends to overestimate the denominator, punishing long constituents. By removing unaligned words, only words that are truly aligned to a different constituent are counted as negative evidence.

Argument filtering. Our last filter applies only to constituent-based models defined over full parse trees. Previous work in shallow semantic parsing has demonstrated that not all nodes in a tree are equally probable as semantic roles for a given predicate (Xue and Palmer, 2004). In fact, assuming a perfect parse, there is a "set of likely arguments", to which almost all semantic roles roles should be assigned. We assume that the set of likely arguments can be characterised as the set of all constituents which are direct children of some ancestor of the predicate, provided that (a), they do not dominate the predicate themselves and (b), there is no sentence boundary between the constituent and its predicate.

This definition is structurally similar, but different from Chomsky's definition of government (Chomsky, 1981): Government assumes a deep (e.g., binary branching) syntax tree and characterises the non-subject complements of a predicate. In contrast, our argument filter assumes a treebank-style flat constituent structures as provided by current probabilistic parsers. In such an analysis, the non-subject complements are just siblings of the predicate; our definition concentrates on retrieving arguments outside the maximal projection of the predicate, such as topicalised elements, or the subjects of control constructions (for verbs) and support constructions (for nouns and adjectives). It can be extended slightly to accommodate coordination. We use argument-based filtering to reduce *target* trees to a set of likely arguments. This is illustrated in Figure 6.4, where the tree nodes VP_2 and S_4 are removed (compare with Figure 6.2). As a result, only the target nodes NP_3 and S_1 participate in the optimisation problem. During optimisation, the small number of target nodes resulting from this filter leads to increased competition between source nodes.

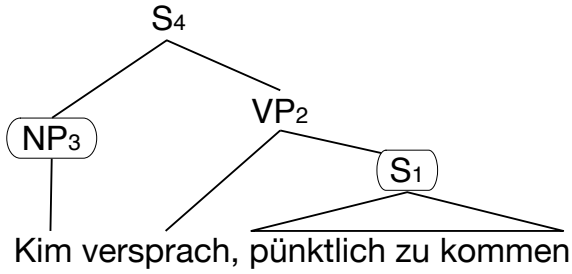


Figure 6.4.: Filtering of unlikely arguments (predicate in boldface, argument candidates after filtering in boxes).

Filter combinations. The filters described above can be used in isolation or in combination. For example, a constituent-based model can be constructed that uses convex complementing, argument-based filtering and both instantiations of word-based filtering. A word-based model can benefit from convex complementing and the filter which removes unaligned words. Non-content words could also be removed provided that information about parts of speech is available. We evaluate the effect of different filtering techniques on the quality of the resulting projections in Section 7.2.

6.4. Summary

This chapter has developed a framework for the induction of semantic roles for new languages within the annotation projection paradigm. The main difficulty of this task is the identification of linguistically plausible target word spans which correspond to source constituents, a problem which we have termed *semantic alignment* (Sections 6.1 and 6.2).

Our framework (Section 6.3) formalises the computation of a semantic alignment as an optimisation problem. When no bracketing information is available, the result is a simple word alignment-based projection. With bracketing information, however, the alignment can be computed as an optimal matching in a bipartite, weighted graph, whose nodes correspond to constituents, and whose edge weights are given by lexical similarity. This model uses syntactic information to restrict the search space to plausible

word spans, and thus has the potential of correcting errors and omissions in the word alignment. The relative weight of the bracketing information can be varied by considering different types of matchings, exploring the tradeoff between more corrective and more expressive matchings. In addition, we have defined a small number of filtering techniques which either repair or eliminate alignment errors.

7. Experimental Evaluation

In this chapter, we provide an experimental evaluation of the framework developed in Chapter 6. We project semantic roles from English onto two different target languages, French and German, using the EUROPARL sample described in Section 3.3 as parallel corpus, and comparing different classes of admissible alignments and filtering methods. In addition, we contrast projection from gold-standard annotation in the source language with projection from automatically obtained annotation.

We begin by detailing our general experimental setup (Section 7.1). Next, we present the actual experimental evaluation for the language pairs English–German (Section 7.2) and English–French (Section 7.3). After reviewing related work (Section 7.4), the chapter concludes with a general discussion (Section 7.5).

7.1. Experimental Setup

In this section, we discuss our parameter exploration procedure, give details on the data, and explain our evaluation measure.

7.1.1. Evaluation as Model Selection

A practical implementation of the projection framework presented in Chapter 6 has to make choices for the three parameters introduced in the last chapter: (a) the linguistic units employed, (b) the class of admissible alignments, (c) the choice of filter for eliminating noise from the word alignment. We consider the following instantiations for each parameter:

- **Units:** words, non-recursive constituents (i.e., chunks), recursive constituents

- **Admissible alignments:** Total alignments (Total), Edge covers (EdgeCover), Perfect matchings (PerfMatch)
- **Filtering:** None, remove non-aligned words (NA), remove non-content words (NC), remove non-arguments (Arg)

These choices lead to a substantial model space, in which we seek to identify the best-performing model. For these reasons, both Experiment 1 (for the language pair English–German) and Experiment 2 (for the language pair English–French) are structured as follows:

Condition 1: Gold standard roles. In this condition, we use manually annotated semantic roles for English. The task of the projection models is to induce semantic roles for German, which are evaluated against German gold standard roles. In order to isolate the best performing models, we follow best practice in machine learning and split our corpus randomly into a development and test set (each 50% of the data).

Step 1: Model selection. In Step 1, we evaluate the model space exhaustively on the development set using different combinations of unit, filter, and admissible alignment together with intersective (automatically produced) alignments.[1] All models use the convex complementing heuristic discussed in Section 6.3.3, which attempts to repair wrong alignments. Also, we do not present results for combinations of filters; in preliminary experiments, these yielded inferior performance to individual filters.

Step 2: Model validation. In Step 2, we validate these findings by applying the best-performing models from Step 1 on the test set. We use intersective and manual word alignments to gauge the extent to which alignment errors influence the overall performance, and particular model classes.

[1]Note that not all combinations of the parameters listed above are possible: (1), the different admissible alignments are only applicable to constituent-based models (which either use chunks or full constituents); (2), the filter that removes non-arguments (Arg) can only be used with a model whose linguistic units are recursive constituents.

Condition 2: Automatically assigned roles. In this condition, the semantic role input is taken from the output of a state-of-the-art automatic shallow semantic parser. Again, roles are projected onto German and evaluated against German gold standard roles. We repeat Step 2 of Condition 1, i.e., the model validation step on the test set, both for intersective and manual word alignment. In this way, we both assess the impact of using noisy input data for annotation projection, and test whether the model selection on gold standard data carries over to noisy input data.

Experiment 2 differs from this outline in that two resources, namely a chunk-based syntactic analysis, and a manual word alignment, were not available for French. In consequence, we omit the consideration of chunk-based models, and compare all models on the basis of the intersective word alignment only.

7.1.2. Data sources

Corpus and Annotation. All our models are evaluated on the trilingual parallel sample corpus (English–German–French). As described in Section 3.3, this corpus was annotated manually with semantic roles. The annotation was performed independently for each language. We also obtained automatic semantic role annotation for the English side of the corpus, using the shallow semantic parser for English by Giuglea and Moschitti (2004) (see Section 7.2.2 for details).

Word alignment. For English–German, we used both the intersective word alignment created by GIZA++ and a manually corrected word alignment (see Section 2.2.2). For English–French, only the automatic intersective alignment was available.

Syntactic analysis. For constituent-based models, we experimented with recursive and non-recursive constituents. English sentences were parsed with Collins' (1997) parser and Abney's (1996) chunker. For German, we used Dubey's (2004) parser and the chunker developed by Schmid and Schulte im Walde (2000). For French, we only had a full parser (Bouri-

gault et al., 2005) at our disposal. Further details on these tools can be found in Section 2.3.1.

Since the parsers are operating out of domain, which is known to degrade parsing accuracy (Gildea, 2001), we obtain an additional assessment of the quality of the recursive syntactic analyses. Since full manual evaluation is beyond our means, we consider a task-specific correlate of analysis quality, namely the degree to which the manually annotated semantic roles in our corpus correspond to single constituents. The choice of this agreement mirrors the intuition that almost all semantic roles correspond to single constituents when manual syntactic annotation is provided (Erk et al., 2003). Deviations from this rule indicate errors in the syntactic analysis. Syntax–role agreement can also be motivated from a projection point of view, since multi-constituent roles require multiple alignments to be correct, and are thus more prone to errors.

The actual agreements on the sample corpus are as follows: In the English text analysed with Collins' (1997) parser, 93% of all roles correspond to single constituents (98% to one or two constituents). In the German text, parsed with Dubey (2004), the numbers are 88% for one constituent, and 95% for up to two constituents. For French, only 82% of the roles were assigned to single constituents; however, 93% correspond to one or two constituents. These numbers indicate that the bracketing for English is best suited for our purposes, followed by German, and finally French. However, all parsers show solid performance in identifying spans that correspond to potential roles, which gives them a good chance of supporting the task at hand.

7.1.3. Evaluation

Evaluation measure. We measure model performance using labelled precision, recall, and F-Score in the "Exact Match" condition in which both the label and the span of a projected role have to match the gold standard annotation for the target language to count as a true positive.[2] We also assess whether differences in performance are statistically significant using stratified shuffling (Noreen, 1989), an instance of assumption-free

[2] We do not consider the second plausible evaluation condition, Headword Match, due to the higher manual effort involved with correcting automatic head word identification.

approximative randomisation testing (see Yeh (2000) for a discussion).

Baseline. As a baseline, we will use the simplest model family, namely word-based projection models. This model family does not require syntactic analysis, but relies exclusively on word alignments for the projection, and has linear time complexity. The use of more elaborate linguistic information in the form of constituent-based models is only warranted if they can deliver significant improvements over word-based models.

Upper bound. In Section 3.3, we have estimated the inter-annotator agreement for the monolingual portions of our corpus. We obtained 0.83 for German, and 0.72 for French in the Span Match condition which measures annotation of the same roles with the same span. As discussed in Section 3.3, these numbers represent an upper bound for the performance of a shallow semantic parser *within* these languages. It is more difficult to determine a ceiling for the *cross-lingual* projection task, since in addition to inter-annotator agreement, we have to take into account the effect of cross-lingual divergence. Our annotation study did provide an estimate of the cross-lingual role divergence (Role Parallelism, see Table 3.3 on page 66). Unfortunately, this number also incorporates inter-annotator disagreement on roles, since the sentences in the main corpus are only annotated by a single coder.

Without a better way of gauging the effect of cross-lingual divergence, we resort to using the monolingual upper bounds: In Condition 1, where manually annotated data is used as input for projection, we adopt the monolingual inter-annotator agreement (Span Match) as an upper bound; in Condition 2, where automatically annotated roles are used, the accuracy of the shallow semantic parser, evaluated against the English gold standard, is used.

7.2. Experiment 1: Language Pair English–German

7.2.1. Condition 1: Projection from Gold Standard Roles

In Condition 1, we use the manually annotated gold standard roles as English input for the projection process. As outlined above in Section 7.1, we first discuss our results on the development set. The best model instantiations are next evaluated on the test set.

Step 1: Model Selection on the Development Set

Word-based models. The results of the word-based models, which we consider as our baseline, are summarised in Table 7.1. Without using any linguistic information or noise filtering, we obtain a modest F-Score of 45.1% (No Filter). Removing non-aligned words (NA Filter) does not alter this performance. This is not surprising, since non-aligned words are not taken into account in word-based models (see Equation (6.7)). As an example, recall the projection error discussed in Section 6.3.1, where the purely word-based projection of the source role MESSAGE in Figure 6.1 (*to be on time*) results in the incomplete target role span (*pünktlich zu*).

Removing non-content words (NC Filter) yields significantly worse results ($p < 0.01$). This is also expected since the word-based models cannot recover words in the target span that have been deleted by a filter. The convex complementing heuristic which we apply to all models can fill gaps in the interior of a role, but still requires that the leftmost and the rightmost word of each role's target span are aligned correctly. Spans like the MESSAGE span in Figure 6.1 thus cannot be repaired.

Chunk-based models. We next examine constituent-based models utilising chunks as linguistic units. Here we can consider different categories of admissible alignments as well as different filtering techniques[3]. The results for chunk-based models are summarised in Table 7.2. Generally, we

[3]Recall that we cannot apply the argument filter since it requires recursive syntactic analysis.

Model	Precision	Recall	F-Score
No Filter	45.6	44.8	45.1
NA Filter	45.6	44.8	45.1
NC Filter	36.4	31.7	33.9

Table 7.1.: Experiment 1, Condition 1: Model comparison for word-based models (intersective word alignment, development set)

	Model	Prec	Rec	F-Score
No Filter	WordBL	45.6	44.8	45.1
	Total	48.5	41.9	45.0
	EdgeCover	**51.6**	**45.1**	**48.1**
	PerfMatch	51.8	41.3	46.0
	UpperBnd	–	–	83.0

	Model	Prec	Rec	F-Score
NA Filter	WordBL	45.6	44.8	45.1
	Total	47.7	38.2	42.4
	EdgeCover	49.7	44.4	46.1
	PerfMatch	**53.0**	**43.9**	**48.0**
	UpperBnd	–	–	83.0

	Model	Prec	Rec	F-Score
NC Filter	WordBL	36.4	31.7	33.9
	Total	45.3	32.0	37.5
	EdgeCover	48.3	41.1	44.4
	PerfMatch	**50.4**	**41.0**	**45.2**
	UpperBnd	–	–	83.0

Table 7.2.: Experiment 1, Condition 1: Model comparison for chunk-based models (intersective word alignments, development set)

observe that global graph-based models outperform the word-based baseline (WordBL). Without any filtering, EdgeCover yields significantly better results than WordBL, Total and PerfMatch ($p < 0.01$). When the NA filter is applied, PerfMatch significantly outperforms all other models ($p < 0.01$). Finally, the NC Filter does not seem to benefit chunk-based models. Here, EdgeCover and PerfMatch yield comparable F-Scores (around 45%; the difference is not statistically significant) but are significantly outperformed by EdgeCover and PerfMatch in the No Filter and NC Filter conditions, respectively ($p < 0.05$).

The overall modest performance of chunk-based models can be explained by the fact that almost all roles span more than one chunk. In fact, in the English gold standard corpus roles span on average almost 5 chunks, thus introducing a large margin for projection errors. In addition, roles assigned to PPs and VPs often do not correspond to chunk boundaries. The following bi-sentence presents a common problem for chunk-based models (syntactic structure in round brackets, semantic roles in square brackets):

(7.1) A director has publicly **stated** that [($_{\text{Clause}}$ the future of the plant is in doubt)]$_{\text{Message}}$.

Ein Direktor hat öffentlich **erklärt**, [($_{\text{NP}}$ die Zukunft der Anlage) sei ungewiss]$_{\text{Message}}$.
A director has publicly **declared**, [($_{\text{NP}}$ the future of the plant) is uncertain]$_{\text{Message}}$.

Ideally, we would like to project the English Message role, which spans the clause *the future of the plant is doubt*, onto the complete German clause *die Zukunft der Anlage sei ungewiss*. However, since there is no German chunk corresponding to this clause, a model using Total alignments will link the English clause to the single best candidate, namely the noun phrase *die Zukunft der Anlage*, resulting in an incomplete role span for German.

This effect can be alleviated – to a small degree – by choosing more restrictive alignments. As shown in Table 7.3, we obtain better results for EdgeCover alignments in the No Filter condition. Recall that EdgeCover forces every source unit to be aligned, thus delivering better recall and precision over Total alignments in the No Filter condition. When the

NA Filter is applied, which removes non-aligned material, even more restricted alignments become profitable so that the one-to-one alignments produced by PerfMatch yields performance gains over EdgeCover.

Full Constituent-based models. Table 7.3 shows the results for constituent-based models when a full parser is employed. First, observe that these models yield substantially better overall results than word- or chunk-based models, with F-Scores ranging between 58.8% and 67.2%. All filter and alignment combinations are significantly better ($p < 0.01$) than either WordBL or the two best chunk-based models (EdgeCover (No Filter) and PerfMatch (NA Filter)).

Note that filtering generally improves the resulting projections. In the No Filter condition, all constituent-based models (i.e., Total, EdgeCover, PerfMatch) yield comparable performance ranging from 58.8% to 60.4% (the differences are not statistically significant). In the NA Filter condition, the best model is PerfMatch. It is significantly better than Total and EdgeCover in the same condition ($p < 0.01$) and all constituent-based models in the No Filter condition ($p < 0.01$). Although the NC Filter delivers improvements over No Filter, its performance is worse when compared to the NA or Arg Filters. The Arg Filter yields results comparable to NA and the highest precision among all models (PerfMatch 80.3%). The latter is offset by a relatively low recall (PerfMatch 47.8%). Recall that the Arg Filter removes all nodes from the target tree except for those deemed likely arguments. With only a few nodes left, PerfMatch enforces a strong competition among source nodes for links to target nodes, resulting in a high-precision alignment. On the other hand, roles assigned to constituents erroneously removed by the Arg Filter (e.g. due to parser errors) cannot be recovered, and therefore recall is low.

Mirroring our observations for chunk-based models, we also find an interaction between the filtering scheme and the optimal admissible alignment for full constituents. Cross-lingual syntactic differences occur often in the unfiltered condition and call for the modelling of one-to-many relationships. In consequence, the more restrictive models have low recall (compare for example Total's recall of 55.9% to PerfMatch's recall of 51.8%). The situation is different for the filtering conditions, since filtering out, e.g., non-aligned words makes the two syntactic trees in the source and target bi-sentence more similar. In this situation, the weaker expressive

	Model	Prec	Rec	F-Score
No Filter	WordBL	45.6	44.8	45.1
	Total	**65.8**	**55.9**	**60.4**
	EdgeCover	64.2	55.8	59.7
	PerfMatch	68.0	51.8	58.8
	UpperBnd	–	–	83.0

	Model	Prec	Rec	F-Score
NA Filter	WordBL	45.6	44.8	45.1
	Total	74.1	56.1	63.9
	EdgeCover	69.0	61.6	65.1
	PerfMatch	**73.3**	**62.1**	**67.2**
	UpperBnd	–	–	83.0

	Model	Prec	Rec	F-Score
NC Filter	WordBL	36.4	31.7	33.9
	Total	64.3	47.5	54.6
	EdgeCover	67.7	57.2	62.0
	PerfMatch	**71.6**	**55.6**	**62.6**
	UpperBnd	–	–	83.0

	Model	Prec	Rec	F-Score
Arg Filter	WordBL	45.6	44.8	45.1
	Total	**69.7**	**60.6**	**64.8**
	EdgeCover	69.4	60.4	64.6
	PerfMatch	80.3	47.8	59.9
	UpperBnd	–	–	83.0

Table 7.3.: Experiment 1, Condition 1: Model comparison for full constituent-based models (intersective word alignments, development set)

power of restrictive alignments (PerfMatch) is sufficient for modelling purposes, and their higher corrective power leads to more precise alignments. Thus, we find that PerfMatch performs best in conjunction with the NA and NC Filters.

In sum, we find that the restrictive models (EdgeCover and PerfMatch) outperform the permissive models (Total) when filtering takes place. The best overall model is PerfMatch in the NA Filter condition. In general, we find that EdgeCover and PerfMatch make similar errors (90% overlap). Differences arise mostly when roles span more than one constituent, e.g., due to misparses. Consider as an example the sentence pair:

(7.2) The Charter is [$_{NP}$ an opportunity to bring the EU closer to the people].

Die Charta ist [$_{NP}$ eine Chance], [$_{S}$ die EU den Bürgern näherzubringen].
The Charter is [$_{NP}$ a chance], [$_{S}$ the EU to the citizens to bring closer].

Ideally, the English NP would be aligned to both the German NP and S. EdgeCover, which can model one-to-many-relationships, acts "confidently" and aligns the NP to the German S to maximise the overlap similarity, incurring both a precision and a recall error. PerfMatch, on the other hand, cannot handle one-to-many relationships, acts "cautiously" and aligns the English NP to a empty node, leading only to a recall error. This means that even though EdgeCover's analysis is partly right, it will be judged worse than PerfMatch, given the current evaluation method.

Step 2: Model Validation on the Test Set

Our experiments on the test set focus exclusively on models employing full constituents as linguistic units, since this model family has outperformed word-based and chunk-based models by a large margin on the development set. We combine each class of admissible alignments with the filtering condition that has worked best on the development set. We verify the performance of these models on the test set, using both intersective and manual word alignments.

	Model	Prec	Rec	F-Score
Intersective	WordBL	45.7	45.0	45.3
	Total (Arg Filter)	72.3	63.1	67.4
	EdgeCover (NA Filter)	71.9	64.6	68.1
	PerfMatch (NA Filter)	**75.5**	**63.5**	**69.0**
	UpperBnd	–	–	83.0

	Model	Prec	Rec	F-Score
Manual	WordBL	61.6	60.4	61.0
	Total (Arg Filter)	72.5	68.6	70.5
	EdgeCover (NA Filter)	71.2	69.1	70.1
	PerfMatch (NA Filter)	**75.2**	**67.2**	**71.0**
	UpperBnd	–	–	83.0

Table 7.4.: Experiment 1, Condition 1: Comparison of the best full constituent-based models on the test set (intersective and manual word alignments)

Our results are summarised in Table 7.4. When intersective alignments are used (top half), we find that the PerfMatch alignment model yields the highest overall F-Score, followed by EdgeCover and Total.[4] Both PerfMatch and EdgeCover are significantly better than Total and WordBL ($p < 0.05$). Considering that two of the three best-performing models use the NA Filter, and that PerfMatch shows the highest precision and recall, we conclude that the best and most robust choice is a constituent-based model with PerfMatch alignment and the NA Filter.

The bottom half of Table 7.4 shows results for the same models, using manually annotated word alignments (see Section 7.1 for details). A comparison of these numbers with the results from the top half indicates by how much the performance of the projection models improves when perfect word alignments are available. First, note that the performance of WordBL increases sharply from F=45.3% to F=61.0%. In contrast, the per-

[4]Our results on the test set are slightly higher in comparison to the development set. The fluctuation reflects natural randomness in the partitioning of our corpus (compare also the differences between development and test set in Experiment 2, Section 7.3).

formance of constituent-based models increases only by 2%–3% F-Score. Performance differences between Total, EdgeCover and PerfMatch are small and not statistically significant. The pronounced improvement of the baseline model underlines the important role of bracketing to correct errors from automatic word alignments. Conversely, models using bracketing information are bound to profit less from cleaner word alignments.

Note that gold standard word alignments alone do not lead to perfect role projection. Even though the impact of bracketing information is less dramatic, it still considerable (around 10% F-Score). The models using manual word alignments also still perform substantially below the upper bound. This holds for both constituent-based (10% F-Score below) and word-based (20% F-Score below) models. Analysis of the projection output identified two main sources of errors: genuine cases of cross-lingual divergence and parsing errors. As an example of cross-lingual divergence, consider the case of pronominal adverbs in German. Many German verbs such as *glauben* ("believe"), which subcategorise for a prepositional phrase, exhibit a diathesis alternation: if the PP has a propositional content, the prepositional phrase *X glaubt an Y* ("X believes in Y") can be replaced by the corresponding pronominal adverb *daran* plus an embedded clause: *X glaubt daran, dass Y* ("X believes that Y"). Even though the pronominal adverb forms part of the complement clause (and therefore also of a role assigned to this clause), it has no English counterpart. Thus, it does not occur in the word alignment, and will therefore not be recoverable in the projection model. Another example is shown below where English has chosen to elide a constituent in an ellipsis, while German does not:

(7.3) We claim X and $[\text{we}]_{\text{Speaker}}$ **say** Y
Wir behaupten X und — **sagen** Y

The word alignment (correctly) aligns the German pronoun *wir* with the first English *we* and leaves the second occurrence unaligned. Since there is no corresponding German word for the second *we*, projection of the role filled by the second *we* fails.

The remaining errors can be mostly traced back to bracketing errors. When the boundaries of constituents do not correspond to role boundaries, projection becomes problematic. In fact, this is often the case when

attachment decisions are involved (see Example (7.2)). Additional compounding factors in this respect are the rather flat syntactic analysis which Dubey's (2004) parser provides for German, and the fact that both the English and German parser are trained on newspaper text, but are applied to transcriptions of spoken text.

7.2.2. Condition 2: Projection from Automatic Roles

In Condition 1, we used manually annotated roles on the English side as input for the projection. Although this setup allows to study the projection framework in detail without worrying about the provenance of the semantic role annotation, it is far from realistic. If annotation projection is to be used for the creation of large scale semantic resources for new languages, then the source side of the parallel corpus will have to be annotated using an automatic role assignment system. In this section, we assess the extent to which the noise in *automatically assigned* semantic roles impacts the quality of projected alignments, using role annotations obtained from a shallow semantic parser as the input for projection.

We trained a state-of-the-art shallow semantic parser developed by Giuglea and Moschitti (2004) on the FrameNet (release 1.2) corpus. The system treats the assignment of semantic roles as a classification task on constituents, which is modelled with a support vector machine.[5] We used Giuglea and Moschitti's "standard" features and not the "extended feature set" which is based on PropBank and would have required us to produce a PropBank analysis of the entire FrameNet corpus. In their paper, Giuglea and Moschitti report an accuracy of 85.2% on a held-out part of the FrameNet corpus, using the "standard" feature set.

We applied the shallow semantic parser to the English side of our parallel corpus to obtain semantic roles, treating the frames as given.[6] Table 7.5 shows an evaluation of the shallow semantic parser's output on our test set against the English gold standard annotation. Since Giuglea and Moschitti's (2004) implementation can currently handle only verbs, we assessed the performance for the complete test set and the subset of

[5] We are grateful to Ana-Maria Giuglea and Alessandro Moschitti for letting us use the system. See Giuglea and Moschitti (2006) for the description of the current release.

[6] This decomposition of frame-semantic parsing is common practice in shallow semantic parsing (cf. Section 1.2).

Evaluation condition	Prec	Rec	F-Score
All predicates	78.1	55.8	65.1
Verbs only	78.1	62.4	69.4

Table 7.5.: Evaluation of Giuglea and Moschitti's (2004) shallow semantic parser on the English side of our parallel corpus (test set)

verbal predicates (87.5% of instances in the test set). The overall performance (65 and 69% F-Score, respectively) is lower than what Giuglea and Moschitti report, but this is expected since our test set differs from the training data in vocabulary, which affects the lexical features, and suffers from parsing errors, which affects the syntactic features. The difference between the complete and verbs-only data sets amounts to 4% F-Score, which are entirely due to the recall missed through unassigned roles for nouns and adjectives. In the following, we report results on the complete test set (including nouns and adjectives) in order to make our evaluation comparable to Experiment 1.

Using the automatically annotated roles as input, we performed projection in the same manner as in Experiment 1. Table 7.6 shows results for PerfMatch (NA Filter), the best projection model from Experiment 1, comparing the performance for intersective and manual word alignments.[7] The performance of this informed model is contrasted to the word-based baseline, and to the performance of the shallow semantic parser, which can be seen as an upper bound of the quality achievable by projection in Condition 2.

Performance on automatically annotated input is approximately 11% F-Score lower than on manual annotation (compare Table 7.6 and Table 7.4). WordBL yields 34.2% F-Score and PerfMatch 56.6% (both on intersective alignments). Note that PerfMatch outperforms WordBL by 22.4%, indicating that the advantages of constituent-based models become even more pronounced on noisy input data. Similarly to Experiment 1, we observe that manual alignments boost performance for WordBL, but

[7] A more detailed comparison of the alignment models verified that the ranking from Condition 1 indeed carries over: The combination PerfMatch with NA Filter significantly outperforms all other models in Condition 2 ($p < 0.05$).

Model (Intersective Word Alignment)	Prec	Rec	F-Score
WordBL	41.3	29.2	34.2
PerfMatch (NA Filter)	**73.7**	**45.9**	**56.6**
UpperBnd	78.1	55.8	65.1
Model (Manual Word Alignment)	Prec	Rec	F-Score
WordBL	49.2	34.9	40.8
PerfMatch (NA Filter)	**73.5**	**47.6**	**57.8**
UpperBnd	78.1	55.8	65.1

Table 7.6.: Experiment 1, Condition 2: Performance of best constituent-based model on the test set, using automatically labeled semantic roles as input

bring relatively small gains for PerfMatch (approximately 2% F-Score).

In order to assess how much errors in the automatic labelling affect projection, we split the shallow semantic parser's output into three bands: (a) instances with no role labelling errors (Error 0, 36% of the test data), (b) instances with one labelling error (Error 1, 34% of the test data), and (c) instances with two or more labelling errors (Error 2+, 31% of the test data). Table 7.7 shows PerfMatch's projection performance for each of these bands when intersective and manual alignments are used. As can be seen, when the output of the shallow semantic parser is error-free, projection performance is relatively good, reaching 80.1% F-Score (intersective). The more errors are found in the automatic output, the worse projection results we obtain. Interestingly, we find that PerfMatch is to a certain degree robust to noise, delivering projections with high precision even in the face of imperfect data (see Error 1 in Table 7.7). The low recall values for bands Error 1 and 2+ are a result of the semantic parser's low recall. Note that the projection of noisy data (Error 2+) is actually worse for manual alignments than for intersective alignments: manual word alignments provide very strong evidence for semantic alignments, which faithfully reproduce the semantic parser's errors in the target language.

In sum, these results are encouraging: They show that projection can yield role annotations from automatically annotated roles with almost the same precision as from manually annotated roles. The restrictive

Band (Intersective)	Prec	Rec	F-Score
Error 0	85.8	75.1	80.1
Error 1	77.4	35.5	48.7
Error 2+	39.1	17.3	24.0
Band (Manual)	Prec	Rec	F-Score
Error 0	87.4	77.8	82.3
Error 1	76.4	37.8	50.6
Error 2+	35.6	16.5	22.5

Table 7.7.: Experiment 1, Condition 2: PerfMatch's performance by errors rate in automatic semantic role labelling per frame (Error 0: no labelling errors, Error 1: one labelling error, Error 2+: two or more labelling errors)

PerfMatch alignment model with its "cautious" strategy (see Section 7.2.1) appears especially suited for this type of input data. Moreover, the shallow semantic parser we employ is not optimised for the EUROPARL corpus, only assigns roles to a subset of our test corpus, and uses a basic feature set. Considering a more elaborate feature space would presumably increase the semantic parser's accuracy further (see Giuglea and Moschitti (2004) for details). However, this was outside the scope of the present evaluation whose aim was to determine the prospects of projection using a state-of-the-art shallow semantic parser off the shelf, without spending effort on possible optimisations within that system.

7.3. Experiment 2: Language Pair English–French

The purpose of this experiment is to determine how well the results we obtain for the language pair English–German generalise to another language pair, English-French, which involves a target language from a different language family: French is a Romance language, while English and German are both Germanic languages.

Model	Precision	Recall	F-Score
No Filter	53.3	48.3	50.7
NA Filter	53.3	48.3	50.7
NC Filter	32.5	28.6	30.4

Table 7.8.: Experiment 2, Condition 1: Model comparison for word-based models (intersective word alignment, development set)

7.3.1. Condition 1: Projection from Gold Standard Roles

In this condition, we use English gold standard roles as input for the projection process. The projected roles are evaluated against the French gold standard. As in Experiment 1, we first compare all possible model instantiations on the development set (Step 1).

Step 1: Model Selection on the Development Set

Word-based models. The results for the word-based baseline models are summarised in Table 7.8. The results are broadly similar to those for German: Without filtering (No Filter), projection works moderately well, yielding an F-Score of 50.7%. Again, due to the nature of the word-based projection model, removing non-aligned words (NA Filter) does not alter this performance. The removal of non-content words by the NC Filter results a sharp performance drop to 30.4%, similar to German. This decrease, which is statistically significant at $p < 0.01$, results predominantly from the removal of role-initial or role-final non-content words such as prepositions (*par* des mafias) or determiners (*un* exemple) which cannot be recovered by the convex complementing heuristic.

In total, the performance of the word-based models is around 5% F-Score better for French than for German, which is reflected in both higher precision and recall values. We ascribe this difference, at least with respect to recall, to the more similar word orders of English and French, which lead to a less sparse word alignment. In fact, the intersective alignment for English–German for our sample contains 12.812 alignment links, while

the English–French alignment consists of 15.517 links.[8]

Recall that no chunk analysis is available for French; we therefore continue directly to considering full constiuent-based models.

Full Constituent-based models. Table 7.9 shows the results for constituent-based models for French when a full parser is employed. By and large, our observations correspond well to the results we obtained in Experiment 1 for German (compare Table 7.3). Constituent-based models substantially outperform word-based models for both languages. For French, we obtain F-Scores of up to 64.2%, an improvement of 13.5% F-Score over the word-based models.

All constituent-based models significantly outperform the corresponding baseline for the filtering scheme ($p < 0.05$). In fact, all models but one significantly outperform the best word-based model, NA Filter. The low performance of the exception, Total Alignment with NC Filter, is due to the overly aggressive NC Filter which removes crucial words, and whose errors cannot be alleviated by the comparatively lenient Total Alignment (see the discussion of admissible alignment classes in Section 6.3.2).

In parallel to German, we find that filtering almost universally improves resulting projections. In the No Filter condition, all constituent-based models yield comparable performance ranging from 53.5% to 55.9%. Using the NC filter results in a substantially higher performance of up to 62% F-Score (with the exception of the Total Alignment model mentioned above). Both the NA and Arg Filters lead to even higher, and more consistent, improvements than the NC Filter. All models for these two filters outperform their counterparts in the No Filter condition highly significantly ($p < 0.01$). These two filters also lead to the two globally best models for French, namely PerfMatch with NA Filter (63.4% F-Score), and EdgeCover with Arg Filter (F-Score 64.2%).

Application of the Argument Filter not only leads to the best overall model, but also to the model with the highest precision (PerfMatch ArgF) at 84.2%. This is the same model instantiation that obtained the highest precision for German (80.3%). The models however share the problem of low recall (47.3% for the French model).

[8] Compare our similar observations in Section 5.1.2.

	Model	Prec	Rec	F-Score
No Filter	WordBL	53.3	48.3	50.7
	Total	57.2	50.2	53.5
	EdgeCover	**60.0**	**52.3**	**55.9**
	PerfMatch -	60.9	49.6	54.7
	UpperBnd	–	–	72.0

	Model	Prec	Rec	F-Score
NA Filter	WordBL	53.3	48.3	50.7
	Total	68.1	50.2	57.8
	EdgeCover	65.9	59.7	62.6
	PerfMatch	**68.9**	**58.7**	**63.4**
	UpperBnd	–	–	72.0

	Model	Prec	Rec	F-Score
NC Filter	WordBL	32.5	28.6	30.4
	Total	60.5	36.6	45.6
	EdgeCover	66.8	57.6	61.9
	PerfMatch	**69.4**	**56.6**	**62.3**
	UpperBnd	–	–	72.0

	Model	Prec	Rec	F-Score
Arg Filter	WordBL	53.3	48.3	50.7
	Total	71.4	58.1	64.1
	EdgeCover	**71.5**	**58.3**	**64.2**
	PerfMatch	84.2	47.3	60.6
	UpperBnd	–	–	72.0

Table 7.9.: Experiment 2, Condition 1: Model comparison for full constituent-based models (intersective word alignments, development set)

Some differences between French and German arise through the more fragmented nature of the French parses: Recall from Section 7.1.2 that we found a higher incidence of syntax–role mismatches (i.e., semantic roles that could not be assigned to a single constituent) in the French data compared with German. This means that reliance on the syntactic bracketing is more likely to yield wrong projections for French, and correspondingly, the results for French are generally somewhat lower in absolute numbers. For example, the best French model (64.2% F-Score) trails the best German model (67.2%) by 3% F-Score. However, compared to the results for word-based models, this absolute difference is rather small. In addition, these results should be interpreted in relation to the upper bound for projection. This ceiling is also considerably lower for French (72%), an effect which can at least be partly attributed to the fragmented nature of the parses (cf. Section 7.1.3). In this context, the results for French look more favourable, with the best model only 8% below the upper bound, while the difference is 16% for German.

The properties of the syntactic analysis for French bracketing also lead to interesting changes in the relative performance of alignment models and filters. As for alignment models, we find that the best model is either EdgeCover or PerfMatch, which underlines the benefits of using a more restrictive alignment model for French as well. However, EdgeCover outperforms PerfMatch in a number of French settings, which does not happen in German. The reason is that PerfMatch cannot produce correct projections for roles which span more than one constituent in the French gold standard. They can, however, be modelled by EdgeCover. In consequence, PerfMatch consistently exhibits a better precision, but worse recall than EdgeCover. Depending on the relative size of the two effects for a particular filtering condition, either one can result in a better overall F-Score. On the level of filtering procedures, another difference to German is that the Arg Filter models are competitive to the NA Filter models. This observation can be traced back to the overall lower recall and precision level for French where the increases in precision for the Arg Filter outweigh the accompanying losses in recall.

Model	Prec	Rec	F-Score
WordBL (NA Filter)	50.6	48.1	49.3
Total (Arg Filter)	68.3	57.9	62.7
EdgeCover (Arg Filter)	68.1	57.9	62.6
PerfMatch (NA Filter)	**66.2**	**60.3**	**63.1**
UpperBnd	–	–	72.0

Table 7.10.: Experiment 2, Condition 1: Comparison of the best full constituent-based models on the test set (intersective word alignment)

Step 2: Model Validation on the Test Set

In this step, we verify the performance of the best-performing models for each class of admissible alignments on the test set. Note that only the intersective word alignment is available for French (cf. Section 2.2.2).

Table 7.10 lists our results. We find that the performance of all models is between 0.5 and 2% lower than on the development set.[9] Otherwise, the advantage conferred by syntactic information remains unchanged: all constituent-based models outperform the word-based baseline by most than 13% F-Score. The difference is highly significant ($p < 0.01$).

We obtain the best result, 63.1% F-Score, for the PerfMatch model with NA Filter, the same model which performed best on the German test set. The two best models on the French development set (Total and EdgeCover with Arg Filter) perform numerically worse, but the difference of 0.5% F-Score is not statistically significant. In total, the test set corroborates our observations on the development set. In particular, application of both the NA Filter and the Arg Filter reliably result in well-performing models.

[9] This is the inverse of the pattern we observed for the language pair English–German, where results on the test set were slighly better. We take this as evidence that there is no systematic bias in the split of our corpus in development and test set.

Model (Intersective Word Alignment)	Prec	Rec	F-Score
WordBL	54.6	38.9	45.4
PerfMatch (NA Filter)	**70.2**	**48.3**	**57.2**
UpperBnd	78.1	55.8	65.1

Table 7.11.: Experiment 2, Condition 2: Performance of best constituent-based model on the test set, using automatically labeled semantic roles as input

7.3.2. Condition 2: Projection from Automatic Roles

In this condition, as in the corresponding English–German condition described in Section 7.2.2, we use noisy English role annotations from the output of a shallow semantic parser as input to the projection process. As we have argued above, this condition is important to gauge the feasibility of role projection for practical (i.e., large-scale) scenarios, since usually no manual annotation will be available for large parallel corpora even in the source language.

We re-used the automatic annotation for English produced for the first experiment with Giuglea and Moschitti's (2004) shallow semantic parser as input for projection. Recall that the shallow semantic parser's performance (65.1% F-Score) can be seen as an upper bound for projection quality. In parallel to Experiment 1, we focus on the projection model which we found to perform best on manual input data in Condition 1, namely PerfMatch with NA Filter.[10].

The results are shown in Table 7.11. Again, we contrast the performance of the constituent-based model to the word-based baseline and to the performance of the shallow semantic parser as an upper bound. In comparison to the results for Condition 1 on the test set (see Table 7.10), the quality of the projection decreases somewhat, but the effect is surprisingly small: Overall performance both for the word-based baseline and the PerfMatch model drops by 5 to 6% F-Score. Furthermore, we find this effect results entirely from a decline in recall. The precision even increases slightly when automatically annotated role information is projected.

[10] Again, further analysis showed that this setting outperforms all other models.

Band (Intersective)	Prec	Rec	F-Score
Error 0	80.7	73.6	77.0
Error 1	73.8	38.3	50.4
Error 2+	39.8	20.5	27.1

Table 7.12.: Experiment 2, Condition 2: PerfMatch's performance by errors rate in automatic semantic role labelling per frame (Error 0: no labelling errors, Error 1: one labelling error, Error 2+: two or more labelling errors)

A comparison of these results for French with the corresponding figures for German (see the top half of Table 7.6) shows that the overall quality is very similar for both languages, with even a slight advantage for French (57.2% over 56.6% F-Score). French shows a slightly higher recall, but lower precision. We attribute this effect to the less sparse word alignment (cf. Condition 1), which supports the projection of more (both correct and erroneous) role information.

The find the same, very similar picture across languages in the more detailed analysis of projection quality in relation on annotation errors, which is shown for French in Table 7.12. The overall stratification is very similar to German (cf. Table 7.7). The projection results are very clean for frames with no labelling errors. With one error, precision is still above average, while recall is below. For two labelling errors, both recall and precision degrade clearly. The smaller amount of variance between bands for French – the Error 0 band is 3% F-Score worse than for German, while the Error 2+ band is 3% F-Score better – can again be attributed to word alignment differences.

In sum, the result for Condition 2 cast a more favourable light on French than the results for Condition 1 (listed in Tables 7.4 and 7.10), where results for German surpassed results for French by a substantial margin. The fact that these differences are confined to the "clean" projection of manually annotated data suggests that they reflect predominantly differences in the manual annotation process. Recall that the inter-annotator-agreement was 83% F-Score for German, but only 72% for French. The lower agreement for French, which can in turn be traced back to the lower agreement be-

tween syntactic and semantic bracketing, makes projection considerably harder, as witnessed by the lower upper bound. In contrast, we find virtually no difference in Condition 2, where projection quality is currently limited primarily by the performance of the shallow semantic parser (65% F-Score). In this situation, the noise in the source data outweighs the differences between the target annotations.

Thus, the experimental results we obtained for Condition 2 correspond well to the predictions from our study on instance-level parallelism in Chapter 3, where we found that German and French role annotations correspond to English approximately equally well. In fact, the differences we encounter between the two target languages are attributable either to differences in the syntactic analysis (overall lower performance for French in Condition 1), or to differences in the statistical word alignment (higher recall and lower precision for French). At the current state of the art in shallow semantic parsing, the performance penalty for French is even masked completely by errors in the automatic source annotation. However, we would expect the effect to become measurable once the performance of shallow semantic parsing attains the same level as the inter-annotator agreement in the target language.

7.4. Related Work

Cross-lingual induction of semantic roles. The study in the literature closest to our work is Johansson and Nugues (2006), who induce frame-semantic role annotations for Swedish also using annotation projection. However, their study relies solely on word alignment for projection, comparable to our word-based projection models. The shortcomings of these alignments are addressed by complementing them with a number of specific heuristics for the language pair English–Swedish. Johansson and Nugues do not evaluate their projected data directly, but concentrate on using them for the induction of a shallow semantic parser for Swedish. Their method yields good results; however, this may be attributable at least partly to the very close relationship between Swedish and English, verified empirically in Koehn (2005). Unfortunately, the identification of heuristics to correct word alignment has to be repeated for every new target languages, and presumably becomes much harder for less closely

related languages.

A completely different strategy was presented by Fung and Chen (2004), who proposed an ontology-based method to induce FrameNet-style roles for Chinese. In a first step, they map English FrameNet entries to concepts listed in HowNet[11], an on-line ontology for Chinese, without making use of parallel texts. In a second step, they search for monolingual Chinese sentences containing predicates instantiating these concepts, and label their arguments with FrameNet roles. While Fung and Chen report high accuracy values, their method relies on the existence of an ontology for the target language whose design decisions do not clash with the frame distinctions made by FrameNet (Ellsworth, Erk, Kingsbury, and Padó, 2004; Ruppenhofer et al., 2005). Such resources are presumably rare, and compliance with the FrameNet principles is difficult to assess.

A third approach was recently proposed by Pitel (2006). Using parallel bilingual corpora, he constructs a bilingual vector space containing content words of both the source (English) and target languages. Within this space, he locates clusters of English lemmas corresponding to the head words for some frame element in the FrameNet example annotations. Target language lemmas within the same cluster are presumed to be typical head words of the frame element in the target language, and can subsequently be used to identify instances of this FE in target language text. Unfortunately, this purely lexical approach on its own appears to be insufficiently discriminative. The three major problems are (a), frame elements which cannot be represented by a semantically coherent class of words (such as MESSAGE, which can have almost any head word), (b), pairs of frame elements with very similar semantic representations (such as BUYER and SELLER), and (c), role instances in the target language filled by semantically light material such as anaphors or pronouns (*this, we*). Nevertheless, the lexical information provided by the bilingual vector space might provide an interesting complement to the more structural projection model presented in this thesis.

Unsupervised induction of semantic roles. An alternative strategy for inducing semantic role information "from scratch" is provided by unsupervised induction methods, which do not require annotated corpora.

[11] See `http://www.keenage.com/zhiwang/e_zhiwang.html`.

A small number of studies develop such methods, but all of them are forced to constrain the roleset they consider to make the induction feasible. For example, Swier and Stevenson (2004, 2005) train a shallow semantic parser for 16 universal semantic roles. They initialise the parser by extracting syntax-semantics mapping rules from the VerbNet (Kingsbury and Kipper, 2003) verb lexicon, and refine it iteratively. They first assign roles to all unambiguous role instances in a parsed, but semantically unannotated, corpus, and then use these instances to estimate a simple probability model which serves to label the ambiguous cases. Unfortunately, the initial reliance of the approach on a large verb lexicon makes its application to low-density languages difficult; in addition, generalisation to larger sets of semantic roles is likely to result in sparse data problems.

A completely unsupervised method was recently proposed by Grenager and Manning (2006), who induce semantic role annotation using the Expectation Maximisation algorithm. The model is able to label syntactically annotated corpora with six indexed semantic roles, and does so with a high accuracy. However, the role annotations carry only limited information: the role labels do not provide any semantic interpretation (they merely serve to distinguish the roles), and do not generalise across verbs. Thus, the semantic roles provided by the model are more accurately characterised as generalised grammatical functions that abstract over diathesis alternations. While generalisation over surface phenomena is a worthwhile task, this representation does not currently offer a semantic characterisation of the predicate-argument structure as usually provided by semantic roles (cf. Section 1.1).

Semantic alignment. On the modelling level, we have proposed a framework for computing semantic alignments which express translational equivalence between constituents of bi-sentences. Semantic alignments are formalised as bipartite graphs and the search for the best alignment as an optimisation problem. The view of alignment as graph matching is relatively widespread in the machine translation literature on word alignment (Melamed, 2000; Matusov, Zens, and Ney, 2004; Tiedemann, 2003a; Taskar, Lacoste-Julien, and Klein, 2005). Despite individual differences, most approaches cast word alignment as a maximum-weight matching problem (cf. Section 6.3), where each pair of words in a bi-sentence is associated with a score representing the desirability of that

pair. The alignment for the bi-sentence is the highest scoring matching under some constraints, for example that matchings must be one-to-one. Our work applies graph matching to the level of constituents and considers a larger class of constraints (see Section 6.3.2) than previous approaches. For example, Taskar et al. (2005) examine solely perfect matchings and Matusov et al. (2004) only edge covers.

The task of aligning constituents of bi-sentences has also been addressed by a number of studies investigating the extraction of phrase-level translation patterns from parallel treebanks. However, most of these models only search for pairs of constituents which are perfectly word-aligned to one another, a strategy that is infeasible for automatically obtained word alignments (Kaji, Kida, and Morimoto, 1992; Imamura, 2001). Other papers address word alignment errors by couching constituent alignment as an optimisation problem, but use greedy search techniques which are not guaranteed to find an optimal solution (Matsumoto, Ishimoto, and Utsuro, 1993; Yamamoto and Matsumoto, 2000). Meyers, Yangarber, and Grishman (1996) impose structural constraints on the alignment which exclude non-isomorphic parts of the source and target trees from the alignment.[12] While this constraint gives the alignment a very strong corrective power (cf. Section 6.3.2), it appears to be most appropriate for similar languages, similar parsing schemes, and high-quality parse trees. When these conditions are not met, it is liable to align only small parts of the two trees. In contrast, the framework we have proposed in this thesis does not explicitly enforce the preservation of dominance relations. We nevertheless obtain high precision alignments, which indicates that such constraints are not strictly necessary in practice.

In this thesis, we have evaluated the semantic alignment models on the semantic role projection task, but we believe that semantic alignments show promise in the context of statistical machine translation as well, especially for systems that use syntactic information to enhance translation quality. For example, Xia and McCord (2004) exploit constituent alignment for rearranging sentences in the source language so as to make their word order similar to that of the target language. They learn tree reordering rules by aligning constituents heuristically with a heuristic

[12] Formally, they require that for each pair of alignment links (c, c') and (d, d'), the alignment maps the lowest common ancestor of c and d onto the lowest common ancestor of c' and d'.

local optimisation procedure similar to our "Total" admissible alignment class. A similar approach is described in Collins, Koehn, and Kučerová (2005); however, there, the rules are manually specified and the constituent alignment step reduces to inspection of the source-target sentence pairs. The different admissible alignment models presented in this thesis could be easily employed for the reordering task common to both approaches. In other work, the rewrite rules are not used as a preprocessing step (e.g., to reorder source strings) but form part of the translation model itself (Gildea, 2003, 2004). Constituent alignments are learnt by estimating the probability of tree transformations, such as node deletions, insertions, and reorderings. These models have a greater expressive power than the models presented here; however, this implies that approximations have to be used to keep the computation feasible.

7.5. General Discussion

In Part III of this thesis, we have presented a general framework for the cross-lingual projection of semantic roles. The framework centers around the notion of semantic alignment which can be parameterised for different units (linguistic entities) of projection. We discussed two broad instantiations of the framework, namely word-based and constituent-based models. We showed that the search for semantic alignments can be formalised as an optimisation problem. Specifically, bi-sentences are conceptualised as bipartite graphs where optimal alignments correspond to a minimum-weight subgraph meeting certain requirements. Solutions to the search problem can be obtained efficiently with well-known graph optimisation methods. In addition, we have introduced a small number of filtering techniques which either repair or eliminate alignment errors. We have evaluated this framework on two different language pairs, namely English–German and English–French.

The proposed framework is both intuitive and general. We can explore a wide variety of models by imposing constraints on the shape of the alignments. This results in models which differ in terms of their expressive and corrective power. Restrictive models can better overcome errors in the word alignment by guiding the search towards linguistically reasonable solutions, but unfortunately cannot account for translational divergences

LingUnit	Similarity	AlignModel	ExpPower	CorPower	Complex
words	binary	WordAlign	(x)	(x)	$O(n)$
constit.	overlap	Total	high	low	$O(n^2)$
constit.	overlap	EdgeCover	medium	medium	$O(n^3)$
constit.	overlap	PerfMatch	low	high	$O(n^3)$

Table 7.13.: Framework instantiations ((x): depending on word alignment)

such as one-to-many or many-to-many alignments. We have derived four models that exemplify our projection framework. The models are summarised in Table 7.13 and vary along the following dimensions: the linguistic units employed (LingUnit), the similarity measure (Similarity), the set of admissible alignments (AlignModel), their expressive (ExpPower) and corrective power (CorPower). We also mention the time complexity of their computation (Complex) in the number of units in the bi-sentence.

We found that the use of constituent information obtained from the output of a full parser yields substantial improvements over word alignments or chunks. Although word-based models could offer a starting point for low-density languages for which parsers are not available, the noisy and often fragmentary nature of word alignments makes it difficult to deliver accurate projections. Chunk-based models perform generally better than word-based ones; however, they also tend to produce incomplete role spans. Our experiments also compared and contrasted three types of semantic alignment models differing in their corrective and expressive power (total, edge covers, and perfect matchings). We found that the perfect matching alignment, the most restrictive model which enforces one-to-one alignments, delivers the highest performance reliably for both language pairs. This suggests that it is more beneficial to choose a more restrictive alignment model that ignores translational divergence, but has the power to correct wrong word alignments. As a side effect, the performance of constituent-based models increases only slightly when manual word alignments are used, which means that role projection yields near-optimal results with automatic word alignments.

As far as different elimination techniques for alignment noise are concerned, we find that removing non-aligned words yields the best results

most consistently across different experimental conditions and target languages, in particular in combination with the PerfMatch class of admissible alignments. This filter is independent of the language pair or the underlying linguistic representations in question and could be potentially useful for projection models beyond the ones we have considered.

Finally, experiments with an shallow semantic parser (Giuglea and Moschitti, 2004) demonstrate that the projection approach delivers satisfactory results even for noisy input, notably preserving high precision in the induced role annotation. Thus, our framework already at present holds promise for large-scale resource creation. Future improvements in shallow semantic parsing have the potential to lead to still better projection results.

Comparing the results for German and French, we find material agreement between the two target languages. Performance is somewhat superior for German when manual role annotation is used as input. However, this difference is almost entirely attributable to technical factors, notably the recall of the statistical word alignment, and the agreement between syntactic bracketing and role spans. The differences furthermore cease to be noticeable when noisy input data is used, and performance becomes virtually identical for both target languages. This corresponds well to the estimates we obtained for instance-level semantic parallelism in Chapter 3, lending further support to the predictive value of the degree of instance-level parallelism for the performance of actual projection.

The projection framework developed here is useful for other semantic role paradigms besides FrameNet, or indeed for other annotations of semantic nature. Potential applications include the projection of PropBank roles (provided the grounding problem discussed in Section 3.2.3 can be solved), discourse structure, or named entities. As mentioned in Section 7.4, the alignment models discussed here could be used in machine translation for the reordering of constituents. An important direction for future work lies in investigating the use of projected annotations for training semantic parsers for the target language (Johansson and Nugues, 2006).

Another important avenue of future work concerns the role of bracketing information in the target language. We assume that accurate projection of a semantic role is considerably simplified when the syntactic analysis for the target sentence contains a single constituent corresponding to the

"ideal" role span. This hypothesis is supported by two observations: first, the large improvement for using recursive constituents (which usually contain ideal spans) over chunks results (which do not); and second, the difference in Condition 1 between German (whose full syntactic analysis almost always offers such constituents) and French (where this is true in a smaller portion of cases). In contrast, the global shape of the parse tree (such as the depth of the analysis) appears to play a negligible role, since filtering has the potential of making trees more similar. Unfortunately, broad-coverage full parsers are not yet available for many potential target languages. We will therefore explore alternative ways of inducing potential role spans for target sentences. Since we do not require hierarchical structure, it might be sufficient to construct pseudo-constituents, either by combining chunks (which are available for a larger number of languages), or by using completely unsupervised aligned-based learning techniques (Geertzen, 2003).

A final option which we have not considered in this thesis is the use of the complete corpus as context to improve alignment. Our current framework optimises the constituent alignment for each sentence individually, and requires the class of admissible alignments to be chosen by the user. While the class of perfect alignments appears to be a robust choice, it has not been optimal for every experimental condition. A strategy which might overcome both of these limitations at the same time is to approach constituent alignment with fertility-based alignment models such as the IBM word alignment models 3–5 (Brown et al., 1993). These models treat the number of alignment links for each source item as a hidden variable, whose value can be estimated together with the other parameters of the model from the corpus itself (see Section 2.2.1 for details). In our case, such a fertility-based model would determine alignments that optimally explain the patterns found in the data. The two central problems which would have to be solved for the application of such a scheme are (a), that word alignment models are designed for aligning (linear) sequences, while constituents are arranged in trees; and (b), that a suitable representation for individual constituents has to be found. Such a representation has to provide enough information to distinguish between constituents (which rules out the use of mere phrase types), but has to allow for the identification of similar patterns across sentences (which precludes the use of complete yields).

7.6. Summary

In this chapter, we have performed an experimental evaluation of the framework for the projection of semantic roles which we have developed in Chapter 6. We have used the trilingual parallel sample corpus whose construction was described in Section 3.3 to induce frame-semantic role information for two target languages, namely German and French.

The results of the experiments have verified the fundamental feasibility of role projection through annotation projection both for German (Section 7.2) and French (Section 7.3), and the main insights hold equally for both languages. First, role projection profits substantially from accurate bracketing information in the target language to recognise linguistically sensible word spans onto which roles can be projected. Results attain between 64% and 70% F-Score. Second, high-quality role projection is feasible even under comparatively noisy conditions, such as when using automatically constructed word alignment, and using input roles from a shallow semantic parser. Projection is notably able to maintain a high precision of over 70%, with F-Scores at around 57%. This is a promising result in particular with respect to resource induction for target languages, since arguably precision is more central for this purpose than recall.

Part IV.

Further Directions and Conclusions

8. Beyond Frame Instance Parallelism: Frame Group Paraphrases

8.1. Motivation

The overall aim of this thesis is to develop methods for the cross-lingual induction of frame-semantic information to create frame and role annotation for new languages. Part III has presented a projection-based models for the transfer of semantic role information. However, as we have established in Section 3.2.2, simple role projection is only possible when frame instance parallelism holds, i.e., when the predicate in the source language evokes the same frame as its aligned translation. The underlying reason is that frame-semantic roles are defined for individual frames (cf. Section 1.1.1). As a result, the methods presented in Part III are not applicable to all bi-sentences in a parallel corpus, or rather, run the risk of producing wrong annotations when applied indiscriminately to all parallel sentences. This limits the applicability of our methods to structurally similar languages with a large number of exactly parallel semantic structures (see Chapter 3 for discussion).

In consequence, this chapter is concerned with analysing and modelling cases where frame instances are not parallel. It addresses the following three questions:

- Are there patterns in frame instance non-parallelism?
- Can we obtain these patterns from a corpus automatically?
- Can we exploit these patterns for role projection?

We will argue that the answer to the above questions is yes. We provide a definition of such patterns, which represent translational equivalences

between frames or frame combinations and mappings between corresponding roles. We call these patterns *frame group paraphrases*, and they are based on *frame groups*, sets of adjacent frames. We provide an algorithm which acquires frame groups from a parallel corpus (Section 8.3), and evaluate it on a small subset of EUROPARL (Section 8.4).

In this study, we lay the foundations for identifying and modelling frame group paraphrases. The methods described here, although developed to be fully automatisable, have so far only been applied and evaluated manually, which results in a very limited evaluation scenario. In Section 8.5, we sketch how frame group paraphrases can be coupled with projection by integrating them as two steps in a general bootstrapping process. We conclude with a review of related work (Section 8.6) and a general discussion, with a focus on the generalisability of our approach (Section 8.7).

8.2. Frame Group Paraphrases

8.2.1. An Example of Frame Non-Parallelism

Recall from Section 1.1.1 that the definition of frames has both a conceptual and a structural component. Slightly simplified, the structural criterion for the inclusion of a predicate in a frame is that its valency allows it to realise all semantic roles belonging to that frame. Thus, it is evident that translation pairs which differ substantially in their valency cannot be analysed using the same frame. In the following example, an English transitive verb (*increase*) is translated as an adjective, *höher (higher)*, which leads to frame non-parallelism:

(8.1) [The drought]$_{\text{Cause}}$ has **increased** [tea prices]$_{\text{Item}}$. (CCPOS)

Die Dürre hat zu **höheren** [Teepreisen]$_{\text{Item}}$ geführt. (CPOS)
The drought has to **higher** [tea prices]$_{\text{Item}}$ led.

The problem is that the transitive verb *increase* has the ability to express both an ITEM, whose value changes, and the CAUSE of this change. It therefore has to be analysed as an instance of a causative frame, CAUSE_-CHANGE_OF_POSITION_ON_A_SCALE (CCPOS). Its German translation, as an adjective, can only only realise the ITEM role due to its restricted

Frame name	Abbrev.	Definition
Cause change of position on a scale	Ccpos	A Cause affects the position of an Item on some scale.
Change of position on a scale	Cpos	An Item's position on a scale changes.
Causation	–	A Cause causes an Effect.

Table 8.1.: Most important FrameNet frames used in Chapter 8 and their definitions (names of semantic roles for each frame printed in small caps)

valency. Thus, it is analysed as evoking the stative counterpart of Ccpos, the Change_of_position_on_a_scale frame (abbreviated Cpos).[1]

With the German predicate only able to realise the Item role, the question arises what happens to the second English role, Cause. Of course, one possibility is that the translation is "lossy" and that the role is not translated at all. Example 3.2, repeated here for convenience, is in fact such a case. Here, the German Cognizer role does not occur in English due to the choice of the passive voice:

(8.2) So I ask that [Ireland]$_{\text{Content}}$ be **remembered** in this particular case.

Ich möchte deshalb darum bitten, dass [man]$_{\text{Cognizer}}$ in diesem
I would like therefore to ask that [one]$_{\text{Cognizer}}$ in this
speziellen Fall auch [an Irland]$_{\text{Content}}$ **denkt**.
particular case also [of Ireland]$_{\text{Content}}$ **thinks**.

This is, however, not what happens in Example 8.1. Figure 8.1, which shows a complete frame-semantic reference analysis for the example, reveals that the German translation introduces an additional predicate, *führen (to lead)*, with its own frame, Causation, and two roles Cause and Effect. It is this new frame which provides the missing role: The German Causation.Cause role[2] corresponds exactly to the English Ccpos.Cause

[1] The definitions of the most important frames used in this chapter are provided in Table 8.1.

[2] In the following, we will use the notation Frame.Role when roles of different frames with the same name need to be distinguished.

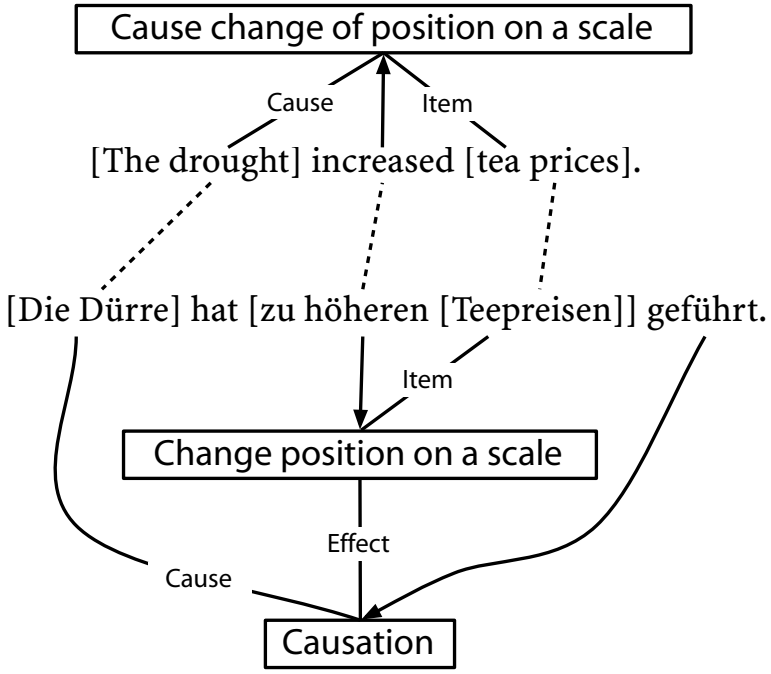


Figure 8.1.: Analysis of the bi-sentence from Example (8.1): Non-parallelism on the level of individual frames

which cannot be realised within the German Cpos frame. Interestingly, the second role provided by Causation, Causation.Effect, is filled by the phrase *zu höheren Teepreisen*, whose head modifier, *höher*, evokes the Cpos frame. This link is an additional indicator for a close relationship between the two German frames Causation and Cpos.

In Chapter 3, we have introduced instance-level parallelism of frames and roles as prerequisite for frame-semantic annotation projection. From this perspective, the bi-sentence in Figure 8.1 poses a real challenge. We find neither frame-instance parallelism (the frames do not correspond), nor role-instance parallelism (which is conditional on frame-instance parallelism). In consequence, the methods developed in Parts II and III are not applicable to this sentence pair. Nevertheless, we have the clear intuition that these two sentences are parallel. They describe the same state of affairs, and they mention the same participants, despite the fact that they do not use the same frames.

Non-parallelism on the frame level is actually a moderately frequent phenomenon, as we have established quantitatively through the evaluation of our manually annotated sample corpus in Section 3.3, where we found around 30% frame instance non-parallelism for both language pairs English–German and English–French. As we have discussed in the analysis of translational shifts (Section 5.7), non-parallelism arises very naturally from the fact that translation does not necessarily proceed word by word; the resulting restructuring of the sentence often breaks instance-level parallelism by changing frames or *re-distributing predications among adjacent frames*. This is exactly what happens in Figure 8.1, where the CAUSE role is once expressed by the CCPOS frame, and by the CAUSATION frame. Not being able to model such cases of non-parallelism therefore limits the coverage of frame-based projection methods on parallel corpora.

8.2.2. Frame Groups and Frame Group Paraphrases

We propose to address the problem of frame non-parallelism by generalising the notion of frame-instance parallelism (cf. Section 3.2.3), which underlies role projection, from individual frames to so-called *frame groups*. A frame group is a set of adjacent frames. The assumption that motivates this step is the intuition that the "redistribution of predications" such as the movement of the CAUSE role from the CCPOS to the CAUSATION role in Figure 8.1 is a largely local phenomenon, which can be modelled in terms of the local semantic context. In this chapter, we consider in particular frame groups of size two, which represent the simplest possible form of frame groups that is more powerful than individual frames.[3]

Our hope is that two-frame frame groups will provide a level of description on which parallelism takes places significantly more often than for individual frames. This idea is illustrated in Figure 8.2: We find two frame groups in the two halves of the bi-sentences: in English, a singleton frame group {CCPOS}, and in German, the frame group {CPOS,CAUSATION}. Our intuition is that these frame groups correspond to one another, effectively forming a *frame group paraphrase*: they describe the same state of affairs, i.e., translationally equivalent, and thus *parallel*, chunks of predicate-

[3]This assumption will be discussed in Section 8.7.

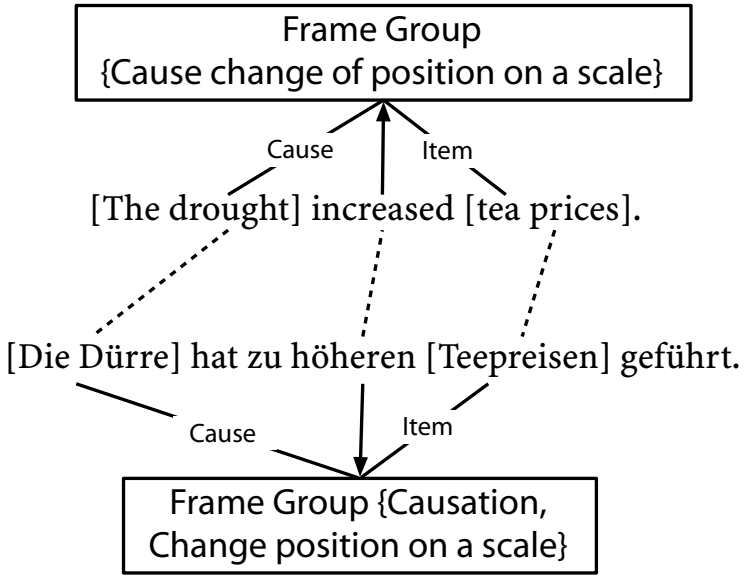


Figure 8.2.: Analysis of the bi-sentence from Example (8.1): Parallelism on the level of frame groups

argument structure. A frame group paraphrase can also be thought of as two frame groups that can be used alternatively.

We now make the concepts of frame groups and frame group paraphrases more precise. First, we consider the definition of frame groups. Clearly, allowing every pair of frame instances in a sentence to form a frame group is too lenient and would result in overgeneration, i.e., the emergence of frame groups which cannot model parallelism in bi-sentences. As mentioned above, our intuition is that frame groups are aimed at capturing *local* redistribution of semantic material; therefore, we require frames in a frame group to be *adjacent*, with no intervening frames.

Frame Group. We define a frame group to be a set of one or two frames. It contains one frames called the *base frame*. If it contains two frames, the frame that is not the base frame is called the *embedding frame*. The embedding frame must have one designated role, whose semantic head[4] is either the FEE of the base frame, or has the FEE

[4]Syntactic and semantic head can differ in the case of transparent nouns: In "he drank a pint of milk", the syntactic head is "pint", while the semantic head is "milk".

of the base frame as a modifier. This role is called the *embedding role*; all other roles of the frame group are called *free roles*.

This definition can be illustrated in Figure 8.1: CCPOS is the base frame, and CAUSATION is the embedding frame. The role CAUSATION.EFFECT is the embedding role: Its span is the phrase *zu höheren Teepreisen*. The embedding role thus "glues together" the two individual frames which act as partial descriptions of the complete state of affairs. For the sake of clarity, our figures depict the embedding role as pointing to the base frame instead of its span.

Note that our present definition of embedding still depends on the existence of syntactic relations between the embedding role and the FEE of the base frame. A more general characterisation is currently emerging in the form of *frame dependency graphs*, dependency graphs which contain non-lexical nodes labelled with frames intervening between predicates and their arguments. Frame dependency graphs have been adopted as the desired output format at the SEMEVAL 2007 task on frame-semantic structure extraction (Baker, Ellsworth, and Erk, 2006).

We next describe frame group paraphrases, pairs of frame groups which describe the same state of affairs in two languages. Our intuition is that such pairs of frame groups need to *express predications of the same entities*; informally, they must "talk about the same things". We also introduce the possibility of restricting the correspondence to a set of *relevant* participants, which makes it possible to derive "lossy" paraphrases as well which preserve only a subset of the original participants.

Frame Group Paraphrase. We define a frame group paraphrase as a pair of frame groups (f, f') on both sides of an aligned bi-sentence, whose base frames are evoked by one source and one target predicate which are translationally equivalent. Furthermore, given a set of *relevant roles*, which form a subset of the free roles of f, it must hold that the translational equivalent of each relevant role of f is labelled by a free role of f', and vice versa. If all free roles of f are relevant, we call the frame group paraphrase *full*, otherwise we call it *lossy*.

According to this definition, and assuming that word alignment can be used as an indicator of translational equivalence, the example in Figure 8.1

shows indeed a full frame group paraphrase. The two base frames, CCPOS and CPOS, are evoked by the predicate pair *increased/höher*, and the free (and thus relevant) roles of CCPOS are mapped as follows: CCPOS.ITEM to CPOS.ITEM, and CCPOS.CAUSE to CAUSATION.CAUSE.

The most immediate application of frame group paraphrases in the context of this thesis is that they allow us to state a generalised definition of instance-level parallelism:

Instance Parallelism for Frame Group Paraphrases. Instance parallelism for frame group paraphrases holds if for all pairs of source and target predicates that are translationally equivalent in a parallel corpus, these predicates evoke a **known frame group paraphrase**.

This definition subsumes the "classical" cases of frame instance parallelism as defined in Section 3.2.2, since individual frames are also frame groups, but its coverage is greater, encompassing also frame groups of size 2 like the case in Figure 8.2. In parallel to classical frame instance parallelism, it implies role instance parallelism. This property is guaranteed by the definition of frame group paraphrases, which requires that the free roles of the two frame groups involved correspond to one another. In the rest of this chapter, we will abbreviate the term of instance parallelism of frame group paraphrases as FGP instance parallelism.

What is still left to be shown, of course, is that this defining instance-level parallelism on the level of frame groups represents an improvement over the old definition on the level of individual frames in Section 3.2.2. This argumentation involves several steps:

- First, note that the definition above mentions "known frame group paraphrases"; therefore, we need to develop a technique to automatically acquire frame group paraphrases from annotated bi-sentences. An algorithm for this task is presented in Section 8.3.

- Next, Section 8.4 provides an evaluation of the degree of FGP instance parallelism, as Section 3.3 did for frame instance parallelism. We assert that frame group paraphrases can indeed improve the coverage over frame instance parallelism (Section 8.4.3). Importantly, Section 8.4.4 provides a complementary qualitative evaluation which verifies that the acquired frame group paraphrases are actually paraphrases, i.e., describe equivalent states of affairs.

- In Section 8.5, we address the main limitation of our acquisition algorithm, namely that it requires frame-semantic analyses for both languages to acquire frame group paraphrases. We sketch how this problem can be solved by interleaving frame group acquisition with annotation projection, using the methods developed in Parts II and III.

8.3. Corpus-based Acquisition of Frame Group Paraphrases

As mentioned in the last section, the new definition of instance-level parallelism requires a set of known frame group paraphrases, i.e. frame groups that can be used interchangeably. In an ideal world, frame group paraphrases would be included in the lexical information encoded in FrameNet. However, due to the high effort involved in manually identifying and encoding frame group paraphrases, only preliminary steps exist in this direction. FrameNet provides some frame group paraphrases for singleton frames through the frame hierarchy (such as the RIDE VEHICLE – OPERATE VEHICLE example discussed in Section 3.1.1), but currently does not include information on frame groups with more than one frame.

A promising strategy for obtaining frame groups is their automatic acquisition from a parallel corpus, such as EUROPARL, when it is annotated with FrameNet frames and roles on both sides. The word alignment in bi-sentences can be adopted as an approximation of translational equivalence.[5] This section presents an algorithm for this task, which produces a set of frame group paraphrases for a given *target frame group*.

8.3.1. The Iterative Matching Algorithm

Notation. To describe the algorithm concisely, we will need to introduce some notation. First, we will need to distinguish frames and frame groups (types) from their instances in a given sentence. Given a frame (group) instance f with roles $r_1, r_2, \ldots$, we will write $\bar{f}$ for the frame (group) type

[5] Alignment and preprocessing can be automatised as seen in Chapter 2. As for obtaining the frame-semantic analysis for both sides, see Section 8.5.

with the role types $\overline{r_1}$,$\overline{r_2}$ that it instantiates. Two frame groups are called *aligned* if the FEEs of their base frames are word-aligned.

We describe correspondences between the roles of two frame groups by their role mappings: Writing roles($\overline{f}$) for the set of free roles of a frame group $\overline{f}$, a role mapping m : roles($\overline{f_1}$) $\rightharpoonup$ roles($\overline{f_2}$) is a partial mapping from roles of a frame group $\overline{f_1}$ to roles of a frame group $\overline{f_2}$. Given two aligned frame group instances f_1, f_2, the *word alignment-induced* role mapping m : roles($\overline{f_1}$) $\rightharpoonup$ roles($\overline{f_2}$) is the role mapping that maps each role $\overline{r_1}$ to a role $\overline{r_2}$ if and only if r_1 and r_2 are aligned according to the word alignment-based methods developed in Chapter 6.

The Algorithm. The algorithm is shown in Figure 8.3. It has two parameters. The first is the **target frame group** $\overline{f_t}$, the frame group for which frame group paraphrases are to be acquired, and the set of relevant roles which have to be filled in all paraphrases. The second is a set of *base frames* that have already been identified as partial paraphrases of the target frame groups, which do not fill all relevant roles of the target frame group, such as Cpos for Ccpos. Base frames thus introduce the necessary knowledge to focus paraphrase acquisition on *plausible* frame groups. However, if nothing is known about base frames for the target frame group, a simplified version of the algorithm can also be produced by removing all checks for consistency with the base frames (see below). Note that in a real-life corpus this is likely to result in a noisier set of paraphrases.

The algorithm iteratively extends a set P of known frame group paraphrases for the target frame group $\overline{f_t}$, along with the role mappings which link paraphrase roles to target roles. It proceeds by browsing the corpus and identifying instances of aligned pairs of frame groups (f_1,f_2) where f_1 is already a known paraphrase of $\overline{f_t}$, but f_2 is not. Provided that the three conditions enumerated immediately below hold, the frame group $\overline{f_2}$ is added to the set of paraphrases for $\overline{f_t}$. The role mapping between $\overline{f_t}$ and $\overline{f_2}$ is constructed by combining the (known) role mapping between $\overline{f_t}$ and $\overline{f_1}$ with the word alignment-induced role mapping between f_1 and f_2.

1. The base frame of f_2 is based on one of the partial paraphrases for the target group. (Line 6)

1: Given: target frame group $\overline{f_t}$ and a set rel $\subset$ roles($\overline{f_t}$) of relevant free roles
2: Given: a set B of base frames with role mappings m_t
3: Set the set of paraphrases $P = \{\overline{f_t}\}$; $m_t(\overline{f_t}) = \{(\overline{r}, \overline{r}) \mid \overline{r} \in \text{rel}$
4: **while** P changes **do**
5: **for** aligned frame group instances f_1,f_2 **do**
6: **if** $\overline{f_1} \in P$ and base($\overline{f_2}$) $\in B$ **then**
7: m is the word alignment-induced role mapping m from f_1 to f_2.
8: **if** $range(m_t(\overline{f_1})) \subseteq dom(m)$ and $m_t(\overline{f_1}) \circ m$ and $m_t(\text{base}(\overline{f_2}))$ coincide on the intersection of their domains **then**
9: $P = P \cup \{\overline{f_2}\}$, where $m_t(\overline{f_2}) = m(\overline{f_1}) \circ m$
10: **end if**
11: **end if**
12: **end for**
13: **end while**

Figure 8.3.: Iterative matching for frame paraphrase acquisition

2. f_2 must realise all the roles that f_1 does, i.e., all relevant roles. (Line 8)

3. The role mapping between $\overline{f_t}$ and $\overline{f_2}$ is consistent with the known role mapping for the partial paraphrase $\overline{\text{base}(f_2)}$. (Line 8a)

The first condition ensures that $\overline{f_2}$ is a plausible paraphrase for $\overline{f_t}$. The second condition is necessary to guarantee that all relevant roles of the target frame group are found. The third condition rules out inconsistent role mappings. Conditions (1) and (3) can be dropped if no base frames are known.

A worked example. Figures 8.4 and 8.5 show a small bilingual corpus with three sentences on which we will demonstrate how frame group

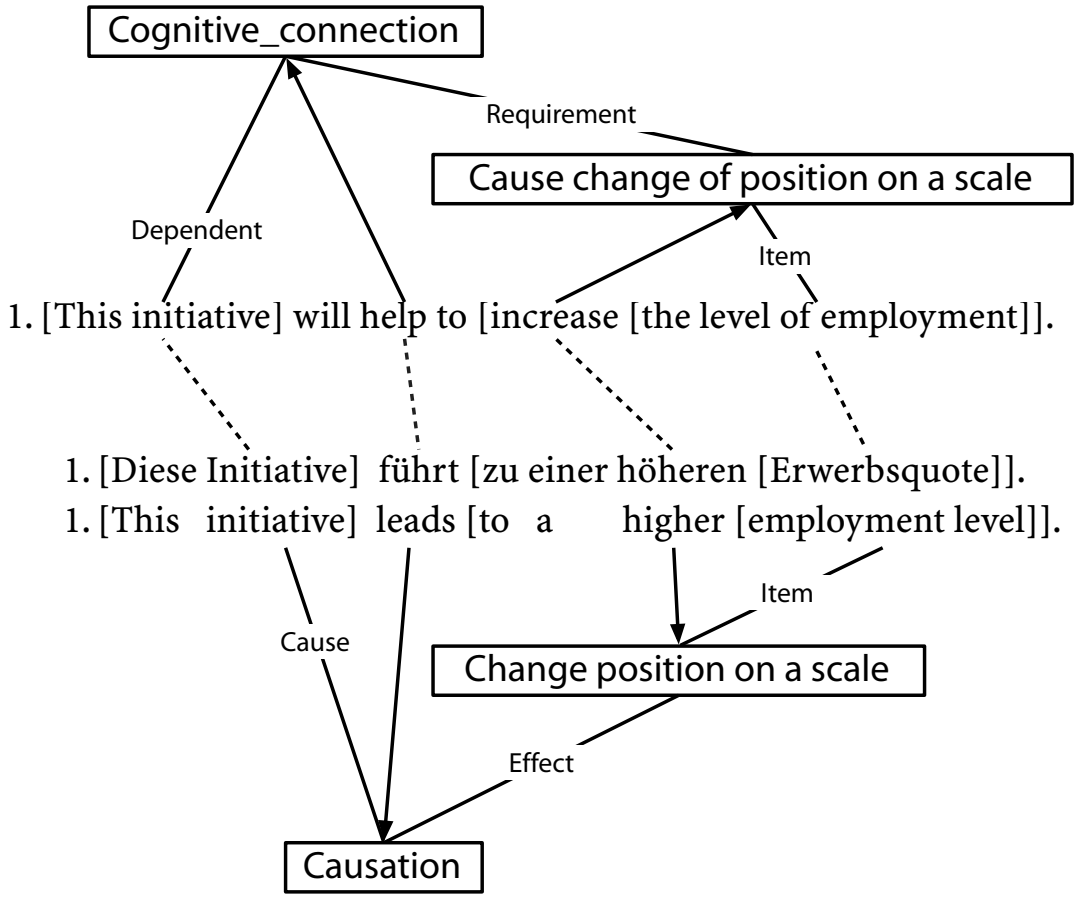


Figure 8.4.: Sentence 1 of example corpus to illustrate the iterative matching algorithm

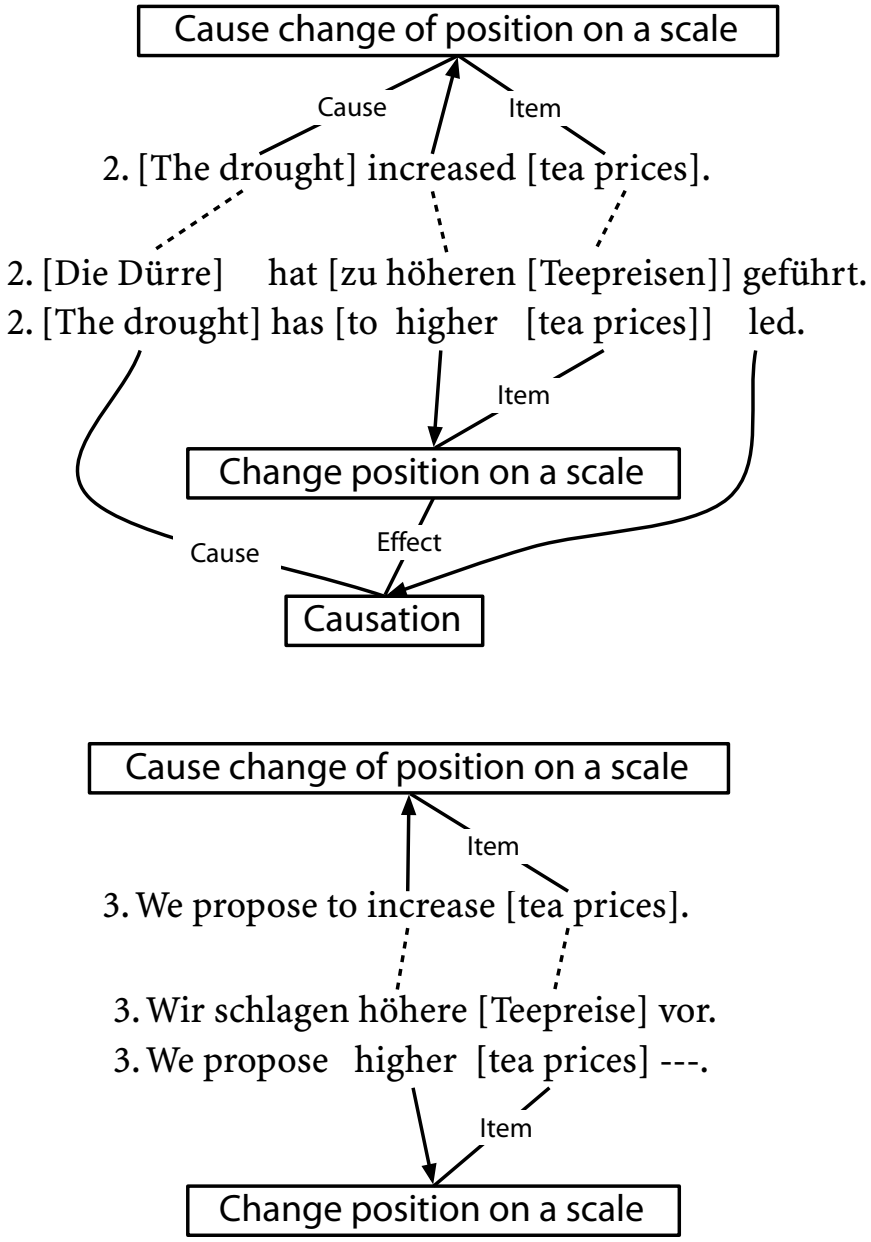


Figure 8.5.: Sentences 2 and 3 of example corpus to illustrate the iterative matching algorithm

paraphrases are acquired iteratively with our algorithm. We initialise the algorithm with the singleton group {CCPOS} as target frame group. We look for full paraphrases, using {CAUSE,ITEM} as relevant roles. Since the target frame group is a singleton, there is no embedding frame. The only partial paraphrase we initially know is the base frame CPOS, with the role mapping {CCPOS.ITEM → CPOS.ITEM}. This captures the intuition that causative change of position can be paraphrased by change of position plus some embedding frame.

In the first iteration, we can only derive "direct" paraphrases, i.e., frame groups that are directly aligned to the target frame group in some bi-sentence. We cannot identify a paraphrase in the first bi-sentence, *This initiative will help to increase the level of employment/Diese Initiative führt zu einer höheren Erwerbsquote*: Even though the English side contains a CCPOS frame, this particular instance does not realise all relevant roles. In consequence, condition (2), which ensures the completeness of acquired frame groups, fails. However, the algorithm works for the second sentence, *The drought increased tea prices/Die Dürre hat zu höheren Teepreisen geführt*. It leads to the addition of the frame group {CPOS, CAUSATION} to the paraphrase list, with the role mapping {CCPOS.ITEM → CPOS.ITEM, CCPOS.CAUSE → CAUSATION.CAUSE}. The third sentence cannot be captured, since it does realise the CAUSE role.

In the second iteration, the algorithm can now analyse the first sentence by taking advantage of the paraphrase {CPOS, CAUSATION} acquired from sentence 2 in the first iteration. Since we find it here aligned to the frame group {CCPOS, REQUIREMENTS}, the latter is also added to the set of paraphrases. This is an example of a "second-order" paraphrase that has been obtained via a direct paraphrase of the target frame group. The role mapping is constructed by functional composition of the (known) mapping from {CCPOS} to {CPOS, CAUSATION} with the (new) word alignment information in the sentence. This results in the following role mapping: {CCPOS.ITEM → CCPOS.ITEM, CCPOS.CAUSE → REQUIREMENTS.DEPENDENT}. The algorithm still cannot analyse the third example in the second iteration. In the third iteration, no new frame group paraphrases are acquired, and the algorithm terminates.

In conclusion, our algorithm has successfully identified full paraphrases in sentences 1 and 2. No full paraphrases was found in Sentence 3, since this instance does not realise all relevant roles. If necessary, the lossy

paraphrase in sentence 3 can be obtained easily by running the algorithm another time, deleting the CAUSE role from the list of relevant roles.

8.4. Experimental Evaluation

8.4.1. Setup and Data

This section evaluates the usefulness of frame group paraphrases to model parallelism of predicate-argument structure on a real-world corpus. We decided to use a manually annotated corpus sample in order to assess the fundamental feasibility of our approach without interference from the noise that necessarily arises from state-of-the art automatic shallow semantic parsing. An additional consideration was that instances of translations across parts of speech (such as *increase–höher*) constitute an especially interesting phenomenon, since they systematically involve changes in valency across languages. Unfortunately, almost all current shallow semantic parsers for FrameNet concentrate exclusively on verbal predicates for English. It therefore turned out to be almost impossible to obtain bilingual, accurate frame-semantic analyses for instances of cross-POS translation pair and their contexts.

Data Choice. We finally decided to analyse a single translation pair for which we expected frame instance parallelism to break down systematically.

Specifically, we investigate the English predicate introduced in Example (8.1), namely *increase*. This predicate exhibits large variation: It can be used either as a verb or a noun. The verb reading furthermore can be used both transitively and intransitively. As shown below, the transitive verbal usage evokes the causative CCPOS frame, even if the CAUSE role is not realised overtly (Example (8.3)), while the intransitive verbal and nominal usages evoke the stative CPOS frame (Examples (8.4) and (8.5)).

(8.3) There is a desire to **increase** public spending. (CCPOS)

(8.4) By 2010, emissions will **increase** by 6%. (CPOS)

(8.5) An **increase** in road traffic in no way automatically implies more pollution. (CPOS)

To maximise the incidence of translation instances that cross the part-of-speech boundary, we paired *increase* with its most frequent German adjectival translation, namely *höher (higher)*. Due to its restricted valency, *höher* can only evoke the stative CPOS frame:

(8.6) Wenngleich der Welthandel einen **höheren** Wohlstand zur Folge hat.
Even though the world trade a **higher** prosperity as consequence has. (CPOS)

Corpus Construction. As basis for our extraction, we used the complete English-German EUROPARL bitext with the automatic lemmatisation and syntactic analysis described in Section 2.3. We applied the automatic intersective word alignment (cf. Section 2.2.2) to find all instances of the translation pair *increase–höher*. This resulted in 122 bi-sentences, which form our sample corpus. For all sentences of this corpus, two annotators hand-corrected word alignment and syntactic structure, and manually assigned FrameNet frames for both languages. All decisions were made unanimously. In analysing the sentences, we detected two frame gaps in FrameNet 1.3 and constructed two new frames in accordance with FrameNet frame construction principles. In the following, these frames are marked with an asterisk. Their definitions are given in Table 8.2.

8.4.2. Method

We applied the iterative matching algorithm (Figure 8.3) to the sample corpus in the same fashion as in the example in Section 8.3.1. We looked for full paraphrases of the target frame group CCPOS, with the relevant role set {ITEM, CAUSE}. We assumed one base frame, CPOS, with the role mapping {CCPOS.ITEM → CPOS.ITEM}. We now describe how the algorithm acquired paraphrases incrementally.

Iteration 1. The algorithm identified new paraphrases only on the German side, since the only paraphrase available for matching, CCPOS, occurred solely on the English side. This resulted in 10 embedding frame types, of which 3 occurred more than once (CAUSATION, GIVING, REQUIREMENTS).

Frame Name	Definition
Allotment*	An Allotment_Event distributes a Theme.
Means*	An entity or state of affairs represents a Means to achieve an Effect.

Table 8.2.: Definition of self-constructed frames

Iteration 2. Frame group paraphrases were now being found on either side, which led to the identification of four more frame paraphrases. In total, 8 embedding frames occurred more than once (the additional ones are Cognitive_connection, Commerce_pay, Deciding, Means*, and Request), and 4 exactly once.

Iteration 3. No new frame paraphrases were found, and the algorithm reached a fixpoint for P. However, some additional instances of known paraphrases were found, resulting in a total of 10 frame group paraphrases which were attested in the corpus more than once (Purpose was the only addition to this group), and 3 with one attestation.

8.4.3. Quantitative Evaluation

In this section, we perform an empirical test of the hypothesis that frame group parallelism holds for a greater number of bi-sentences than parallelism of individual frames.

We first assess the coverage of the baseline hypothesis, namely frame instance parallelism (cf. Chapter 3), which predicts that all aligned source–target predicate pairs in the corpus evoke the same frame. To test this hypothesis, Table 8.3 lists the frame pairs evoked by the translation pair *increase/höher* in the sample corpus. Not surprisingly, the results show that there is a significant amount of frame mismatch: All German instances are adjectival, and therefore evoke the Cpos frame. This corresponds well to the English nominal and adjectival (past participle) cases; however, the majority of verb instances in English is transitive, and invokes the (mismatching) Ccpos frame. The first line of Table 8.3 shows that frame instance parallelism can account only for 73 of 122, or roughly 60%, of all

English	German	Frequency
CPOS (36 n, 13 v, 24 adj/part)	CPOS (adj)	73
CCPOS (49 v)	CPOS (adj)	49

Table 8.3.: Cross-lingual breakdown of single frame pairs evoked by *increase/höher*.

cases.

We next investigate an alternative hypothesis, namely whether the frame group paraphrases acquired by our iterative matching algorithm show a high coverage. Table 8.4 shows the frequencies of the different cross-lingual patterns (frame groups or individual frames) which we observed in the corpus. Since the identification of paraphrases for the current dataset hinges on the existence of CAUSE-like frame elements, we organised the table according to the existence of realised CAUSE roles on either side. The rightmost column in the table indicates for each pattern whether it has been acquired by our algorithm, and can therefore be modelled successfully by FGP instance parallelism. Successful acquisition is denoted by '+', failure by '-'.

The first row (0+0) covers examples where a CAUSE exists neither in English nor in German. Slightly more than half of all instances (65 of 122) fall into this category. About two thirds are simple matching CPOS cases, but one third consists of German CPOS frames that correspond to CAUSE-less English CCPOS frames, such as infinitives and participle constructions without subject. The 0+0 cases were not retrieved in our experiment, since they do not realise a CAUSE role, and thus cannot constitute full paraphrases. They are therefore set in parentheses. However, as discussed in Section 8.3.1, they could be acquired by a simple repetition of the experiment, looking for lossy paraphrases with only {ITEM} as the set of relevant roles.

The second row (1+1) contains examples where a CAUSE exists in both languages (45, i.e., about 40%) and for which our algorithm established full paraphrases by identifying matching CAUSE roles. In German, the CAUSE is always contributed by an embedding frame, since *höher* cannot introduce one itself; in English, the CAUSE is realised either directly in

Cause	English	German	Frequency	Success
0+0	CPOS	CPOS	45	(+)
	CCPOS n.c.	CPOS	20	(+)
1+1	CPOS Frame Group	CPOS FG	22	+
	CCPOS	CPOS FG	14	+
	CCPOS FG	CPOS FG	9	+
1+0	CPOS FG	CPOS	4	-
	CCPOS	CPOS	2	-
	CCPOS FG	CPOS	3	-
0+1	CPOS	CPOS FG	2	-
	CCPOS n.c.	CPOS FG	1	-

Table 8.4.: Cross-lingual breakdown of frames and frame groups evoked by *increase/höher* (FG: as base frame of frame group; n.c.: instances without realised CAUSE role; Success: Successfully acquired Frame Group Paraphrase).

CCPOS, or is provided by an embedding frame. All of the 1+1 cases but one could be analysed as frame group paraphrases of CCPOS by our algorithm. The sentence that could not be handled will be discussed in Section 8.7.[6]

The lower half of the table shows 12 cases (10%) where a CAUSE exists on one side only. 11 cases are instances of proper cross-lingual divergence: In 9 cases, an English CAUSE does not have a German counterpart, and vice versa in 3 cases. We interpret this asymmetry as a slightly stronger preference in German to conceptualise situations as events without an overt CAUSE (e.g. "es entsteht" - it arises). Clearly, our algorithm failed to acquire paraphrases for these cases, due to the role mismatch.

In total, we find that our algorithm acquires frame group paraphrases that account for parallel structure in all 0+0 and 1+1 cases, 110 of 122 or roughly 90% of all instances. Compared to the result for simple frame instance parallelism (60% of instances), we find that FGP instance parallelism covers substantially more instances.

[6]Six of the 45 instances would be problematic to identify in a real-world scenario, though. Their CAUSE roles are not directly word-aligned, but have to be recovered by anaphora resolution.

Frame	Cause FE	Embedding FE	Frequency (German)
Accomplishment	Agent	Goal	1
Allotment*	Allotment_event	Theme	1
Attempt_suasion	Addressee	Content	1
Cognitive_connection	Concept_1	Concept_2	5
Causation	Cause	Effect	15
Commerce_pay	Buyer	Money	2
Deciding	Cognizer	Decision	1
Giving	Donor	Theme	3
Means*	Means	Effect	3
Purpose	Means	Goal	3
Request	Speaker	Message	3
Requirements	Dependent	Requirement	5

Table 8.5.: Identified frame paraphrases for CCPOS which contribute a CAUSE role (* = self-defined frame)

8.4.4. Qualitative Evaluation

This section addresses the question of how felicitous the frame group paraphrases acquired from the corpus are, i.e., how well they express similar states of affairs. Table 8.5 lists the 13 frames our matching algorithm identified as embedding frames in frame group paraphrases for the target frame group CCPOS as well as their frequencies. In addition, Table 8.6 lists normalised surface forms for all paraphrases we obtained.[7]

All frames listed in Table 8.5 provide a role which our iterative matching algorithm has mapped onto the CAUSE role of the CCPOS frame. Therefore, it makes sense to evaluate how well these frames express causation, in addition to a general assessment of how well they paraphrase CCPOS. To do so, we compare our paraphrases to a standard account of the linguistic realisation of causativity from the area of cognitive linguistics, namely

[7]The slightly larger number of German expressions does not necessarily mean that there is more variation in German; this is merely a reflection of the experimental setup which allows English to express the complete state of affairs as a single frame (CCPOS), which is not possible in German.

Frame	English	German
Accomplishment	–	C erreicht höheres I
Allotment*	–	C sieht höheres I vor
Attempt_-suasion	C is encouraged to increase I	C zu höherem I anspornen
Cognitive_-connection	C helps to increase I; I is not unrelated to increasing C; C means increasing I	höheres I hat mit C zu tun; das mit C verbundene höhere I; C gewährleistet höheres I; C bedeutet höheres I
Causation	C leads to increasing I; C brings increasing I; C produces increasing I; increasing I results from C; C has the effect of increasing I	C führt zu höherem I; C hat höheres I zur Folge; C bewirkt höheres I; C ruft höheres I hervor; höheres I entsteht durch C; C bedingt höheres I; C bringt höheres I
Commerce_pay	–	C leistet höheres I
Deciding	C decides to increase I	C legt höheres I fest
Giving	C gives higher I; C provides higher I	C zahlt höheres I ein; C verleiht höheres I; C stellt ein höheres I zur Verfügung
Means*	C is a measure to increase I; C is an instrument to increase I	C ist Massnahme für ein höheres I; C ist Instrument für ein höheres I
Purpose	C is intended to increase I	Ziel von C ist höheres I; Zweck von C ist höheres I
Request	C demands a higher I; C's request for a higher I	C fordert höheres I
Requirements	C necessitates a higher I; C requires a higher I; C needs a higher I	C ist notwendig für höheres I; C bedarf höheren Is; C macht ein höheres I erforderlich

Table 8.6.: English and German paraphrases for CCPOS identified by iterative matching (C = CAUSE; I = ITEM; * = self-defined frame).

Talmy (2000). Talmy performs a cognitively motivated study on the expression of causal relationships in natural language, and concludes that causation is ubiquitous in our conceptualisation of reality. In consequence, causal relationships are not only present when they are expressed overtly, but form part of the understood meaning of many other predicates. Depending on the properties of these meaning components, he distinguishes different causation situations.

Talmy's *basic causative situation* describes the overt expression of causativity, where one event results from another event. This situation is described by the Causation frame, the most frequent embedding frame in our corpus sample. A check of the expressions in Table 8.6 confirms that all of these instances are straightforward paraphrases of Ccpos.

Next, five of the frames in Table 8.5 are related to Talmy's *agentive causation*: The frame Accomplishment (an Agent achieves a Goal) is just *agentive causation* itself; Attempt_suasion and Request can be classified as *caused agency*; Purpose and Allotment* (some Theme is allotted to an intended Recipient through some Allotment_event) can be grouped under *purpose and uncertain fulfilment*. Disregarding modality issues, these frames can also still be accepted as general paraphrases of Ccpos.

An interesting, but less straightforward, embedding frame is Cognitive_connection (a Concept_1 contributes to an Concept_2). While structurally very similar to Causation, the frame Cognitive_connection does not imply that the Concept_1 is the most important factor in bringing about the Concept_2. The frequent use of Cognitive_connection provides an empirical handle on the problem of *gradience in causation*: In almost all real-world states of affairs, more than one cause is involved in bringing about an effect, and intuitions may diverge on whether a given cause is seen as especially prominent (warranting a Causation frame) or not (warranting a Cognitive_connection frame). Talmy notes (p. 544) that "one of the more significant issues wanting attention pertains to the existence of gradience in causative concepts", but does not offer a detailed analysis. Our study, and the methods we have developed, offer an approach to pursue this direction of research: Through exploiting the cross-lingual differences in conceptualisation, it would be possible to conduct a corpus-based study of this gradient causativity and inspecting the expressions which evoke Cognitive_connection in

Table 8.6.

The remaining five embedding frames are less applicable as general paraphrases, showing an increasing degree of context dependence. This mirrors the general pattern that we found in Section 5.7 when analysing translational shifts, which is not very surprising. In that discussion, we established that translations often differ with respect to the verbalisation of separate, but frequently co-occurring events, where – at least in certain contexts – a mention of one of the events carries a strong default assumption that the other event takes place as well.

For three of the five frames, there is still a fairly general relation between the two events, namely one of prerequisite, or, symmetrically, consequence. These are the MEANS* (something is a MEANS for achieving an EFFECT), REQUIREMENTS and DECIDING. Even though they do not directly fall into any of Talmy's groups, they show considerable similarity to the *purpose and uncertain fulfilment* class, with the difference that what is described is just an intention, while it is unclear that the following causing event itself takes place. For example, in the general parliamentary context from which EUROPARL is drawn, DECIDING for an increase of something to happen has the default consequence that the increase will take place, making the COGNIZER of the DECIDING event effectively a CAUSE.

The last two frames, GIVING and COMMERCE_PAY, express paraphrases for a specific situation type, namely situations of transfer where the ITEM is a money-related concept such as *payment, premium, contribution*. In this context, an increase of the ITEM can be paraphrased as being caused by the person or entity providing the money.[8]. However, this relationship is too idiosyncratic to be explained well within Talmy's typology of causality.

In sum, we find that the majority of frame group paraphrases acquired by our algorithm are acceptable general paraphrases, and fit well with Talmy's (2000) classification of causation. One of the somewhat problematic paraphrases, COGNITIVE_CONNECTION results from a semantic problem, gradience of causation, which is presumably especially prominent in our sample. We identified one more general problem, namely that we acquired situation-specific paraphrases, an effect that is very similar to the translation shifts we discussed in Section 5.7. This problem makes it necessary to provide a means of *determining the applicability* of a para-

[8]Of course, this mirrors – as discussed above – a simplified view of the causal chain.

phrase for a given sentence, when using automatically acquired frame groups for role projection (see Section 8.5 below). Our first observation is that the most frequent paraphrases (cf. Table 8.5) are all general. In conjunction with the skewed frequency distribution (the three most frequent paraphrases account for about 60% of all cases), this suggests that a simple frequency cutoff can be useful to separate general from specific paraphrases. A more principled, but also more tentative, strategy could take advantage of information present in the FrameNet database: It can be checked whether the headwords of roles assigned by potential frame groups are in fact likely fillers of these roles. Such a check can take place by comparison to the role's selectional preferences (cf. the discussion of the monetary transfer situation above). In English, selectional preferences can be obtained directly from the annotated example sentences in the FrameNet database (Gildea and Jurafsky, 2002); they can be projected onto other languages e.g. using Pitel's (2006) approach (see Section 7.4 for a discussion).

8.5. Cross-lingual Transfer as a Bootstrapping Cycle

At the outset of this chapter, we motivated our interest in frame groups with their potential use in a projection scenario, analogously to Parts II and III. However, the algorithm we have presented in Section 8.3 presupposes the existence of a parallel, bilingual corpus which provides frame-semantic annotation both for the source and the target languages. This seems to run counter to the idea of projection, where we assumed previously that nothing was known about the semantic structure of the target language; or rather, that all information about the target language could be obtained from the source language, and transferred to the target language by taking advantage of frame instance and role instance parallelism.

However, it is exactly this parallelism assumption which we have generalised in this chapter to frame group paraphrases. As we have seen in the Section 8.4, this allows us to cover a considerably higher portion of translation pairs. The price we pay is that we have to provide more knowledge: Frame group paraphrases can be interpreted exactly as de-

scriptions of the variation in overt frame-semantic structure over which we aim to generalise, and this information is best acquired from parallel bi-sentences with frame-semantic analysis for both languages.

Fortunately, this does not necessarily require manual frame-semantic annotation in the target language. In a nutshell, what we propose is to set up a bootstrapping cycle which interleaves (generalised) projection, the induction of monolingual shallow semantic parsers, the application of these shallow semantic parsers, and the acquisition of frame group paraphrases. The fundamental idea here is to use shallow semantic parsing to generalise markup from "easy", i.e., frame-parallel and role-parallel instances to non-parallel instances from which frame group paraphrases can be acquired. We now describe the individual steps in detail:

Step 1: Projection. As usual, we assume that we have a parallel corpus whose source side is annotated with frames and roles. In this step, we restrict ourselves to projecting semantic information for two cases: (a), parallel frames, and (b), parallel known frame group paraphrases. The cases of type (a) can be handled with the methods developed in Parts II and III. The cases of type (b) require new methods; however, since in the first iteration of the bootstrapping cycle no frame group paraphrases are available, we defer the description of projection for type (b) to below. In any case, the result of this step is a partially annotated corpus in the target language in which all frame-preserving translations are analysed. For example, a translation of *to lead to* results in the analysis of *führen zu* as FEE for the CAUSATION frame; analogously, the translation pair *higher* – *höher* causes *höher* to be recorded as FEE for CPOS.

Step 2: Application of a shallow semantic parser. In this step, the annotation projected in Step 1 is used to annotate *other* examples of the same predicates in the target language. To this purpose, well-understood supervised learning techniques to induce shallow semantic parsers can be used (see Section 1.2 for details). The crucial contribution of this step is, as noted above, that it can provide annotation for those target instances whose semantic structure is not parallel to their translation on the source side. For example, an analysis will be constructed for the predicates *höher* and *führen* on the German side in Figure 8.1.

Step 3: Acquisition of frame group paraphrases. After the successful completion of Step 2, we have a partially annotated bilingual corpus, in which the semantic structures of the source and target sides will disagree for a number of bi-sentences; again, Figure 8.1 is an example. These instances can serve as input for the acquisition of frame group paraphrases.

The completion of the cycle results not only in a set of frame group paraphrases available. When the cycle starts over, the case (b) of Step 1 introduced above applies for the first time: projection can now also take place for predicates whose translation is not frame-preserving, but forms part of a frame group. Note that we are now no longer projecting roles frame by frame, as laid out in Section 3.2.3, but frame group by frame group. Therefore, determining whether projection can take place requires checking whether the projected frames and roles make up a well-formed frame group, as defined in Section 8.2.2. In the following, we sketch the necessary checks, assuming that we have a frame group paraphrase (f_1, f_2) and a bi-sentence (s, t).

1. The frame group f_1 is evoked in the source sentence s, and its base frame is evoked by a source predicate p.

2. p is aligned to a target predicate which can evoke the base frame of f_2.

3. If f_2 also consists of an embedding frame, t must contain a predicate that can evoke this embedding frame, and there must be a constituent which (a), is a plausible realisation of the embedding role, and (b) whose semantic head is either the FEE of the base frame, or has the FEE of the base frame as a modifier.

The third condition is aimed at verifying that any embedding frame provides a proper embedding role for the base frame (cf. the definition of frame groups in Section 8.2.2). This check is difficult to perform properly, since embedding roles in frame groups, such as the EFFECT role in Figure 8.1 (page 192), are not part of the role mapping – they simply do not have to have a corresponding role. Therefore, information about the embedding role cannot be projected when frame group paraphrases apply.

Condition 3 is therefore currently phrased heuristically, using monolingual information on the target side to emulate the structural check, but clearly requires more research.

In the bootstrapping cycle sketched above, frame group projection and shallow semantic parsing can be seen as complementary means of extending the coverage of single-frame projection: Shallow semantic parsing exploits the available monolingual evidence, and frame group projection exploits bilingual evidence. This suggests the bootstrapping cycle can succeed in a "cautious", as opposed to a "greedy", fashion: The availability of two strategies to extend the set of known frame annotations means that both can restrict themselves to processing (i.e., projecting and labelling) those examples which can be treated with the highest confidence. This strategy has thus a good chance of avoiding "contamination" of its datasets, a pervasive problem with bootstrapping approaches (Riloff and Jones, 1999).

8.6. Related Work

The twofold nature of frame group paraphrases as characterisations of synonymous or near-synonymous phrases, and as descriptions of translational equivalence, means that related work falls into two broad categories, one related to each aspect.

Paraphrase identification. In recent years, the acquisition of paraphrases from large parallel datasets has received a lot of attention. Usually, paraphrases were identified using distributional evidence on different linguistic levels from surface strings (Barzilay and Lee, 2003; Bannard and Callison-Burch, 2005) to syntactic structures (Pang, Knight, and Marcu, 2003). To our knowledge, our study is the first one to assess the potential of using *semantic* structures to characterise paraphrases. One important benefit of characterising paraphrases in terms of frame groups is that each frame itself is an abstract representations of a number of surface realisations, of which the paraphrases shown in Table 8.6 are mere examples. It is therefore possible to create paraphrases unseen in the training data.

An important difference of the current study to almost all other paraphrase studies, with the exception of Bannard and Callison-Burch (2005),

is that we use bilingual corpora to identify paraphrases. This is crucial for applicability, since bilingual parallel corpora are much less scarce than monolingual parallel corpora. Nevertheless, this does not mean that we can only acquire cross-lingual paraphrases. Due to the high degree of concept-level parallelism of frames (cf. Section 3.2.1), frame group paraphrases are a substantially *language-independent paraphrase model*. By using the FEE lists from frame-semantic lexicons, language-specific surface instances for these paraphrases can be generated. Using one lexicon results in monolingual paraphrases; several lexicons lead to cross-lingual paraphrases. It is a matter of further research to determine whether the same paraphrases arise from monolingual and cross-lingual parallel corpora, or whether there is variation.

The central criterion with which we identify frame group paraphrases is that all *relevant roles* have to be filled. This criterion is similar to the "matching argument slots" method used by Barzilay and Lee (2002, 2003). However, the background is different: Barzilay and Lee use multiple sequence alignment on several raw comparable corpora, identifying argument slots as stretches with high variability across corpora. Since comparable corpora do not allow for direct structural matching, matches between argument slots are then induced from similarities between fillers for these sequences found in texts from a short time span (e.g., a few days).

Clearly, a major difference between such paraphrase induction methods and ours is that they are *framework-neutral*, which has important consequences on two different levels.

Status of induced argument slots. Empirically induced argument slots do not necessarily correspond to semantic arguments in the sense of FrameNet; they rely more on the quality and properties of the assumed corpora. For example, Barzilay and Lee's heuristic of identifying arguments as stretches of high variability might miss argument slots which show little variation, e.g. due to low overall frequency. Thus, there is no guarantee that argument slots are identified consistently across diathesis alternations. On the other hand, place and time expressions are likely to be identified as arguments, while they are not "Core" elements in the sense of FrameNet (cf. Section 3.2.1).

Applicability. On the other hand, much fewer resources are necessary with a completely data-driven approach: all Barzilay and Lee re-

quire for the induction of paraphrases is a large number of comparable corpora, while our method requires bilingual frame-semantic analyses. While this makes theory-neutral paraphrase induction methods definitely more easily applicable to new languages, note that the bootstrapping cycle sketched above (Section 8.5) is aimed at alleviating exactly this problem.

Transfer-based machine translation. The aim of transfer-based machine translation is to construct translation models for specific language pairs by identifying a level of representation where translational equivalences can be represented compactly as correspondences between structures of either language (Hutchins and Somers, 1992). The underlying assumption is that such correspondences are easier to specify on syntactic or semantic structures than on the surface structure itself, since these levels provide a higher degree of instance-level parallelism, while fully language-independent representations (i.e., interlinguas) are difficult, if not impossible, to obtain.

As in our study, it is usually assumed that these correspondences can be obtained from contrastive analyses of bilingual corpora. One instance of transfer-based machine translation which is close in spirit to our approach is Dorna and Emele (1996), who model English-German translations with correspondences between sets of argument relations of a Neo-Davidsonian event semantics (Dowty, 1989). In fact, a number of Dorna and Emele's motivations for transfer on the semantic level, such as the preservation of correct predicate-argument relations, and the mismatch between lexical items and semantic entities, arise naturally from the frame-semantic analysis we assume.

However, our aim is clearly more modest than full-fledged transfer-based machine translation. Our representation on the frame level is much weaker than Dorna and Emele's full logical analyses. They also introduce a calculus to express directed transfer, which is not possible in our current framework, which only identifies matching chunks of predicate-argument structure that can be used interchangeably. Importantly, their transfer approach was integrated into a large-scale MT system (the Verbmobil system) which could use the output of the transfer component to generate target language text. In contrast, generation from frame-semantic representations is a research question that has yet, to our knowledge, to be

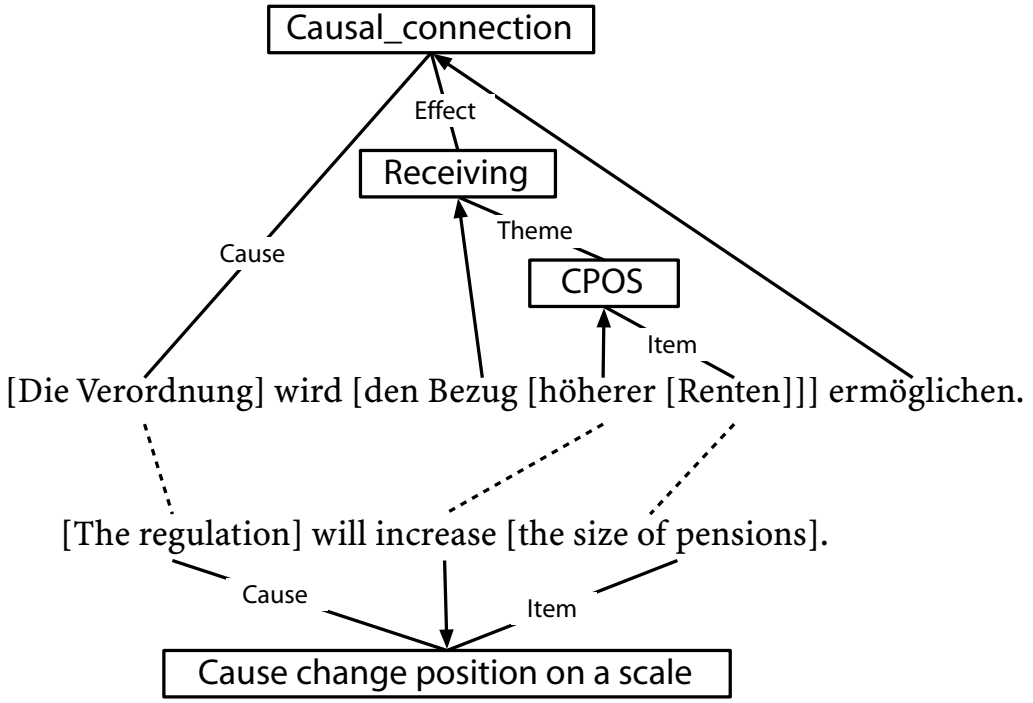


Figure 8.6.: Problematic bi-sentence: A German frame group involving more than two frames

addressed.

8.7. General Discussion

Our pilot study has made three main simplifying assumptions: we have limited our attention to frame groups of size two, assumed perfect linguistic analyses, and concentrated on one translation pair. These assumptions cannot be upheld if our methods are to be applied for automatic, broad-coverage frame group paraphrase acquisition. In this section, we discuss the consequences, with the understanding that the empirical evidence on which we base our discussions is limited.

We first consider the size of frame groups. Recall that in our current study, for which we only used frame groups of size 2, we were able to acquire frame group paraphrases which accounted for 90% of all bi-sentences. In Section 8.4.3, we noticed that all bi-sentences which could

not be analysed were cases of role mismatch, except for one. This instance is shown in Figure 8.6 and illustrates the limits of the current definition of frame groups. The German sentence can be paraphrased as *The regulation will enable* **the drawing** *of higher pensions*. In contrast to the instances we have previously seen, there is not only redistribution of existing semantic material, but introduction of new material (marked in bold in the transcription). On the level of frame analysis, the German frame group corresponding to the English frame CCPOS thus consists of three frames: the frame COGNITIVE_CONNECTION embeds an intermediate frame, RECEIVING, which in turn embeds CPOS.

In principle, our algorithm and our definition of frame groups is capable of dealing with frame groups of more than two frames, provided that the definition of embedding can be suitably generalised (see Padó and Erk (2005)). However, it is not clear whether such a generalisation is indeed desirable. The example in Figure 8.6 is already a fairly specific paraphrase which is only applicable if the ITEM in question is scheduled for a transfer in connection with its change of scalar position. We speculate that larger frame group paraphrases in general would tend to model more context-specific paraphrases. In sum, we have currently no evidence that allowing larger frame groups would significantly increase the yield of generally applicable frame group paraphrases.

The second assumption mentioned above is the idealisation of linguistic analyses. Clearly, an automatic application of our algorithm will encounter problems introduced by errors in parsing, semantic analysis, and word alignment. While we do not rely much on the syntactic structure per se, syntactic errors can in turn lead to errors in the semantic analysis. Also, errors in word alignment can lead to incomplete role mappings. Still, we are confident that our algorithm is applicable to noisy data, since it pursues a "high precision, low recall" strategy: It imposes strong checks on potential frame group paraphrases (cf. Section 8.3), especially the test for matching roles. Since noise tends to break systematic matches, we can be fairly confident that if a given frame group paraphrase meets all conditions, it is in fact correct. As discussed in Section 8.6, contamination is therefore not a large issue. This high-precision strategy is feasible because it can exploit corpus redundancy: Since our algorithm works on the level of individual instances, one correct analysis in the whole corpus is in principle enough to acquire a frame group paraphrase.

With respect to applying the algorithm to more data, we feel that the good performance of the algorithm in this study is a promising result. We do not foresee any difficulties, on the part of the algorithm, in applying it other translation pairs. The most serious limitation for wide-coverage frame group paraphrase acquisition appears to be the availability of high-quality shallow semantic parsers which can analyse all predicates in free text. This problem can be further divided into two parts: (i), missing coverage of predicates which can be described with existing FrameNet frames; and (ii), missing frames in FrameNet. While a number of approaches have been proposed for (i), it currently appears that (ii) can only be addressed by further lexicographic efforts (compare also the discussion in Section 5.5).

8.8. Summary

This chapter has provided first steps towards providing an account of translation pairs with non-parallel frames. These cases cannot be treated by the standard annotation projection-based methods discussed above, despite representing a considerable portion of all corpus instances. Even though cases of non-parallel frames in their entirety constitute a very difficult and heterogeneous class, our intuition is that *local rearrangements* of semantic material during translation are a frequent reason for frame instance non-parallelism that can be modelled relatively straightforwardly.

To capture this class of cases, Section 8.2.2 has introduced the concept of *frame group paraphrases*, groups of up to two connected frames on either side of a bi-sentence which describe the same state of affairs by virtue of having corresponding sets of semantic roles. Section 8.3 has provided a semi-supervised algorithm for identifying new frame group paraphrases in parallel, word-aligned corpora. Next, in Section 8.4, we have verified, within the limits of a manual pilot study, that the acquired frame group paraphrases (a), can model the parallelism of a higher number of translations than individual frames; and (b), can be seen as language-independent paraphrase templates, all of which are linguistically interpretable. Building on these results, Section 8.5 has outlined how parallelism on the level of frame group paraphrases can be used to extend annotation projection to cases of translation pairs with non-parallel

frames. Sections 8.6 and 8.7 conclude with discussing related work and our results.

In sum, this chapter demonstrates that the annotation projection approach proposed in this thesis is not confined to treating only the relatively easy cases of frame-preserving translations. It can in fact account for translations involving a higher degree of deviation from the source expression as well.

9. Conclusions

In this chapter, we summarise the main results of this thesis, discuss the relation between the projection methods proposed in this thesis and manual annotation, and outline issues for future research.

9.1. Contributions

Large role-semantic resources, which are necessary to train shallow semantic parsers, exist only for English and a small number of other languages (see Section 1.1). In this thesis, we have developed methods that alleviate this resource scarcity by automatically constructing frame-semantic annotations for languages other than English. Our models make use of annotation projection, a framework which uses parallel corpora to transfer linguistic annotations from a source to a target language. Our methods are suitable even for resource-poor target languages, since the only prerequisites of basic annotation projection are (a), a parallel corpus, and (b), automatically obtainable word alignments.

We have successfully evaluated our methods using the English FrameNet database to obtain frame-semantic resources for two target languages, German and French. We have found that the semantic generalisations made by frame semantics carry over to a considerable degree from English to other languages. This is particularly encouraging, since it means that it is possible to apply the projection methods developed in this thesis for automatically creating frame-semantic resources for many other languages. To the degree to which this property holds for other role-semantic frameworks, our methods can be used for their projection as well (see Section 3.2.3 for a discussion).

Semantic parallelism. We have verified empirically that frame-semantic annotation exhibits a high degree of cross-lingual parallelism, that is, a

large majority of corresponding pairs of source and target predicates in an aligned corpus agree in their frame (i.e., semantic class) and set of semantic roles. On a trilingual sample corpus (English – French – German) with independent annotation for each language, we found a degree of cross-lingual instance-level parallelism of around 70% for frames. For matching frames, around 90% of the semantic roles agreed. This result demonstrates that the central precondition of high-quality annotation projection, namely a high degree of cross-lingual parallelism on the instance level, is met in practice.

Frame Projection. We have shown that it is possible to induce a small, high-precision frame-semantic predicate classification relatively quickly by combining word alignments with a small number of shallow filters. At a size comparable to the English FrameNet database, the induced classifications yield reasonable precision (approximately 70% for German and 65% for French). The classifications can serve for a number of tasks in the target language, notably to generally alleviate sparse data issues with class-based smoothing (Lapata et al., 2001), and to restrict role projection to instances where it is appropriate, i.e., instances with parallel frames.

The main problem in the projection of semantic classes is to distinguish instances where the target predicate of a translation pair preserves the frame evoked by the source predicate (e.g., *give – geben (give)*) from instances where this is not the case (e.g., *give – bekommen (receive)*). This task benefits most from morphosyntactic and distributional information, such as the translational entropy of the translation pair (Melamed, 1997). Syntactic information is not necessary for this task, since the single-word alignment between source and target predicate is sufficient to guide the projection process. Furthermore, the projection of frames suffers more from precision errors than from recall errors in the word alignment, since the redundancy in the corpus can make up for missing alignment links to a large degree, but erroneous links weigh down on precision.

Semantic Role Projection. We have also found it feasible to use word alignments as primary information source for the cross-lingual projection of semantic role annotations. The induced role annotations can be used as training data for shallow semantic parsers, which can subsequently

provide role-semantic analyses for free text in the target language.

The role projection task requires different types of knowledge from frame projection. It relies predominantly on the *semantic alignment*, i.e., the identification of counterparts for role-labelled source word spans in the target sentence. The main obstacle for this task is posed by gaps in the word alignment (i.e., missing links), since these lead to the projection of roles to incomplete spans in the target sentence. We have alleviated this effect by introducing bracketing information from automatic parsers into the projection process, to help enforce projection onto complete constituents. This has led to an improvement in the precision of projected roles from around 50% (using just word alignments) to up to 80% for both German and French.

We have found this scheme to be most effective when roles could be projected onto single constituents. This requirement rules out shallow chunk-based analyses that do not offer constituents spanning complete embedded sentences or complex NPs. Moreover, our experiments have indicated that the syntactic formalism used to represent tree structures does not seem to have a large effect on projection: we were able to compute high-precision alignments between constituent-based Penn Treebank-style analyses for English, the combined constituent/dependency-based TIGER analyses for German, and dependency-inspired analyses for French. An important factor for the success of our methods are our filtering mechanisms which reduce syntactic trees to their essentials and thus make them more similar.

One potential drawback of our method is its reliance on full parse trees. Fortunately, full parsers are becoming increasingly available for a wide range of languages. This development is driven both by refined techniques for inducing parsers from small treebanks (see e.g. the CoNLL 2006 shared task on multilingual dependency parsing (Buchholz and Marsi, 2006)) and by successes in unsupervised parsing (see e.g. Klein and Manning (2004)). In consequence, we expect that the reliance of role projection on full parse trees will cease to limit the range of possible target languages in the future.

Importantly, our experiments have demonstrated that role projection can even tolerate noisy input, such as role annotations produced by a shallow semantic parser. When we used a state-of-the-art shallow semantic parser (Giuglea and Moschitti, 2004), precision of the projected roles

remained constant, with losses confined to recall. This is a promising result, since high precision is arguably more vital for resource induction than high recall. Even in this condition, precision of the projected roles remained above 70% for both French and German.

Translational Divergences. We have also attempted to provide an account of translation pairs with non-parallel semantic classes. The latter cannot be treated by the standard annotation projection-based methods discussed above, although they represent a considerable portion of all corpus instances. Semantically non-parallel translation pairs form a very difficult and heterogeneous class, but we have identified a subclass of these cases, called *frame group paraphrases*, to which a generalised version of annotation projection can be applied. We have provided a semi-supervised algorithm for the identification of new frame group paraphrases which is again based on word alignments, and have outlined how it can be integrated in a bootstrapping cycle with standard annotation projection for parallel cases. In a manual pilot study, our algorithm has been able to acquire almost all frame group paraphrases, thus covering a substantial majority of all translations in our sample. The acquired frame group paraphrases can be seen as language-independent paraphrase templates, all of which are linguistically interpretable.

Our study of frame group paraphrases thus demonstrates that the annotation projection approach proposed in this thesis is not confined to treating only the relatively easy cases of frame-preserving translations. It can in fact account for translations involving a higher degree of deviation from the source expression as well. In addition, it can capture deep lexical-semantic information in the form of paraphrase templates.

9.2. Projection vs. Manual Resource Creation

There is ongoing interest in the computational linguistics community to develop role-semantic resources for a larger range of languages.[1] Projection methods are likely to play an important role in this endeavour because

[1] This interest is documented by the Global FrameNet group as well as by other events such as the Romance FrameNet Workshop in 2005 and the Scandinavian and Baltic FrameNet Workshop in 2007.

of their potential to substantially reduce the amount of human effort necessary; however, the quality of projected annotations still trails manual annotation. It is therefore an important practical question how projection methods can support manual resource creation most effectively.

In the initial phase of such a resource creation projection, projection methods can be used to create a repository of frame-semantic annotations in the target language automatically, without manual effort. Such a dataset is crucial to survey the challenges of the task and to prepare a smooth setting for manual annotation, in particular if there is no prior experience with frame-semantic annotation for the target language. Notably, projected data can be used to identify phenomena and predicates in the target language which are problematic with respect to the generalisations made by FrameNet (cf. Section 3.1.1), and to extract lists of predicates suitable for manual annotation.

In the annotation phase itself, projected annotations can be exploited using two broad strategies. The first possibility is to treat them as *annotation proposals*, all of which have to be reviewed by some human annotator (Swift, Dzikovska, Tetreault, and Allen, 2004). Alternatively, projected annotations can be employed directly to train a shallow semantic parser (cf. Section 1.2) which is able to analyse target language texts beyond the parallel corpus used for projection (Johansson and Nugues, 2006). Such shallow semantic parsers can be subsequently improved through active learning (Cohn, Atlas, and Ladner, 1994). These two strategies can be implemented in a large number of different ways; the optimal choice in each situation depends on the requirements and goals of the particular project, and on the amount of manpower available for annotation or correction.

9.3. Avenues for Future Work

We have already discussed some directions for future work in individual chapters. At this point, we outline two more general avenues for research which generalise different aspects of the bilingual annotation projection paradigm we have adopted in this thesis.

Annotation projection involving more than two languages. In this thesis, we have concentrated on using bitexts. A natural generalisation is

to consider *multitexts*, i.e., parallel corpora for more than two languages. While multitexts are rarer than bitexts, their number is steadily increasing, at least in the areas of politics, economy, and international newswire. Recall that the EUROPARL corpus itself is a multitext comprising 11 languages. Multitexts have already been exploited for example by Kuhn (2004) in the context of grammar induction, where he uses evidence from word alignment between multiple languages to bias the unsupervised recognition of constituents in a source language.

For the purposes of annotation projection, a very simple way of taking advantage of a multitext is to retain one designated source–target language pair, and to use the other languages to provide additional *redundancy*. The idea is similar in nature to *classifier combination* in machine learning, where several classifiers with different "views" on a problem (e.g., using different feature sets) can be combined to obtain a single, more accurate solution to the problem (Dietterich, 2000). In our scenario, the different "views" are contributed by the additional languages, or more specifically, by the projections of an instance of source language annotation into these languages. A comparison and combination of these views has the potential of making the projection more robust by alleviating errors in individual languages on all levels, e.g., word alignment, part-of-speech tags, or syntactic structure.

On the level of semantic classes, this idea can be instantiated concretely by generalising the filtering procedures from Section 4.4.2, notably the reliability filter, to model properties of translation pairs over the complete multitext. In the bitext situation, we found a correlation between translational reliability and the frame-preservation of the translation pair that was strong, but not perfect. Our hypothesis is that being able to estimate translational reliability for more than one language pair in the multitext will allow an even more accurate identification of frame-preserving translations.

As for semantic roles, the alignment model which we have found to yield the best and most robust optimal alignments, namely weighted bipartite matching (Section 6.3), can also be generalised to the multilingual case. Solutions to the resulting *weighted multipartite matching problem* are sets of alignment links that connect not pairs, but tuples of constituents from the different languages of the multitext. Here as well, our hypothesis is that an alignment computed on multiple languages is more reliable

than a bilingual one. Note that the multipartite matching problem is NP-complete (Karp, 1972); however, there exists a large body of work on algorithms which solve NP-complete problems efficiently enough for practical applications (see e.g. Althaus, Karamanis, and Koller (2004)).

Bidirectional information flow. Throughout this thesis, we have limited our attention to standard annotation projection in the sense that we transferred information in one direction, from the source to the target language. However, it is worthwhile to investigate strategies which feed projected information back into the source language.

This idea can be applied particularly easily to semantic classes. Its most naive application is to project induced predicate classes from the target language back onto the source language, and thus to obtain new and previously unknown frame-evoking elements in the source language. In this way, it represents a first step towards addressing the coverage problems associated with FrameNet that we have found to hamper unidirectional projection (cf. Section 4.4.1).

The bidirectional projection of semantic classes can be repeated in the form of a *bootstrapping cycle*, i.e., projection is performed alternately in either direction, with each projection step adding to the set of known frame-evoking elements for its target language. Unfortunately, this strategy runs the well-known risk of data contamination, i.e., the acquisition of wrong FEEs at some stage which misdirect succeeding iterations (Riloff and Jones, 1999). A promising alternative is to cast the bidirectional flow of semantic class information as a clustering problem in a bipartite graph (Bollobás, 1998). In this model, the partitions of the graph correspond to the source and target language predicates. The incomplete predicate classification for the source language imposes a clustering on a subset of the nodes of the first partition. Corpus-derived word similarity information can now be used to extend this clustering both to other nodes of the same partition (i.e., unknown source predicates), and to nodes of the other partition (i.e., target predicates). This approach does not rely on the existence of large linguistic resources, as monolingual methods for the extension of frame-semantic predicate classifications usually do (see Section 4.2). It should thus be applicable more easily to a wide range of languages.

Part V.

Appendix

A. Guidelines for the Evaluation of Projected FEE Candidates

(These are the original English guidelines given to annotators.)

A.1. Introduction

The theory of Frame Semantics holds that it is possible to define so-called **frames**, representations of prototypical situations, and that for each frame there is a set of words (**lexical units**, usually verbs, nouns or adjectives) which can **evoke** this frame. Each frame specifies a set of semantic roles (**frame elements**) which conceptually represent the participants and objects relevant for the situation depicted by the frame. Consider as example the definition of the Frame ARRIVING given in Table A.1. When a token of the frame's lexical units in actual text evokes a token of this frame, it is called a **frame-evoking element** (FEE) of that frame token. The frame elements (roles) can then be used to characterise the entities in the linguistic context of the FEE by describing which entity "fills" which semantic role. For example, in the sentence

- [Peter]$_{\text{Theme}}$ **came** [home]$_{\text{Goal}}$ late.

the FEE "came" introduces the Frame ARRIVING known from above. "Peter" fills the role of THEME since he is the one moving, and "home" fills the role of GOAL since this is where he ends up. Note that not all frame elements have to be realised with an occurrence of a frame; in the example

- [Mary]$_{\text{Theme}}$ **returned** from her trip.

Definition	An object Theme moves in the direction of a Goal
Frame Elements	**Goal (Go)**: Goal is any expression that tells where the Theme ends up, or would end up, as a result of the motion. **Theme (Th)**: Theme is the object that moves.
Lexical units	approach.n, approach.v, arrival.n, arrive.v, come.v, enter.v, entrance.n, get.v, make.v, reach.v, return.n, return.v, visit.n, visit.v

Table A.1.: Example definition for the frame ARRIVING

(again Frame ARRIVING), the GOAL is not realised overtly. What is important is that it could, in principle, be added to the sentence ("...to her hometown").

Often, lexical units are **polysemous** in terms of frames, that is, they may be able to evoke different frames in different constructions. Examples:

- Ich **frage** mich, wie spät es ist. (Frame COGITATION)
- Ich **frage** ihn nach einer Zigarette. (Frame REQUEST)
- Ich **frage** ihn, was er gestern gemacht habe. (Frame QUESTIONING)

Note that FrameNet is not complete; but in general, it is at least possible to determine that some example does **not** evoke a given frame, even if the "right" frame it should evoke does not exist (yet).

A.2. The Annotation Task

The present study attempts to automatically determine which words can evoke which frames in a new language (the **target language**) by taking into account knowledge about another language (the **source language**). Currently, we will concentrate on English/German and English/French.

The output of the computer program is a list which contains, for a number of frames, target language **FEE candidates** and their English **supports**, i.e. the words that the FEE candidate is supposed to be a translation of. The following is an example for frame ATTEMPT:

```
versuchen          try, attempt
Versuch            attempt, try
bemühen            try, attempt
```

The task of the annotator is first to decide whether the FEE candidate is **appropriate** for the frame (see below for a definition of appropriateness) or not; and if not, to determine the type of error. For an overview of the decision process, see Figure A.1.

The annotator can use the list as described above, and a list of concordances which list for every FEE candidate-support the first concordance in the corpus.

Annotate the tag directly after the FEE candidate, separating candidate and tag by one space.

A.2.1. Decision 1: When Is a FEE Candidate Appropriate for a Frame?

A FEE candidate is appropriate (Annotation `OK`) for a frame if two conditions are fulfilled:

1. It has parallel semantics to (at least one of) its supports and
2. It can evoke the frame

Parallel Semantics

A FEE candidate corresponds to a support if it is a reasonable translation, i.e. if they **can introduce the same semantics** into a sentence. In case of doubt, this can be checked with a bilingual dictionary (although dictionaries usually don't list all the possible translations), or with the concordances provided (although these contain only the first co-occurrence of FEE candidate and support, and should be treated with caution).

Usually, in order to introduce the same semantics, the FEE candidate and support both have to be either (a) single-word predicates or (b) the semantic heads of **multi-word expressions** (MWE). However, they **do not need to be the same part of speech**.

I take multi-word expressions to be expression whose meaning is not compositional (i.e. *to push up the daisies* is a MWE, but *to raise a question*

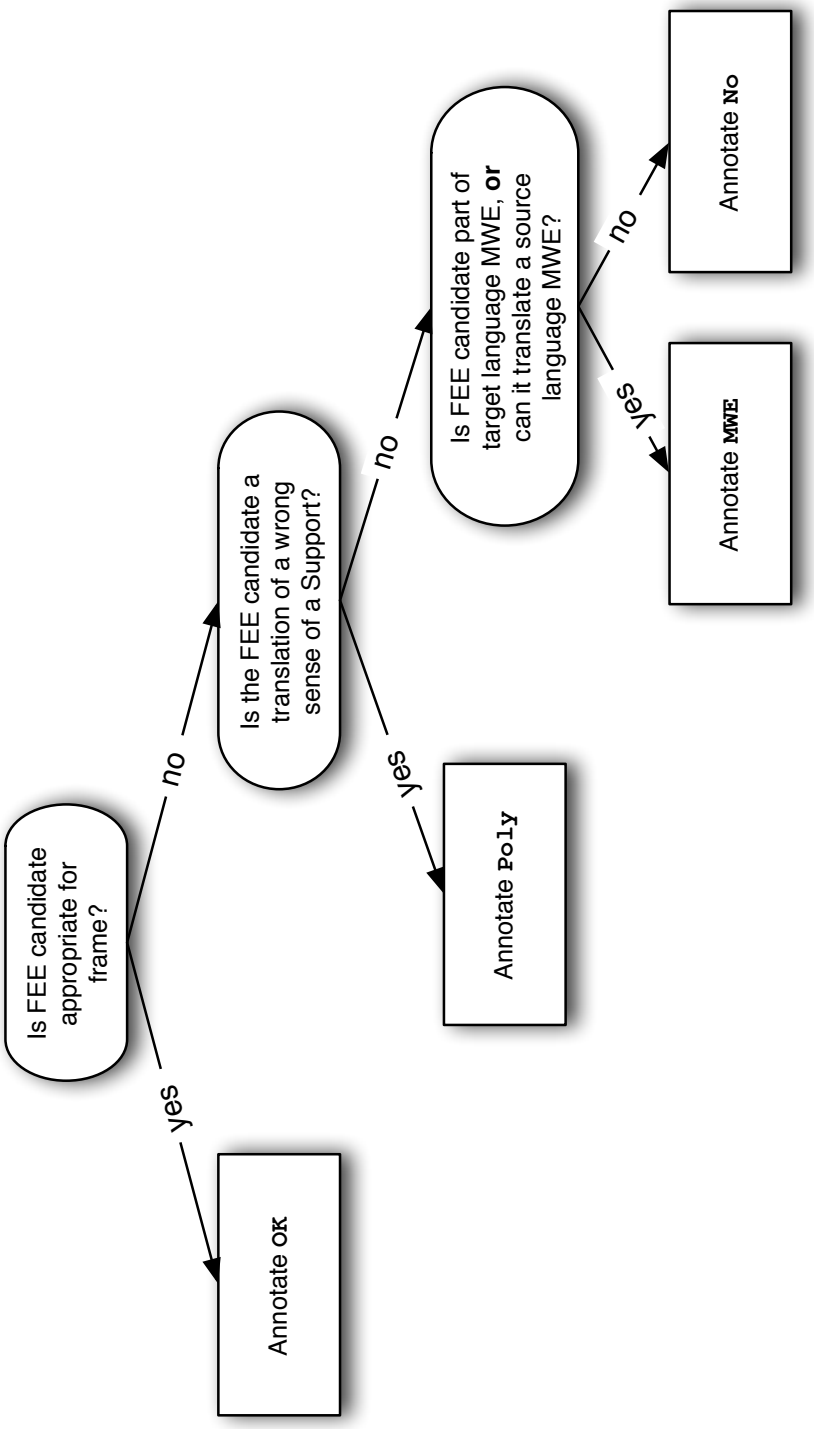


Figure A.1.: Decision Tree for the FEE candidate annotation

is not. In German, it usually makes sense to treat particle verbs and noun compounds as MWEs).

MWEs have to be treated **depending on their meaning**:

- FEE candidates are **appropriate if they are the semantic head of the MWE**, i.e. the meaning of the MWE is (nearly) the same as the meaning of the single candidate word, i.e. if they are the semantic head of the multi-word expression.

  This case occurs for this frequently for nouns and adjectives in frames that denote events where one has to add a "light" verb for syntactic reasons, but where the noun still retains the main semantic contribution. In the following example, **Siedepunkt** is appropriate for the APPLY_HEAT frame, because it can be used in the phrase **zum Siedepunkt.n bringen**, where "bringen" only adds aspectual information:

Siedepunkt	**boil**	APPLY_HEAT	→	**OK**

- However, the FEE candidate is **not appropriate if the frame can only be evoked by the combined semantics of all words of a multi-word expression.**

  This frequently happens for idiomatic expressions, as for the following, where "den Löffel abgeben" has nothing to do with "Löffel" any more:

Löffel	**die**	DYING	→	**MWE**

Evoking a Frame

The FEE candidate has to be able to introduce the Frame, as defined on the FrameNet website [1] under FN data / frames (disregarding any syntactic definitions for English). An example for a clear case:

[1] `http://www.icsi.berkeley.edu/\~framenet/`)

kochen	**cook**	COOKING_CREATION	→	**OK**

It is also okay if a FEE candidate is very specific, as long as it conforms to the definition of the frame:

Verkehrslärm noise	SENSATION	→	**OK**

It also sometimes happens that the semantic structure of the example sentences is rather parallel, but that the predicate has undergone a change of perspective:

erhalten	**give**	GIVING	→	**OK**

As long as the meaning is the same, this should be fine. If the semantic structures diverge too much, **No** should be annotated (see below).

There might be problems, though:

- **Frame definitions:** Some frame definitions tend to be vague. In that case, it may also help to browse the annotation examples for the lexical units of that frame on the FrameNet web site (FN data / lexical units) to get a better feeling of which examples are actually annotated with the current frame. Still, often there is a grey area in which it is unclear whether the FEE candidate is still appropriate or not.

- **MWEs**: Recall that MWEs evoke a frame if the FEE candidate is the semantic head, and only minor (non-frame-changing) information is added by the additional elements of the MWE.

- **Semantic association:** The automatic methods used for the current task tend to bring up associated, but not synonymous words as candidates; **FEE candidates are not appropriate if they are just semantically related to the support**. In the following example, **Pfanne** is not appropriate for the Frame APPLY_HEAT because in order to make it evoke the frame, you have to add a lexical unit that could evoke the frame itself (such as **braten in der …**).

Pfanne	**fry**	APPLY_HEAT	→	**No**

A.2.2. Decision 2: Does the FEE Candidate Result from a Wrong English Sense?

A FEE candidate is a result of a wrong English sense of a support (Annotation `Poly`) if

1. It has parallel semantics to (at least) one of its supports and
2. It corresponds to a sense of the support which cannot evoke the current frame.

With regard to parallel semantics, see Section A.2.1.

Other Sense of English Support

`Poly` stands for polysemy error (i.e. an error introduced by additional senses of the support): This happens frequently, because of the incompleteness in FrameNet and the resulting problems for frame disambiguation.

This frequently happens for verbs. For example, the verb "raise" which can evoke
CAUSE_CHANGE_OF_SCALAR_POSITION (CCPOS) also has different sense, as in "raise a question":

stellen	**raise**	CCPOS	→	**Poly**

The differences may be more subtle, though. As an example, consider the frame ABUSING and the German FEE candidates for the support **abuse**:

Missbrauch	**abuse**	ABUSING	→	**OK**

Misstand	**abuse**	ABUSING	→	**Poly**

The frame definition of ABUSING makes specific reference to domestic violence. Since **Misstand** cannot be used in the domestic violence sense, but is a translation of a more general sense of **abuse** as "a state of exploitation", it is not appropriate for that frame and receive the label `Poly`. On the other hand, "Missbrauch" can be used in that sense can should receive the label `OK`.

`Poly` is also appropriate if the meaning of the FEE candidate goes into the general direction of the frame, but does not quite fit. Example:

Wachstum	**increase**	CCPOS	→	**Poly**

Wachstum does not indicate Causation; rather, it is inchoative; therefore, **Poly** is the right label.

A.2.3. Decision 3: Are Misaligned Multi-word Expressions Involved?

A FEE candidate is the result of a misaligned multi-word expression (Annotation `MWE`) if

1. The FEE candidate and its supports are **not** semantically parallel and
2. At least one of them is part of a multi-word expression

Recall that MWEs are only expressions which receive their meaning only by virtue of their specific combination of words. Therefore

- Phrases whose meaning is compositional are **not** MWEs
- Compounds, particle verbs, idioms are MWEs

Clearly, MWEs can occur both in English and in the target language. English MWEs lead to target FEE candidates which are not translations of their English supports. As an example, consider the following example:

Kettensäge	**chain**	ACCOUTREMENT	→	**MWE**

where Kettensäge is most probably a translation of the English multi-word expression **chain saw**. This happens frequently for German compound nouns.

Target-language MWEs lead to FEE candidates which cannot evoke the frame alone. Consider the example

werfen	**raise**	COMM._RESPONSE	→	**MWE**

where the German FEE candidate is missing its particle; it should read **einwerfen** instead.

A.2.4. The Rest: Noise

This leaves an incoherent set of remaining cases, which mostly fall into one of the following categories and should be annotated with `No` (for `Noise`):

- Misalignments not involving proper MWEs. This is rather frequent and involves all kinds of words that co-occur frequently enough:

Frage	**raise**	COMMUNICATION	→	**No**

tief	**veiled**	ACCOUTREMENTS	→	**No**

- Lemmatiser errors

Reisekost	**travel**	TRAVEL	→	**No**

- Cases where the alignment is OK, but the aligned words introduce different concepts (i.e. when the translation is too liberal)

Behauptung	**accusation**	SUSPICIOUSNESS	→	**No**

B. Guidelines for Frame-Semantic Annotation

B.1. Introduction

These guidelines describe the annotation of English and German drawn from the EUROPARL sample with frames and roles according to Berkeley FrameNet.[1]

B.1.1. Frame Semantics

Frames describe *prototypical situations*, including the associated participants and props. For example, the frame REQUEST in the following table describes a situation in which a question or demand is put forward. This frame can be introduced by English words such as *urge, request* and German words such as *fordern, auffordern, Forderung*. These words are called *frame-evoking elements (FEEs)*, and are printed boldface in the examples. Table B.1 shows the participants for REQUEST situations: a SPEAKER, who utters the question; the ADDRESSEE of the question; and the MESSAGE, the part of the sentence expressing the content of the question. The message can be replaced by a TOPIC, a short characterisation of the question's subject area. Finally, the MEDIUM can express the way in which the question was transmitted.

Being based on (sets of) participants, frame semantics is a theory of semantic roles: The frame elements introduced above constitute the roles. They describe the semantic arguments of verbs, nouns, and adjectives, and are specific to frames. This means that Frame Semantics is located

[1] This is an English translation of the original German guidelines given to annotators. It is a simplified variant of the annotation guidelines that were used in the SALSA project at Saarland University (Burchardt et al., 2006b) in early 2005.

FE	Example
Speaker	**Pat** urged me to apply for the job.
Addressee	Pat urged **me** to apply for the job.
Message	Pat urged me **to apply for the job**.
Topic	Kim made a request **about changing her appointment**.
Medium	Kim made a request **in her letter**.

Table B.1.: The frame Request

somewhere between two extremes:

- On one side, there are "classical" thematic role theories. These theories assume a small set of roles such as Agent, Patient, Theme, Experiencer which are applied to the describe the arguments of all verbs equally. However, there is little consensus on the exact set of roles.
- On the other side, some theories have introduced verb-specific roles. This means that every verb introduces its own roleset. While this makes annotation more manageable, this model does not straightforwardly allow any generalisation across word boundaries.

Frame Semantics appears to constitute a realistic "middle way".

B.1.2. FrameNet

The FrameNet project in Berkeley[2] is concerned with constructing a lexicon for the English base vocabulary that provides a frame-semantic description for each word. In particular, FrameNet provides the following resources:

- A list of frames. The description of each frame contains a characterisation of the prototypical situation represented by the frame, a description of the frame elements, like the ones shown in Table B.1, and information about possible syntactic realisations of frame elements in English.

[2]Homepage: `http://www.icsi.berkeley.edu/~framenet/`

- A lexicon of English predicates (predominantly verbs, nouns, and adjectives) that can evoke frames. All appropriate frames are listed for every word, together with some annotated example sentences from the British National Corpus[3] which are chosen to illustrate "typical" usages of particular predicate-frame combinations.

Important: Frame semantics assumes that frames have cross-lingual validity, at least to a certain degree. Thus, most of the frames developed by FrameNet for English can also be used for the annotation of German. Consequently, the SALSA project (which is concerned with frame-semantic annotation for German), has produced (incomplete) lists with German frame-evoking elements for frames, which can be used as guideline for the annotation of the parallel corpus (see Section B.2.2).

Important: Unfortunately, the FrameNet lexicon is incomplete both in terms of words (some words are not listed at all) and in terms of the different word senses (for some words, only rather marginal senses are listed in FrameNet). See also Section B.3.

B.1.3. Syntax

The sentences which are to be annotated with semantic roles have already received a syntactic analysis (constituent structure). A constituent is the set of words dominated by a common node. Figure B.1 shows an example of an English syntax tree. The inner nodes of the syntactic tree are shown as circles, and the words of the sentences are the squares at the bottom (the terminals). *Syntactic categories* are indicated by node labels. For example, the subject of the sentence is the NP (nominal phrase) *And those who T have remained*.

This example already contains a special terminal symbol, namely `T`, which stands for *Trace*. Such nodes appear at places where another constituent would be located, had it not been displaced to some other position in the sentence. This can happen for a variety of phenomena, such as relative clauses, ellipses, coordinations, or control constructions. German syntax trees look similar, but are *flatter* and do not use traces.

[3] `http://www.natcorp.ox.ac.uk/`

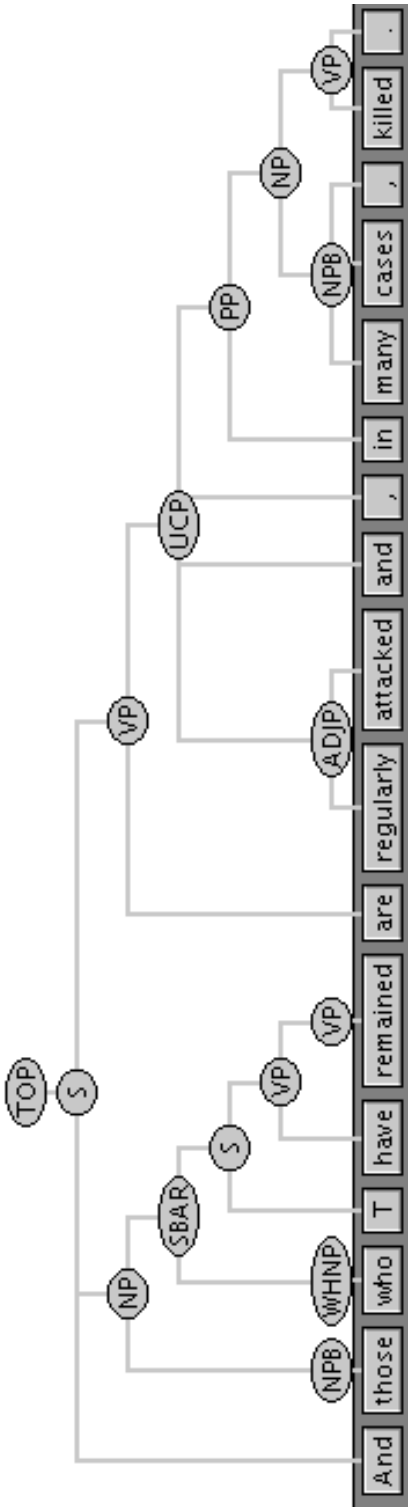


Figure B.1.: English syntactic tree (Collins parser)

Important: To generalise over this difference, traces should *not* be annotated (see Section B.2.3).

Important: The syntactic structure of the corpus was produced automatically and can contain errors. For example, German sentences often have the right local structures, but they are assembled to complete sentences in the wrong way.

Important: Section B.5 contains a list explaining the most important syntactic categories in English and German syntax trees we use.

B.2. Annotation Procedure

B.2.1. Starting the Annotation Tool

The annotation tool is available on all linux servers at the institute. It runs on sun servers as well, but is rather slow (not recommended). To start the tool, enter the following commands:

```
[pado] (../CE/pado) $ cd /proj/llx/Software/SALSA
[pado] (../Software/SALSA) $ ./salsa.sh
```

Then log in with your account name.

B.2.2. Annotation of a New File

Annotation with the annotation tool proceeds file-wise: each individual file is a **subcorpus**. Every subcorpus contains all relevant occurrences of a frame-evoking element (FEE, predicate). Thus, the annotation proceeds **FEE-wise**.

You can load a new file by clicking on the file's symbol in the "work" folder of the "user" window (lower left on the screen). During loading, a warning about the TIGER corpus pops up, which can be ignored.

For every subcorpus, the annotation tool keeps a **list of available frames** that can be annotated in this subcorpus. At the outset, when a new subcorpus is being loaded, the list is empty – the first thing that has to be done is to add frames. There are two main sources of "appropriate" frames for German and English predicates:

- For English words, the FrameNet website lists all known frames for predicates:
  `http://www.icsi.berkeley.edu/~framenet/`
- A list with German words and matching frames can be found here:
  `http://www.coli.uni-saarland.de/~pado/pub/annotation/German_fees_frames.txt`[4]

Important: Due to the methodology of FrameNet, **the list of frames on the FrameNet website is incomplete for virtually every English lemma**. A workaround is to browse the German list for plausible translations of the missing senses, and to use the frames listed there. (Besides, this is a good opportunity to get acquainted with the FrameNet resource and to become aware of the different senses of the current FEEs.)

Technically, adding frames works by opening the "Corpus" menu and activating the menu items "Edit Frames" / "Add Frames". This opens a dialogue window in which frames can be added to the current list. As soon as the list contains at least one frame, you can right-click on any word in the sentence, and the resulting context menu will contain the command "Evoke Frame". Figure B.2 shows the window for subcorpora selection (left) and the dialogue window for adding frames (right).

B.2.3. Annotation of Individual Sentences

Annotation of individual sentences proceeds in three steps:

Step 1: Identification of the FEE

First, the relevant FEE (i.e., the current instance of the lemma that is the name of the subcorpus) has to be identified.

[4] Due to inconsistencies in the FrameNet database, two of the frames in the German list are not accessible through the list of frames on the FN website. However, they can be accessed indirectly through the pages of the FEEs they evoke. This concerns the frames Discussion *(discuss)* and Reliance *(rely)*.

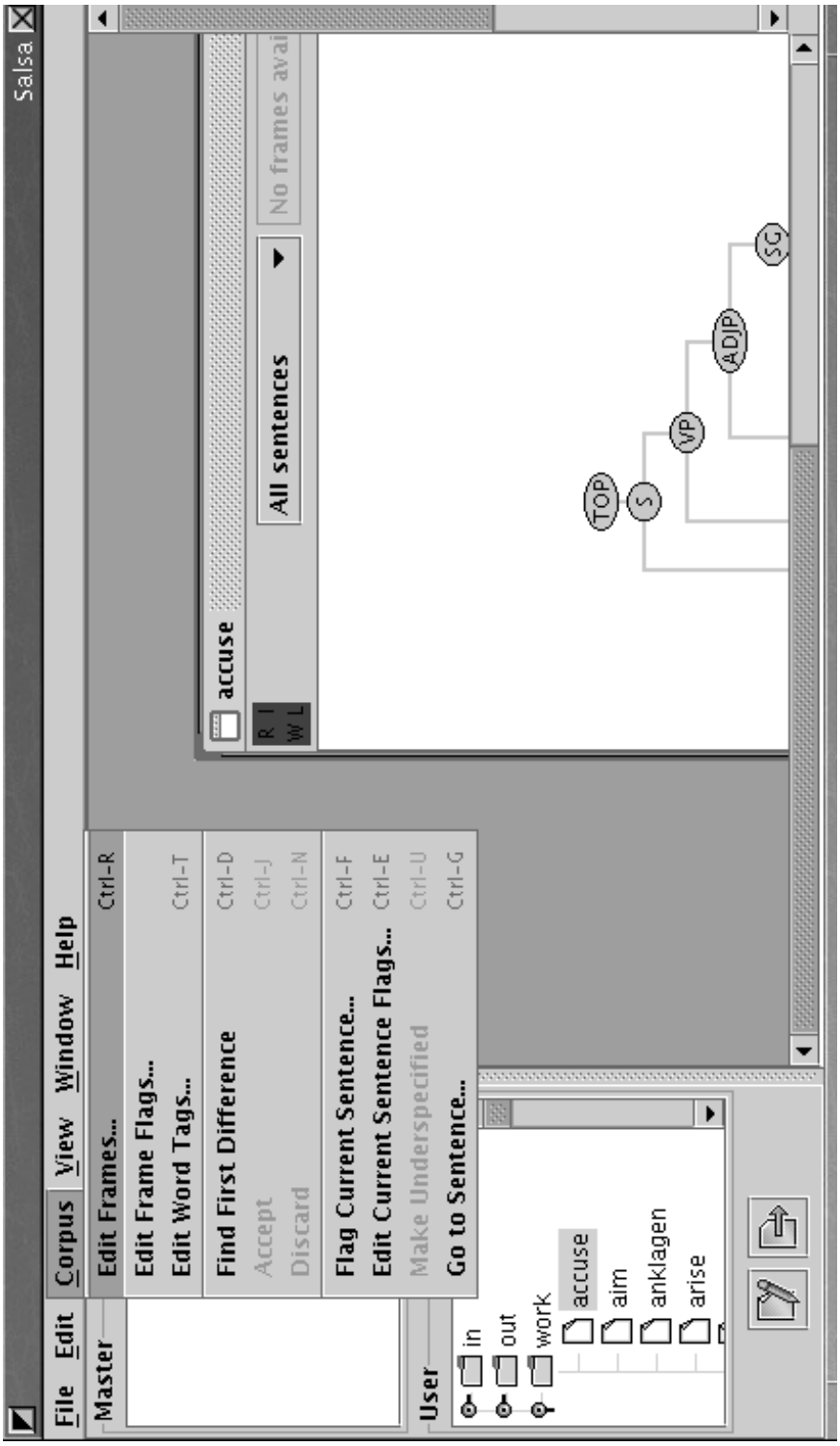


Figure B.2.: Choosing a corpus and adding frames

Step 2: Frame Annotation

The second step consists in choosing the right frame, i.e. determining **which conceptualised situation** is described by the current FEE. In most cases, the appropriate frame can be found in the list of possible frames for that word:

- The list of *lexical units* on the FrameNet webpage (English)
- The list of FEE-frame-combinations on my webpage (German)

(The URLs are given in Section B.2.2 above.)

Often, more than one frame is available for a given word. The following resources can support the decision process:

- The frame description on the FrameNet website which contains a textual definition of the frame's meaning as well as the list of available frame elements for this frame
- The FEE lists mentioned above (can be used for paraphrase tests, see below)
- The annotated FrameNet example sentences (English only)

For example, the German FEE *fragen* can introduce at least the following frames:

- Questioning. FrameNet definition: The words in this frame have to do with a Speaker asking an Addressee a question which calls for a reply (as opposed to making a request which calls for an action on the part of the Addressee).
  - Example: Vor vielen Jahren schwärmte ich von diesem ersten meiner Lieblingsfilme, als einer der Zuhörer mich **fragte**, ob Fellinis Film Einfluss gehabt habe auf die Titelgebung eines Stueckes – The Road – , das gerade seine Uraufführung erlebt hatte .
- Request. FrameNet definition: In this frame a Speaker asks an Addressee for something, or to carry out some action.

  - Example: Immer wieder **fragt** die junge Frau nach einem Arzt, besteht darauf, untersucht zu werden .

- COGITATION. FrameNet definition: A person, the Cognizer, thinks about a Topic over a period of time. What is thought about may be a course of action that the person might take, or something more general.

  - Example: Das **frage** ich mich manchmal selbst.

Some more detailed hints:

- It is generally a good idea for FEEs which can introduce more than one frame to try and paraphrase the sentence with FEEs which can only evoke a single frame f. If the sentence cannot be paraphrased with FEEs particular to f, it is highly probable that f is not a good description for the sentence.

- The list of a frame f's frame elements provided by FrameNet can be compared to the subcategorisation of the current predicate. For example, if the current predicate can have an object (realised or unrealised) which does not correspond to any frame element of f, f is not appropriate. Using this argumentation, we can exclude the frame FEELING for the sentence "Sie empfanden die Entscheidung als unfair" since "die Entscheidung" cannot be mapped onto any frame element of FEELING.

- There are cases in which it is very difficult to make a decision between two frames, since they correspond to **different meaning components** of the current sentence. For example, consider the frames AWARENESS (to have a fact in his/her mental representation, as a belief or as knowledge) and CERTAINTY (to be certain of a fact). The sentence "Ich glaube, dass X", without additional context, has elements of both meaning components. In such a case, it helps to:

  1. Browse the annotated examples for either frames and check whether examples which work analogously to the current sentence are annotated for one frame only. (i.e. if one of the meaning components is classified by FrameNet as purely inferred.)

2. If this is not the case, a decision should be made which of the meaning components is **dominant** in the example at hand. If the example "ich glaube X" focuses on expressing the **content** of the belief/knowledge (like "ich weiss"), AWARENESS is more appropriate; if the main information is about the **degree of certainty** of the belief (like "ich bin sicher, dass"), it is a case of CERTAINTY.

Unfortunately, these are case-to-case decisions for which no guidelines can be provided and which are completely in the annotator's power. Section B.6 lists the most notoriously difficult frame distinctions and provides paraphrases of the meaning components the individual frames emphasise.

On a technical level, right clicking onto the word node and then selecting "Invoke Frame" shows a list of all available frames for the current subcorpus (cf. Section B.2.2; if "Invoke frames" is not displayed in the menu, no frames have been made available for the subcorpus yet); clicking on a frame name then finally introduces the frame.

Important: Frame assignment can also be difficult for other reasons, the most important ones being gaps in FrameNet (no suitable frame is available), and multi-word expressions (the meaning arises from the combination of several words at once). Section B.3 discusses these cases.

Step 3: Annotation of Frame Elements

As soon as a frame has been chosen, the annotation tool displays it as a rectangle above the sentence that is "anchored" to the frame-evoking element, and the frame elements are shown as smaller rectangles hovering around the frame. Frame elements are assigned by dragging them onto a syntactic constituent and releasing. Before releasing, the tool highlights in blue the subtree corresponding to the current constituent, i.e. the part of the sentence the role is being assigned to.

Fundamentally, all frame elements which can be recognised **with certainty** should be annotated. Sometimes, several plausible annotation options exist and it is difficult to choose one over the other. The following criteria can be used to decide such cases:

- **Annotate maximally and locally.** In principle, *exactly* the lexical material that belongs to some role should be annotated as part of it. Optimally, the complete lexical material is located below one, the so-called *maximal*, constituent. Problems arise through sentential constructions that allow more than one expression to refer to the same semantic entity. Examples of such constructions are relative pronouns, relative clauses, pronouns, and so on. As an example, consider Figure B.1, where the victims that have been attacked are referred to by three different expressions: a pronoun (*those*), a relative pronoun (*who*), and an empty element or trace (`T`), plus the constituents dominating these words. Which is the correct annotation choice?

  To answer this question, apply the **locality principle**:

  > Search from the FEE outward in the tree, and annotate the first suitable constituent, unless it is a trace.

  In more concrete terms, this means first testing whether the sister nodes of the FEE are potential FEs. If not, test the sister nodes' children and do so recursively. Only if none of these is suitable, climb one node up from the FEE and test its sisters and their descendants. Repeat until suitable constituent is found (or FE remains unaligned). Put more simply, choose the suitable constituent *with the shortest path to the FEE*.

  In the example in Figure B.1, that would be the NP with the yield "those who T have remained", and whose path to the FEE is ADJP-UCP-VP-S-NP.

  This example demonstrates one important side effect of the locality principle alluded to in the first sentence of this paragraph, namely maximality: choose the complete lexical material belonging to a node. This is enforced by locality, since all syntactic subtrees are traversed in a top-down sequence, so that nodes higher up (maximal constituents) are more likely to be chosen.

- **Traces cannot act as frame elements** The only exception to the locality principle is the case of "empty" elements. As an example, consider again the frame Self_Motion in Figure B.1. Here, the

trace "T" is a placeholder for the entity that moves (the role of SELF_MOVER); but traces should not be annotated. In the example, the two other realisations of the SELF_MOVER are "who" and "those". Since "who" is closer to the FEE"remain"than "those", the relative pronoun "who" is annotated as FE.

- **Every frame element occurs at most once.** No frame element occurs necessarily; there may be frames which don't have a single realised frame element (example for FEE "eat": "Eating is fun."). On the other hand, multiple occurrences of frame elements are licensed only in quite specific syntactic conditions, namely when the lexical material of a role is distributed over several syntactic constituents (see next item).

- **Errors in the syntactic structure.** Since the syntactic structure was created automatically, it can contain errors. Figure B.3 shows the relevant bits of an example sentence for the FEE "accused": "Mr President" is just an address and does not semantically form part of the role of ACCUSED. This leads to a conflict between two principles: on one hand, maximal constituents should be annotated; on on the other, exactly the relevant lexical material should be annotated.

  In such cases, the "maximal constituent" criterion is demoted, and consequently the complete NP is not annotated as ACCUSED. Instead, the two words "country" and "rapporteurs" have to be annotated individually with the same role. On a technical level, this can be achieved by dragging the FE box over the first word to assign it. A further right click on the frame box makes a context menu appear with an entry called "Add Element/<FE name>". Choosing this entry adds a new instance of the FE which can then be assigned to the second word, and so on. Figure B.3 shows the resulting annotation.

Important: Section B.4 contains further instructions for the annotation of problematic frame elements.

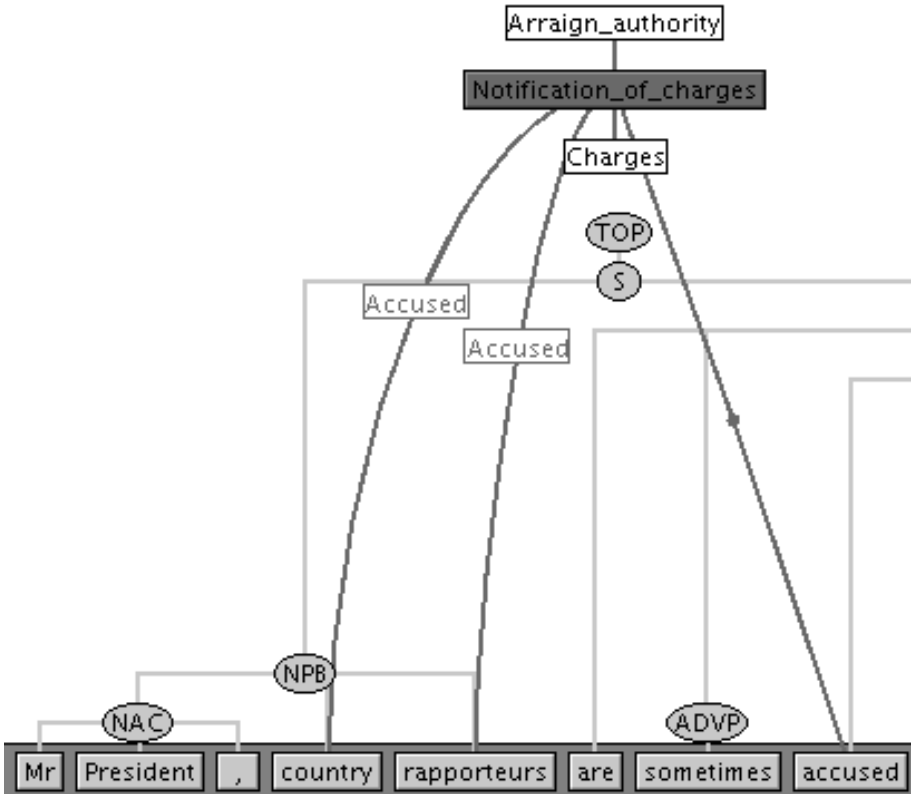


Figure B.3.: Error in the syntactic structure

B.2.4. Finishing the Annotation of a File

When a file is annotated completely, save it first. Then move the file to the "out" folder. To do so, left click on the "document arrow" symbol at the bottom left (let the mouse hover over buttons for about two seconds to get a description).

Once a corpus has been moved to the "out" folder, it cannot be modified any more. At the same time, this means for me that I can start evaluating. Therefore, please put only properly finished files into the "out" folder.

B.3. Linguistic difficulties – Frame Choice

As described in Section B.2.2, the frame-predicate lists for English and German may propose more than one frame for a give predicate, and a

decision has to be made. However, it is also possible that **none of the proposed frames is appropriate**. There are two main reasons for this:

B.3.1. Gaps in FrameNet

We assume here that the predicate is an individual word which introduces some meaning by itself and is understood literally (for more information on multi-word expressions and metaphors, see Section B.3.3). However, FrameNet does not list the evoked meaning. An example of this type is the English verb *have*, for which FrameNet lists only GIVING_BIRTH, as in *have a child*. Clearly, this does not do for sentences such as *He had three dollars*.

There are two possible scenarios: either a frame exists for this meaning, and it is only the lexical unit that is missing in the frame (*lexical gap*), or no frame exists in FrameNet which describes this meaning (*conceptual gap*). These two cases can be distinguished by looking up related cases in the provided resources, either monolingually by searching synonyms in the FrameNet predicate list, or cross-lingually by searching translations in the SALSA predicate lists. If synonyms and/or translations exist and evoke a frame that is appropriate, that frame can be used (lexical gap). If, however, there appears to be a conceptual gap in that there aren't any synonyms or translations, please add a new frame called UNKNOWN (uppercase U, lowercase nknown) for the current predicate (menu entry "Corpus/Edit Frames/Add Frame") and annotate it.

The example above (*He had three dollars*) can be handled by finding the synonyms *possess, own*, which lead to the frame POSSESSION; this is the case of a lexical gap. In contrast, there is currently no FrameNet frame for the "social behaviour" sense of *treat* as in *He treated the conductor badly.*, as can be seen by the absence of synonyms like *behave*. This is a real conceptual gap.

Important: For lexical gaps, it is especially important to check if the frame identified through synonyms is really appropriate; maybe it is not listed for the predicate for a reason. A specific problem in this respect are occasional differences between the *definition* of a FrameNet frame and the annotated example sentences. Sometimes, the definition is *narrower* and does not even license the annotated example sentences. For example, the definition of DEPARTING states that the frame is only meant to describe

physical movements; however, the annotation set also includes examples such as *He had WITHDRAWN behind his shield of self-possession*, a clear metaphorical usage. When in doubt, the general tendency for the purposes of the present annotation is to **interpret the frame definition more liberally, i.e., in line with the annotated example sentences**. This makes it possible to annotate as many EUROPARL instances as possible using FrameNet frames instead of Unknown.

B.3.2. Frames: Definitions vs. Example Sentences

For some frames (especially frames about movement and/or cognitive processes), there is a mismatch between the frame definition and the example sentences. Usually, the frame definition refers to very concrete instances and appears to restrict the use of the frame to literal instances of physical movement and human cognition/perception. In contrast, the lexical units of such frames often appear in more abstract or "generalised" instances. Examples are

- John's freedom **disappeared**
- Let's **look** ahead!
- to **distance** oneself from something

All of these cases are not "literal" usages of the respective frames, but are annotated as instances in FrameNet nevertheless. In the present annotation, **generalised uses are to be annotated with matching frames whenever they are backed by example sentences even if the frame definition is phrased more restrictively.**

B.3.3. Multi-word Expressions and the Problem of Readings

There are easy cases, and there are difficult cases. Whenever the frame is introduced by the literal reading of a single word (such as *fire*) plus possibly a particle (*pick up*) or a reflexive pronoun (*sich fragen*), frame annotation is usually (comparatively) easy, despite missing frames in FrameNet.

In contrast, multi-word expressions (MWEs) which introduce frames are often troublesome. They can be grouped into three broad categories: support verbs, metaphors, and idioms. They can be distinguished primarily by the differences in the relationship between their *literal or compositional* reading and their *nonliteral or understood* reading.

- **Support verbs**: These are usually the combination of a noun (or a prepositional phrase), which is the "semantic center" of the expression, and a verb which only contributes a specific nuance of the meaning, e.g., perspectivisation or aspect.

  – Examples: Risiko scheuen/eingehen, to undergo/perform an operation

  One criterion for recognising support verbs is that the meaning of the "semantic center" and the complete expression are (almost) the same:

  – Example: Peter underwent an operation – Peter's operation

- **Metaphors**: Metaphors are expressions which have a literal as well an understood meaning, and where both levels are recognisable.

  – Examples: Perot läuft im weissen Haus gegen eine Wand, Perot walks against a brick wall in the White House

  Here, the literal meaning "to walk against a wall" can be described using the frame MOTION; the understood meaning "not succeed, get stuck" introduces the frame ENCOUNTER_OBSTACLE. Knowledge of the literal frame helps to understand the understood frame.

- **Idioms**: Similar to metaphors, idioms have an "understood" meaning; however, the relation to the original literal meaning is not present any longer. Often, idioms are "fixed" expressions which cannot be modified, while metaphors (as the example above shows) can be made up on the fly.

  – Examples: zugrunde gehen, to push up the daisies

This is evidently an instance of the DYING frame; even though, diachronically speaking, there will have been a "literal" meaning evoking the frames MOTION and CAUSE_MOTION, respectively, these are not necessary to understand the intended meaning - they do not even help.

Such multi-word expressions are problematic especially because they are very difficult to distinguish from one another: Even though there are clear-cut cases, there are also large "grey areas".

- Is *unter Druck setzen / to put under pressure* still a support verb construction, since the semantics is determined mostly by "under pressure", and the verb contributes only the aspect? Or is it already an idiom?

- Is *Flagge zeigen ("to show one's flag")* in the sense of "to show one's loyalty" an idiom, or a metaphor? Annotators who have a picture of a ship showing its flag in mind will presumably argue for the metaphor analysis, others might tend towards idiom.

For the purpose of the present annotation, therefore, all multi-word expressions are treated in the same manner:

1. Decision: It is a single-word FEE or a multi-word FEE?

    - Single-word FEEs are annotated with their normal frame, according to the standard rules
    - All multi-word FEEs are, in principle, to be annotated with their **understood** meaning. Note that the search for understood meanings can take a long time, since neither the English nor the German frame lists are necessarily helpful. Often, it helps to come up with **paraphrases of the understood meaning** which are themselves "literal" and can be looked up in FrameNet.

      Be aware that choosing a frame is the most important part of your annotation – and in the case of multi-word expressions, definitely the most difficult. Therefore, please take extra care to be consistent in your decisions.

If no frame that applies can be found, or if you run out of time (I recommend that you do not spend more than a quarter of an hour on any single instance), please annotate Unknown (cf. Section B.3.1).

B.4. Linguistic difficulties – Frame Elements

B.4.1. Which Frame Elements to Annotate?

It is possible for more frame elements show up in the description of a frame on the FrameNet web page or its annotated example sentences than are available for actual annotation in the annotation tool. The reason for this is that FrameNet distinguishes

- **Core** frame elements which express central concepts of the situation in question, and are therefore frame-specific, and
- **Additional** frame elements such as location, time, etc., which are applicable to a large range of situations, and are thus not specific to particular frames

This distinction is shown in the FrameNet definitions of individual frame elements, which are defined as either **core** or **peripheral/extrathematic**. Only core elements are made available for annotation by the annotation tool; and only core elements are to be annotated. This means that Time, Place, Means, Manner, are not to be annotated. Exceptionally, Place, Direction, and Path are core elements of movement-type frames since they express central concepts of movements.

B.4.2. Certain and Uncertain Frame Elements

As mentioned in Section B.2.3, only frame elements recognised as certain are supposed to be annotated. At times, this is an issue of interpretation, in particular when it comes to nouns, but also for more complex verbal constructions. Consider the following example for the Frame Commerce (FEE **kaufen**):

- Example: Besitzer von Einfamilienhäusern, [NP die vor 1994 [VP **gekauft** haben]]

In this sentence, the author has not realised the direct object of *gekauft*, the thing bought. It is rather probable that the goods in question are the single-family homes, but this interpretation is defeasible: In a sufficiently specific context, for example one that talks about the stock market and its interaction with the housing market, this interpretation can be overridden, and the sentence can be interpreted as talking about the purchase of stocks.

Guideline: FEs are to be annotated if they are reasonable in neutral context, i.e. if they are instantaneously understood. In the case above, *Einfamilienhäuser* can therefore be annotated as the Goods FE of Commerce.

B.4.3. Particular Syntactic Constructions in German

There are a couple of syntactic constructions in German which may lead to problems in the assignment of semantic roles and which are therefore discussed here. This list may not be complete.

- ***mit* constructions.** Consider the following example:

  - Example: Mit Donnerstag, 9 Uhr, wurde ein unglücklicher Termin gewählt.

  Is *Mit Donnerstag, 9 Uhr,* a frame element? We do not treat it as such – *mit* phrases are mostly best seen as *depictives*, complements that provide a more detailed description of the proper frame element *ein unglücklicher Termin*. This can also be seen by considering a possible paraphrase:

  - Example: Es wurde ein unglücklicher Termin gewählt, und zwar Donnerstag, 9 Uhr.

  In this form, it is even clearer that the *mit* phrase is just an addition.

- **Reflexives.** There are different uses of reflexives in German. If they realise proper arguments of predicates they have to be annotated;

if not, not. The "argumenthood" can be tested by replacing the reflexive pronoun with another NP: if this is possible, the reflexive fills a proper argument slot.

- Example: Er **rasiert** sich $\longrightarrow$ Er **rasiert** Peter.

This works, therefore *sich* fills a frame element.

- Example: Das Brot **verkauft** sich gut $\not\longrightarrow$ Das Brot **verkauft** ihn gut.

This does not work – sich is not a proper argument. Note that sometimes the replacement works, but the meaning changes in the process. This is taken as negative evidence for argumenthood:

- Ich frage mich, ob X. (Cogitation) $\not\longrightarrow$ Ich frage Peter, ob X. (Questioning)

B.5. Reference: Important Syntactic Categories

English

- ADJP: adjectival phrase
- ADVP: adverbial phrase
- NP(B): nominal phrase
- PP: prepositional phrase
- S(BAR): sentence
- VP: verbal phrase
- WHNP: wh-question phrase

German

- AP: adjectival phrase
- NP: nominal phrase
- PN: proper noun
- PP: prepositional phrase
- S: sentence
- VP: verbal phrase
- VZ: infinitival phrase with "zu"

In addition, the German trees contain categories with names of the form "C<x>". These are "coordinating" nodes, i.e. nodes which directly dominate two nodes of the category "<x>", typically linked by an "and" or "or". For example, "CS" is a sequence of two conjoined sentences ("S").

B.6. Reference: (Some) Difficult Frame Distinctions

- **Achieving_first / Invention**: **Achieving_first** covers, in addition to feats achieved for the first time, inventions of concrete artifacts. **Invention** focuses on conceptual breakthroughs.
- **Awareness / Certainty**: **Awareness** stresses the content of a believing/knowing/thinking event; **Certainty** stresses the degree of (un-) reliability.
- **Getting / Receiving**: In **Receiving** events, the transfer is initiated volitionally by the **Giver**.
- **Hostile_Encounter / Quarreling**: If the situation is an obvious communication situation in which the participants exchange verbal arguments, **Quarreling** is the right frame; all other cases are covered by **Hostile_Encounter**.

- **Judgment_Communication / Notification of Charges**: **Notification_of_Charges** refers specifically to the situation in the judicial court; **Judgment_Communication** covers all verbal accusations.

Bibliography

Abney, Steven. 1996. Partial parsing via finite-state cascades. *Journal of Natural Language Engineering* 2(4):337–344.

Agirre, Eneko, and Philip Edmonds, eds. 2006. *Word Sense Disambiguation: Algorithms and Applications*. Text, Speech and Language Technology 33, Springer.

Althaus, Ernst, Nikiforos Karamanis, and Alexander Koller. 2004. Computing locally coherent discourses. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, 399–406. Barcelona, Spain.

Baeza-Yates, Ricardo A., and Berthier A. Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Baker, Collin F., Michael Ellsworth, and Katrin Erk. 2006. Frame Semantic Structure Extraction: Description of the SEMEVAL 2007 task. `http://nlp.cs.swarthmore.edu/semeval/tasks/task19/description.pdf`.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, 86–90. Montreal, QC.

Bannard, Colin, and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 597–604. Ann Arbor, MI.

Barzilay, Regina, and Lillian Lee. 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing*, 164–171. Philadelphia, PA.

———. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the joint Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 16–23. Edmonton, AL.

Baumgartner, Peter, and Aljoscha Burchardt. 2004. Logic Programming Infrastructure for Inferences on FrameNet. In *Logics in Artificial Intelligence, JELIA 2004*, vol. 3229 of *Lecture Notes in Artificial Intelligence*, 591–603. Springer Verlag.

Bentivogli, Luisa, and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Journal of Natural Language Engineering* 11(3): 247–261.

Boas, Hans C. 2002. Bilingual FrameNet dictionaries for machine translation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands.

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

Bollobás, Béla. 1998. *Modern graph theory*. Graduate Texts in Mathematics, Springer.

Bourigault, Didier, Cécile Fabre, Cécile Frérot, Marie-Paule Jacques, and Sylwia Ozdowska. 2005. Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*. Dourdan, France.

Bourigault, Didier, and Cécile Frérot. 2005. Acquisition et évaluation sur corpus de propriétés de sous-catégorisation syntaxique. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*. Dourdan, France.

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

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics* 16(2):79–85.

Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2):263–311.

Buchholz, Sabine, and Erwin Marsi, eds. 2006. *Proceedings of the CoNLL shared task: Multilingual dependency parsing*. New York City, NY.

Budanitsky, Alexander, and Graeme Hirst. 2001. Semantic distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In *Proceedings of the ACL Workshop on WordNet and Other Lexical Resources*, 29–34. Pittsburgh, PA.

Burchardt, Aljoscha, Katrin Erk, and Anette Frank. 2005a. A WordNet Detour to FrameNet. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, vol. 8 of *Computer Studies in Language and Speech*, 408–421. Frankfurt, Germany: Peter Lang.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006a. Consistency and coverage: Challenges for exhaustive semantic annotation. Presentation given at 28. Jahrestagung, Deutsche Gesellschaft für Sprachwissenschaft. Bielefeld, Germany.

———. 2006b. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

Burchardt, Aljoscha, and Anette Frank. 2006. Approaching Textual Entailment with LFG and FrameNet frames. In *Proceedings of the 2nd Workshop on Recognising Textual Entailment*. Venice, Italy.

Burchardt, Aljoscha, Anette Frank, and Manfred Pinkal. 2005b. Building text meaning representations from contextually related frames – a case study. In *Proceedings of the 6th International Workshop on Computational Semantics*. Tilburg, The Netherlands.

Burnard, Lou. 1995. *User's guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Services.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. In *Proceedings of the COLING Workshop on Grammar Engineering and Evaluation*, 1–7.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics* 22(2):249–254.

Carlson, Greg. 1984. Thematic roles and their role in semantic interpretation. *Linguistics* 22:259–279.

Carreras, Xavier, and Lluís Màrquez, eds. 2004. *Proceedings of the CoNLL shared task: Semantic role labelling*. Boston, MA.

———. 2005. *Proceedings of the CoNLL shared task: Semantic role labelling*. Ann Arbor, MA.

Charniak, Eugene. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on Artificial Intelligence*, 598–603. Providence, RI.

Chklovski, Timothy, and Patrick Pantel. 2004. VerbOcean: Mining the Web for fine-grained semantic verb relations. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, 33–40. Barcelona, Spain.

Chomsky, Noam. 1957. *Syntactic Structures*. MIT Press.

———. 1981. *Lectures on Government and Binding*. Dordrecht, The Netherlands: Foris Publications.

Cohen, Kevin B., and Lawrence Hunter. 2006. A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics* 7(Suppl. 3):S5.

Cohn, David A., Les Atlas, and Richard E. Ladner. 1994. Improving generalization with active learning. *Machine Learning* 15(2):201–221.

Collins, Michael. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 16–23. Madrid, Spain.

Collins, Michael, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 531–540. Ann Arbor, MI.

Cormen, Thomas, Charles Leiserson, and Ronald Rivest. 1990. *Introduction to algorithms*. Cambridge, MA: MIT Press.

Cyrus, Lea. 2006. Building a resource for studying translation shifts. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

Davis, Tony. 1996. Lexical semantics and linking in the hierarchical lexicon. Ph.D. thesis, Stanford University.

Dempster, Arthur, Nan Laird, and Donald Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39(1):1–22.

Diab, Mona, and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 255–262. Philadelphia, PA.

Dietterich, Thomas G. 2000. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, 1–15. London, UK: Springer-Verlag.

Dorna, Michael, and Martin C. Emele. 1996. Semantic-based transfer. In *Proceedings of the 16th International Conference on Computational Linguistics*, 316–321. Copenhagen, Denmark.

Dorr, Bonnie. 1995. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics* 20(4):597–633.

Dowty, David. 1989. On the semantic content of the notion of 'thematic role'. In *Property Theory, Type Theory, and Natural Language Semantics*, ed. G. Chierchia, B. Partee, and R. Turner, 69–130. Dordrecht: Reidel.

———. 1991. Thematic proto-roles and argument selection. *Language* 67:547–619.

Dubey, Amit. 2004. Statistical parsing for german: Modelling syntactic properties and annotation differences. Ph.D. thesis, Saarland University.

———. 2005. What to do when lexicalization fails: Parsing German with suffix analysis and smoothing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 314–321. Ann Arbor, MI.

Eiter, Thomas, and Heikki Mannila. 1997. Distance measures for point sets and their computation. *Acta Informatica* 34(2):109–133.

Ellsworth, Michael, Katrin Erk, Paul Kingsbury, and Sebastian Padó. 2004. PropBank, SALSA and FrameNet: How design determines product. In *Proceedings of the LREC Workshop on Building Lexical Resources From Semantically Annotated Corpora*. Lisbon, Portugal.

Erk, Katrin. 2005. Frame assignment as word sense disambiguation. In *Proceedings of the 6th International Workshop on Computational Semantics*. Tilburg, The Netherlands.

———. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the joint Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*. New York City, NY.

Erk, Katrin, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2003. Towards a resource for lexical semantics: A large German corpus with extensive semantic annotation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 537–544. Sapporo, Japan.

Erk, Katrin, and Sebastian Padó. 2005. Analysing models for semantic role assignment using confusability. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 668–675. Vancouver, BC.

———. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

Fillmore, Charles J. 1968. The case for case. In *Universals in linguistic theory*, ed. Bach and Harms, 1–88. New York: Holt, Rinehart, and Winston.

———. 1982. Frame Semantics. In *Linguistics in the morning calm*, 111–138. Seoul, Korea: Hanshin.

———. 1985. Frames and the semantics of understanding. *Quaderni di Semantica* IV(2):222–254.

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

Fillmore, Charles J., Charles Wooters, and Collin F. Baker. 2001. Building a large lexical databank which provides deep semantics. In *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*, 3–26. Hong Kong.

Fleischman, Michael, and Eduard Hovy. 2003. Maximum entropy models for FrameNet classification. In *Proceedings of the 8th Conference on Empirical Methods in Natural Language Processing*, 49–56. Sapporo, Japan.

Frank, Anette, and Katrin Erk. 2004. Towards an LFG Syntax-Semantics Interface for Frame Semantics Annotation. In *Proceedings of the Fifth International Conference on Intelligent Text Processing and Computational Linguistics*. Seoul, Korea.

Frank, Anette, Hans-Ulrich Krieger, Feiyu Xu, Hans Uszkoreit, Berthold Crysmann, Brigitte Jörg, and Ulrich Schäfer. 2007. Question answering

from structured knowledge sources. *Journal of Applied Logic* 5(1):20–48.

Fredman, Michael L., and Robert E. Tarjan. 1987. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM* 34(3):596–615.

Fung, Pascale, and Benfeng Chen. 2004. BiFrameNet: Bilingual frame semantics resources construction by cross-lingual induction. In *Proceedings of the 20th International Conference on Computational Linguistics*, 931–935. Geneva, Switzerland.

Gale, William A., and Kenneth Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1):75–102.

Geertzen, Jeroen. 2003. String alignment in grammatical inference: what suffix trees can do. Master's thesis, ILK, Tilburg University, Tilburg, the Netherlands.

Gildea, Daniel. 2001. Corpus variation and parser performance. In *Proceedings of the 6th Conference on Empirical Methods in Natural Language Processing*, 167–172. Pittsburgh, PA.

———. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 80–87. Sapporo, Japan.

———. 2004. Dependencies vs. constituents for tree-based alignment. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, 214–221. Barcelona, Spain.

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

Giuglea, Ana-Maria, and Alessandro Moschitti. 2004. Knowledge discovery using FrameNet, VerbNet and PropBank. In *Proceedings of the Workshop on Ontology and Knowledge Discovery at the 15th European Conference on Machine Learning*. Pisa, Italy.

———. 2006. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, 929–936. Sydney, Australia.

Green, Rebecca, and Bonnie Dorr. 2004. Inducing a semantic frame lexicon from WordNet data. In *Proceedings of the ACL Workshop on text meaning and interpretation*. Barcelona, Spain.

Green, Rebecca, Bonnie Dorr, and Philip Resnik. 2004. Inducing frame semantic verb classes from WordNet and LDOCE. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, 375–382. Barcelona, Spain.

Grenager, Trond, and Christopher Manning. 2006. Unsupervised discovery of a statistical verb lexicon. In *Proceedings of the 11th Conference on Empirical Methods in Natural Language Processing*, 1–8. Sydney, Australia.

Gruber, Jeffrey S. 1965. Studies in lexical relations. Ph.D. thesis, Massachusetts Institute of Technology.

Hajičová, Eva. 1998. Prague Dependency Treebank: From analytic to tectogrammatical annotation. In *Proceedings of TSD 1998*, 45–50. Brno, Czech Republic.

———. 2000. Dependency-based underlying-structure tagging of a very large Czech corpus. *T.A.L.* 41(1):47–66.

Hakkani-Tür, Dilek, Gokhan Tur, and Ananlada Chotimongkol. 2004. Using semantic and syntactic graphs for call classification. In *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, 24–31. Barcelona, Spain.

Hawkins, John A. 1986. *A comparative typology of English and German*. London and Sydney: Croom Helm.

Hi, Chenhai, and Rebecca Hwa. 2005. A backoff model for bootstrapping resources for non-English languages. In *Proceedings of the joint*

Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 851–858. Vancouver, BC.

Hutchins, John W., and Harold L. Somers. 1992. *An Introduction to Machine Translation*. Academic Press.

Hwa, Rebecca, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluation translational correspondance using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 392–399. Philadelphia, PA.

Hwa, Rebecca, Philipp Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Special Issue of the Journal of Natural Language Engineering on Parallel Texts* 11(3):311–325.

Imamura, Kenji. 2001. Hierarchical phrase alignment harmonized with parsing. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, 377–384. Tokyo, Japan.

Johansson, Richard, and Pierre Nugues. 2006. A FrameNet-Based Semantic Role Labeler for Swedish. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, 436–443. Sydney, Australia.

Johansson, Richard, David Williams, Anders Berglund, and Pierre Nugues. 2004. Carsim: A system to visualize written road accident reports as animated 3D scenes. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, 57–64.

Jonker, R., and A. Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* 38:325–340.

Kaji, Hiroyuki, Yuuko Kida, and Yasutsugu Morimoto. 1992. Learning translation templates from bilingual text. In *Proceedings of the 14th International Conference on Computational Linguistics*, 672–678. Nantes, France.

Kanamaru, Toshiyuki, Masaki Murata, Kow Kuroda, and Hitoshi Isahara. 2005. Obtaining Japanese lexical units for semantic frames from Berkeley FrameNet using a bilingual corpus. In *Proceedings of the 6th International Workshop on Linguistically Interpreted Corpora*, 11–20. Jeju Island, Korea.

Karp, Richard M. 1972. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, ed. R.E. Miller and J.W. Thatcher. New York, NY: Plenum Press.

Katz, Jerrold J., and Jerry A. Fodor. 1964. The structure of a semantic theory. In *The structure of language*, ed. Jerrold J. Katz and Jerry A. Fodor. Prentice-Hall.

Kay, Paul, and Charles J. Fillmore. 1999. Grammatical constructions and linguistic generalizations: the What's X doing Y? construction. *Language* 75:1–33.

Keenan, Edward L., and Bernard Comrie. 1977. Noun phrase accessibility and universal gramamr. *Linguistic Inquiry* 8:63–99.

Kilgarriff, Adam, and Joseph Rosenzweig. 2000. Framework and results for English Senseval. *Computers and the Humanities* 34(1–2):15–48.

Kingsbury, P., and K. Kipper. 2003. Deriving verb-meaning clusters from syntactic structure. In *Proceedings of the HLT/NAACL Workshop on Text Meaning*. Edmonton, Canada.

Kipper, K., M. Palmer, and O. Rambow. 2002. Extending PropBank with VerbNet semantic predicates. In *Workshop on applied interlinguas at AMTA 2002*. Tiburon, CA.

Klein, Dan, and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, 478–485. Barcelona, Spain.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the joint Human Language Technology Conference and Annual Meeting of the North Amer-*

ican Chapter of the Association for Computational Linguistics, 48–54. Edmonton, AL.

Koehn, Phillip. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit X*. Phuket, Thailand.

Koeling, Rob, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 419–426. Vancouver, BC.

Kuhn, Jonas. 2004. Experiments in parallel-text based grammar induction. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, 470–477. Barcelona, Spain.

Kučera, Henry, and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Lapata, Mirella, Frank Keller, and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgements. In *Proceedings of acl 39th Annual Meeting of the Association for Computational Linguistics*, 354–361. Toulouse, France.

van Leuven-Zwart, Kitty M. 1989. Translation and original: Similarities and dissimilarities. *Target* 1(2):151–181.

Levin, Beth. 1993. *English verb classes and alternations*. Chicago, IL: University of Chicago Press.

Lopatková, Marketa, and Jarmila Panevová. 2005. Recent developments in the theory of valency in the light of the Prague Dependency Treebank. In *Insight into Slovak and Czech corpus linguistic*, ed. M. Šimková. Bratislava, Slovakia: Veda.

Manning, Christopher, and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2):313–330.

Markou, Markos, and Sameer Singh. 2003. Novelty detection: A review, part i: Statistical approaches. *ACM Signal Processing* 83:2481–2497.

Matsumoto, Yuji, Hiroyuki Ishimoto, and Takehito Utsuro. 1993. Structural matching of parallel texts. In *Proceedings of acl 31st Annual Meeting of the Association for Computational Linguistics*, 23–30. Columbus, OH.

Matthiessen, Christian M.I.M. 2001. The environments of translation. In *Exploring translation and multilingual text production: Beyond content*, ed. Erich Steiner and Colin Yallop, 41–124. Text, Translation, Computational Processing, Berlin: Mouton de Gruyter.

Matusov, Evgeny, Richard Zens, and Hermann Ney. 2004. Symmetric word alignments for statistical maching translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, 219–225. Geneva, Switzerland.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics*, 279–286. Barcelona, Spain.

Melamed, I. Dan. 1996. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas*. Montreal, Canada.

———. 1997. Measuring semantic entropy. In *Proceedings of the ANLP SIGLEX Workshop on tagging text with lexical semantics*. Washington, DC.

———. 1998a. Annotation style guide for the Blinker project. Tech. Rep. IRCS TR #98-06, IRCS, University of Pennsylvania.

———. 1998b. Manual annotation of translational equivalence: The Blinker project. Tech. Rep. IRCS TR #98-07, IRCS, University of Pennsylvania.

———. 2000. Models of translational equivalence among words. *Computational Linguistics* 2(23):221–249.

Merlo, Paola, and Suzanne Stevenson. 2001. Automatic Verb Classification based on Statistical Distribution of Argument Structure. *Computational Linguistics* 27(3):373–408.

Meyers, Adam, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating Noun Argument Structure for NomBank. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Meyers, Adam, Roman Yangarber, and Ralph Grishman. 1996. Alignment of shared forests for bilingual corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*, 460–465. Copenhagen, Denmark.

Mihalcea, Rada, and Phil Edmonds, eds. 2004. *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain.

Miller, George, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on WordNet. *International Journal of Lexicography* 3(4):235–312.

Montague, Richard. 1974. The proper treatment of quantification in English. In *Formal philosophy: selected papers of Richard Montague*, ed. Richmond H. Thomason. New Haven, CN: Yale University Press.

Moschitti, Alessandro, and Cosmin A. Bejan. 2004. A semantic kernel for predicate argument classification. In *Proceedings of the 8th Conference on Natural Language Learning*, 17–24. Boston, MA.

Narayanan, Srini, Charles J. Fillmore, Collin F. Baker, and Miriam R. L. Petruck. 2002. FrameNet meets the Semantic Web: A DAML+OIL frame representation. In *Proceedings of the AAAI Workshop on Semantic Web meets language resources*. Edmonton, Canada.

Narayanan, Srini, and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th International Conference on Computational Linguistics*, 693–701. Geneva, Switzerland.

Noreen, E. 1989. *Computer-intensive methods for testing hypotheses: An introduction*. John Wiley and Sons Inc.

Och, Franz Josef, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1): 19–52.

Ohara, Kyoko Hirose, Seiko Fujii, Toshio Ohori, Ryoko Suzuki, Hiroaki Saito, and Shun Ishizaki. 2004. The Japanese FrameNet project: An introduction. In *Proceedings of the LREC Workshop on Building Lexical Resources from Semantically Annotated Corpora*. Lisbon, Portugal.

Padó, Sebastian, and Gemma Boleda Torrent. 2004. The influence of argument structure on semantic role assignment. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, 103–110. Barcelona, Spain.

Padó, Sebastian, and Katrin Erk. 2005. To cause or not to cause: Cross-lingual semantic matching for paraphrase modelling. In *Proceedings of the EUROLAN Workshop on Cross-Linguistic Knowledge Induction*. Cluj-Napoca, Romania.

Padó, Sebastian, and Mirella Lapata. 2005a. Cross-lingual bootstrapping for semantic lexicons. In *Proceedings of the 22nd National Conference on Artificial Intelligence*, 1087–1092. Pittsburgh, PA.

———. 2005b. Cross-lingual projection of role-semantic information. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 859–866. Vancouver, BC.

———. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the joint Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, 1161–1168. Sydney, Australia.

Pallotta, Vincenzo, ed. 2005. *Proceedings of the EUROLAN Romance FrameNet Workshop*. Cluj-Napoca, Romania.

Palmer, Martha, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1):71–106.

Pang, Bo, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: extracting paraphrases and generating new sentences. In *Proceedings of the joint Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 102–109. Edmonton, AL.

Peters, Wim, Piek Vossen, Pedro Diez-Ortas, and Geert Adriaens. 1998. Cross-linguistic alignment of WordNets with an inter-lingual-index. *Computers and the Humanities* 32(2-3):221–251.

Petruck, Miriam R. L. 2005. Steps towards Hebrew FrameNet. Presentation given at the NAPH Conference on Hebrew Language and Literature. Stanford, CA.

Pianta, Emanuele, and Luisa Bentivogli. 2004. Knowledge intensive word alignment with KNOWA. In *Proceedings of the 20th International Conference on Computational Linguistics*, 1086–1092. Geneva, Switzerland.

Pitel, Guillaume. 2006. Using bilingual LSA for FrameNet annotation of French text from generic resources. Presentation given at the Workshop on Multilingual Semantic Annotation: Theory and Applications. Saarbrücken, Germany.

Postolache, Oana, Dan Cristea, and Constantin Orasan. 2006. Tranferring coreference chains through word alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61:127–159.

Resnik, Philip, and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Lingustics* 29(3):349–380.

Riloff, Ellen, and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence*, 474–479. Orlando, FL.

Riloff, Ellen, Charles Schafer, and David Yarowsky. 2002. Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of the 19th International Conference on Computational Linguistics*, 828–834. Taipei, Taiwan.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R.L. Petruck, and Christopher R. Johnson. 2005. FrameNet: Theory and Practice. `http://www.icsi.berkeley.edu/~framenet/book/book.html`.

Saint-Dizier, Patrick. 2005. PrepNet: a Framework for Describing Prepositions: preliminary investigation results. In *Proceedings of the 6th International Workshop on Computational Semantics*. Tilburg, The Netherlands.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, 44–49. Manchester, UK.

Schmid, Helmut, and Sabine Schulte im Walde. 2000. Robust German noun chunking with a probabilistic context-free grammar. In *Proceedings of the 18th International Conference on Computational Linguistics*, 726–732. Saarbrücken, Germany.

Schulte im Walde, Sabine. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics* 32(2): 159–194.

Sgall, Petr. 2000. English Syntax in Functional Generative Description, Topic-focus articulation (information structure) of the sentence, Syntax and semantics. In *Rudiments of English Linguistics*, ed. Pavol Štekauer, 225–265. Prešov: Slovacontact.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspect*. Dordrecht, The Netherlands: Reidel.

Smith, David A., and Noah A. Smith. 2004. Bilingual parsing with factored estimation: Using English to parse Korean. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, 49–56. Barcelona, Spain.

Subirats, Carlos, and Miriam R. L. Petruck. 2003. Surprise! Spanish FrameNet! In *Proceedings of the Workshop on Frame Semantics, XVII. International Congress of Linguists*. Prague, Czech Republic.

Surdeanu, Mihai, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 8–15. Sapporo, Japan.

Swier, Robert S., and Suzanne Stevenson. 2004. Unsupervised semantic role labelling. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, 95–102. Barcelona, Spain.

———. 2005. Exploiting a verb lexicon in automatic semantic role labelling. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 883–890. Vancouver, BC.

Swift, Mary D., Myroslava O. Dzikovska, Joel Tetreault, and James F. Allen. 2004. Semi-automatic syntactic and semantic corpus annotation with a deep parser. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal.

Talmy, Leonard. 2000. *Towards a cognitive semantics*. Cambridge, MA: MIT Press.

Taskar, Ben, Simon Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 73–80. Vancouver, BC.

Tatu, Marta, and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 371–378. Vancouver, BC.

Thompson, Cynthia A., Roger Levy, and Christopher Manning. 2003. A generative model for FrameNet semantic role labeling. In *Proceedings of the European Conference on Machine Learning 2003*, 397–408. Cavtat, Croatia.

Tiedemann, Jörg. 2003a. Combining clues for word alignment. In *Proceedings of the 16th Meeting of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

———. 2003b. Recycling translations. Ph.D. thesis, Uppsala University.

Truffaut, Louis. 1997. *Traducteur tu seras*. Brussels: Éditions du Hazard.

Tufiş, Dan. 2002. A cheap and fast way to build translation lexicons. In *Proceedings of the 19th International Conference on Computational Linguistics*, 1030–1036. Taipei, Taiwan.

Vauquois, Bernard. 1975. *La traduction automatique à Grenoble*. Paris: Dunod.

Viberg, Åke. 2006. What one verb can do: The Swedish verb göra in a crosslinguistic perspective. *SKY Journal of Linguistics* 19:243–257.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, 836–841. Copenhagen, Denmark.

Weeds, Julie. 2003. Measures and applications of lexical distributional similarity. Ph.D. thesis, University of Sussex.

Xia, Fei, and Michael McCord. 2004. Improving a statistical MT system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, 508–514. Geneva, Switzerland.

Xue, Nianwen, and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, 88–94. Barcelona, Spain.

Yamamoto, Kaoru, and Yuji Matsumoto. 2000. Acquisition of phrase-level bilingual correspondence using dependency structure. In *Proceedings of the 18th International Conference on Computational Linguistics*, 933–939. Saarbrücken, Germany.

Yarowsky, David, and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, 200–207. Pittsburgh, PA.

Yarowsky, David, Grace Ngai, and Roger Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st Human Language Technology Conference*, 161–168. San Francisco, CA.

Yeh, Alexander. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics*, 947–953. Saarbrücken, Germany.

Akademischer Lebenslauf