PARAPHRASES IN LFG-BASED BROAD-COVERAGE SEMANTICS

Sina Zarrieß and Jonas Kuhn
Universität Potsdam Universität Potsdam

Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications

http://csli-publications.stanford.edu/

Abstract

This paper adresses the problem of modelling paraphrases in a deep linguistic processing framework where the meaning construction component is based on an LFG grammar. We present a syntax-based approach to paraphrase extraction that operates on shallow dependency analyses in a parallel corpus. By means of an XLE-based conversion routine, we generate transfer rules for the automatically acquired semantic correspondences. These rules can be used as an additional component in the rule-based process of meaning construction which will augment the meaning representation with entailments that hold for complex phrasal units.

1 Introduction

This paper[1] [2] deals with the induction of a paraphrase lexicon for a rule-based meaning construction component which is based on a wide-coverage LFG grammar. We describe a technique for extracting paraphrases from a parallel corpus that exploits several broad-coverage analysis tools. The output of the paraphrase extraction is then fed to an XLE-based conversion routine that automatically derives meaning representations for phrasal expressions. The resulting paraphrase lexicon is implemented in the framework of LFG-based meaning construction outlined in Crouch and King (2006). The lexicon can be used as an additional module in the process of meaning construction.

Crouch and King's meaning construction system makes use of XLE's term rewrite engine to derive semantic representations from LFG F-structures. In addition to a hand-crafted rule component, the system integrates modules that augment the representation with lexical entries obtained from external resources. For instance, the meaning representation of a sentence containing the verb *see* would be enriched with a semantic predicate which asserts the meaning equivalence between *see* and its synonyms *watch, perceive*, and *notice*. This strategy of explicitly augmenting the meaning representation with all possible entailments can be considered as a process that derives the "deductive closure" of a given semantic analysis of a sentence. Given the "deductive closure" of two meaning representations, the computation of entailment between them boils down to a matching problem and no inference module is required. This strategy of "deductive closure" makes the system particularly suitable for semantic applications that need to deal with the problem of textual entailment — see Bobrow et al. (2007) for a question answering application that is built on top of Crouch and King's meaning construction.

The effectiveness of the strategy of "deductive closure" depends on the quality and the coverage of the captured semantic correspondences. However, whereas the coverage of currently available resources is often limited to single lexical items,

[1]The work reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in the Emmy Noether project PTOLEMAIOS, on Grammar Induction from Parallel Corpora.

[2]The current affiliation of the first author is Universität Stuttgart, IMS.

real-world semantic applications (like search or textual entailment) need to capture complex, phrasal correspondences. As an example, the following pair of sentences illustrates a synonymy relation between the phrase *put obstacles in the way of* and the simplex verb *impede*.

(1) The European Union puts obstacles in the way of importing genetically-modified products.

(2) The European Union impedes the import of genetically-modified products.

State-of-the-art approaches to paraphrase extraction usually do not focus on the further use of the resulting paraphrase resources in the framework of deep linguistic processing systems.[3] For instance, the extraction methods presented in Bannard and Callison-Burch (2005) would represent the semantic equivalence between phrases as the correspondence of their surface strings. As we will discuss in section 2, this simplifies the problem of semantic equivalence of phrases in an inappropriate way, since syntactic or semantic argument relations cannot be captured. Moreover, it is not clear how to integrate knowledge about surface string correspondence into a meaning representation that abstracts away from surface strings.

Regardless of the final use of the paraphrase resource, one could argue that the paraphrase extraction itself should not be exclusively based on deep processing provided by a particular linguistic formalism. A reasonable recall is essential for the paraphrase resources to be of any practical relevance for the already mentionned real-word applications. Therefore, our approach to the induction of a paraphrase lexicon aims to combine shallow and deep linguistic processing techniques: (i) The paraphrase extraction exploits shallow dependency analyses in addition to word alignments. In section 2, we show that a minimum of syntactic information is needed in order to establish well-formed semantic correspondences later. (ii) The derivation of deep meaning representations for paraphrases is carried out in an LFG setting. In the framework of Crouch and King (2006), we can map complex phrasal expressions that may relate argument slots at various levels in a hierarchical embedding structure to a simple semantic predicate with corresponding argument slots. Our proposal includes a routine for generating such transfer rules (which do get unwieldy for larger phrasal units) automatically from text instances, exploiting the XLE parsing system.

The rest of the paper is structured as follows: In section 2, we will discuss some examples of verb paraphrases, found in the Europarl corpus. These examples motivate our approach to paraphrase extraction, presented in section 3. In the first part of section 4, we will briefly introduce the LFG-based semantic framework proposed by Crouch and King (2006), which constitutes the underlying formalism for the implementation of our paraphrase representation. In the second part of

[3] A prototypical application context for this type of resource would be (phrasal) statistical machine translation, where additional data-driven components, such as a statistical (n-gram) language model, impose additional constraints on the usability of the induced paraphrases.

section 4, we describe the implementation of the conversion routine that produces semantic transfer rules for arbitrary types of semantic paraphrases.

2 Semantic Correspondences in Parallel Corpora

This section gives a general, non-technical overview of the strategy we exploit to induce meaning representations for phrasal expressions from parallel data.

The idea to acquire lexical semantic knowledge from translational data has been particularly pursued in the field of word sense disambiguation and acquisition (Resnik and Yarowsky, 1999; Ide et al., 2002; Dyvik, 2004). Crosslingual models of word sense inventories mainly exploit the fact that a lexical item in a source language usually has a (large) set of possible translations since its different senses are likely to translate to different words in a target language.

The main idea we propose in this paper is to extend this view to translational correspondences where a single lexical item in the source language corresponds to a complex expression in the target language. Zarrieß and Kuhn (2009) use these complex translational correspondences to identify multiword expressions. They assume that a phrase which has a simplex translation in another language can be considered a (at least partially) non-compositional multi-word. The semantic compositionality of phrases is also highly relevant for application-oriented semantic systems that need to account for inference relations. As an example, consider the German-English sentence pair in (3)-(4) from Europarl and the corresponding meaning representations derived by the transfer semantics where the English analysis corresponds to Crouch and King (2006) and the German analysis is produced as described in Zarrieß (2009). The meaning representations can basically be seen as flat, DRT-style analyses; for further detail, see section 4.

(3) Mit dem Gesetz wurde die Lage verschlimmert.
With the law was the situation aggravated.

HEAD (verschlimmern)
PAST (verschlimmern)
ROLE (Agent,verschlimmern,pro)
ROLE (Theme,verschlimmern,Lage)
ROLE (prep(mit),verschlimmern,Gesetz)

(4) The law made the situation even worse.

HEAD (make)
PAST (make)
ROLE (Cause,make,law)
ROLE (Experiencer,make,situation)
ROLE (Pred,make,bad)
COMPARATIVE-DIFF (bad,situation,unspecified)

The meaning representation for the English sentence in (4) would not permit the inference that there is an *aggravate*-relation between the Cause and the Experiencer since the predicative construction has been assigned a compositional meaning. The fact that the *make worse* construction can also be assigned a non-compositional meaning can be directly read off the meaning representation of its German translation where an *aggravate* relation holds between the corresponding Agent and Theme. On the other hand, the German meaning representation would not permit the inference that the instrumental *with*-PP acts as a Cause in the sentence, information that is explicit in the English meaning representation. The English representation also makes explicit the fact that the mentioned situation is compared to some previous, presupposed situation. This information remains implicit in the German representation and, therefore, could not be inferred. One could argue that the predicative construction (4) explicitly decomposes the lexical semantics of the German main verb whereas the German verb reflects the semi-compositional status of the predicative construction. Thus, both sides of the paraphrase inform each other.

In Zarrieß and Kuhn (2009), we find that complex translations of simplex words actually occur very frequently in Europarl. This observation can be directly exploited for paraphrase extraction, i.e. the extraction of monolingual semantic correspondences. Bannard and Callison-Burch (2005) use the source language of a parallel text as a pivot providing contextual features for identifying semantically similar expressions in the target language. Following Bannard and Callison-Burch (2005), we relate the meaning of some English expressions if they can translate to an identical German expression and vice versa. This line of reasoning is illustrated in figure 1 that exemplifies two translation instances of the German main verb *verschlimmern* ('aggravate'). From the fact that the verb has been translated by two different English phrases we make the assumption that their meanings correspond to each other. This means that the representation of a *make worse* predication can be enriched by the semantics obtained for an *exacerbate* predication, which results in the representation at the bottom of the figure.

To demonstrate the contrast between the deep and surface-based semantic correspondence extraction, figure 2 shows an example output of the system described in Bannard and Callison-Burch (2005) when paraphrases for the English verb *exacerbate* are looked for in the English-German Europarl section.[4] First of all, it can be noted that (at least in this particular case) the system has problems with phrasal correspondences as it proposes paraphrase pairs like *exacerbate* - *worse* or *exacerbate* - *made*. Moreover, for the pair *exacerbate* - *deteriorate*, it is unclear how the arguments or roles of *exacerbate* correspond to the arguments of *deteriorate*. It might be possible that the subject Experiencer of the latter corresponds to the object Patient of the former. The form in which the correspondences in figure 2 are given does not allow us to derive deep meaning representations for them.

[4]We used the code kindly made available by the authors on `http://www.cs.jhu.edu/~ccb/howto-extract-paraphrases.html`

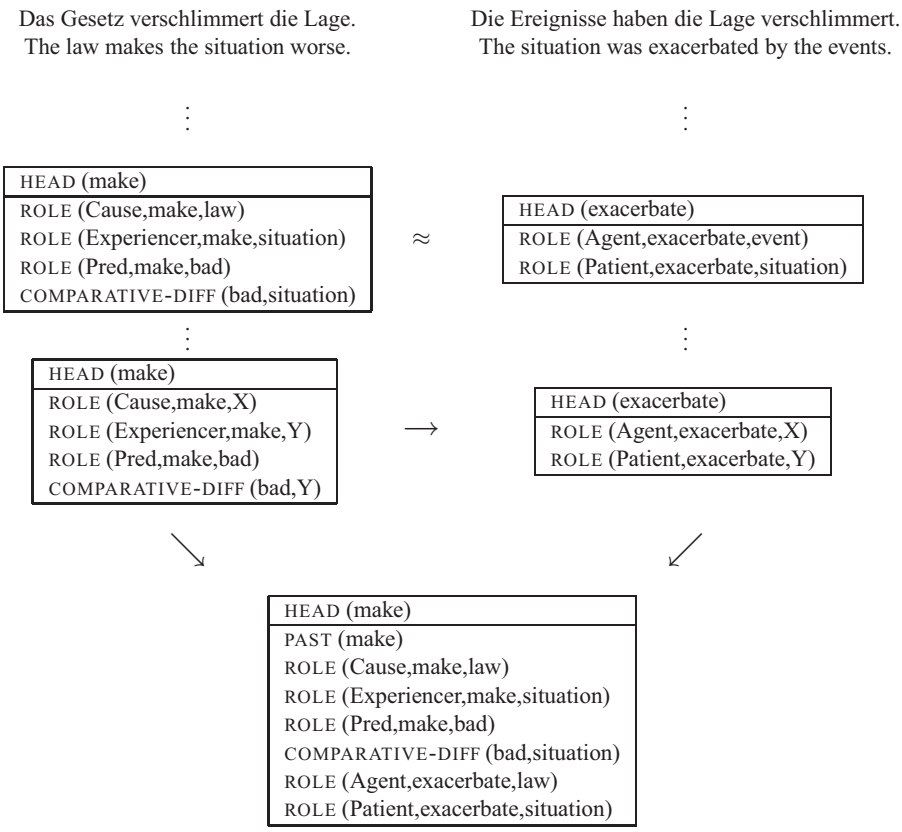

Figure 1: Inducing deep, monolingual meaning equivalences from translations

exacerbated	worse
exacerbated	increased
exacerbated	intensified
exacerbated	more
exacerbated	undermined
exacerbated	better
exacerbated	stepped
exacerbated	worsened
exacerbated	made
exacerbated	reinforced
exacerbated	waived
exacerbated	deteriorated
exacerbated	improved
exacerbated	increased
exacerbated	triggered
exacerbated	compounded
exacerbated	aggravated

Figure 2: Examples of paraphrases of *exacerbate* found in Europarl by the method proposed in Bannard and Callison-Burch (2005)

In contrast to surface-based paraphrase extraction, deep accounts of complex meaning equivalences need to capture correspondences between argument slots. These correspondences can be easily derived if we have some information about the syntactic or semantic parallelism between the two sides of the paraphrase. In figure 1, we can establish a correspondence relation between the Cause of *make worse* and the Agent of *exacerbate*, because they are aligned to the same German verb's Cause. In general, we extend the concept of a pivot, defined as a surface string in Bannard and Callison-Burch (2005), to a semantic relation that occurs in a particular argument frame. These frame correspondences can serve as lexical entries which can be integrated into a deep representation of sentence meaning.

The "pivot approach" to paraphrase induction has several limitations. First of all, it ignores the problem that the meaning equivalence of two expressions might be very context dependent. Thus, paraphrases that are valid in some specific context do not necessarily have exactly the same set of senses and selectional preferences such that, in other contexts, their substitution might not be possible. A related question is the directionality of paraphrase rules (one expression might have a more general meaning than the other) which cannot be captured by the naive pivot approach. Unfortunately, the automatic modelling of the context-dependence of paraphrase relations can be considered an unsolved problem (see e.g. Erk and Pado (2009)) and reliable systems are not yet available.

Being aware of the limitations of a paraphrase acquisition method that does not deal with the context dependence of meaning and is, therefore, noisy to a certain extent, we consider the corpus-based extraction of meaning correspondences an attractive way to supplement existing, hand-coded resources. To assure the quality of the paraphrase lexicon in the context of a high-precision semantic representation,

one could manually inspect the automatically acquired lexical entries, still benefiting from the improved recall offered by corpus-based paraphrase acquisition.

To summarize, this section first discussed examples of translations which suggest that translational correspondences on the phrase level can decompose the semantics of a lexical item and make explicit some of its inferences in the meaning representation. Second, by means of the pivot approach, translations can be exploited for the acquisition of monolingual correspondences. We proposed an extension to the pivot approach used by Bannard and Callison-Burch (2005) to discover inference relations between two complex meanings. In order to capture correspondences between semantic relations for deep meaning representation, the paraphrase extraction has to capture correspondences between argument slots.

3 Extraction of Syntactic Correspondences from Parallel Corpora

This section will describe the implementation of the paraphrase extraction that has been discussed theoretically in the last section.

Crucially, our approach only relies on flat dependency analyses that can be obtained from currently available, statistical state-of-the-art parsers. This shallow syntactic information is used to approximate information about argument slot correspondences needed by the meaning representation derivation. Moreover, the extraction method exploits the syntactic information as an indicator of the reliability of the translation candidate. The final conversion from syntactic to semantic correspondences is treated in section 4.

The section is structured as follows: The data preprocessing is described in section 3.1. Section 3.2 deals with the extraction of crosslingual syntactic correspondences from shallow dependency analyses. In 3.3, we discuss the mapping of crosslingual correspondences onto monolingual ones.

3.1 Parallel Data for Paraphrase Extraction

We base our investigations on the German and English portion of the Europarl corpus (Koehn, 2005) which is available in a sentence-aligned, tokenized format. To produce word-alignments for the German-English parallel text, we used the wide-spread, open-source GIZA++ tool (Och and Ney, 2003). We employed the standard settings for alignments in both directions (viterbi alignments, IBM model 4) and the refined alignment heuristics for bidirectional alignment.

To obtain robust syntactic analyses for the two portions of the parallel corpus, we used MaltParser (Nivre et al., 2007), a data-driven dependency parsing system which is freely available.[5] In comparison to other statistical state-of-the-art parsing systems, MaltParser has especially proven succesful for a broad range of languages. The English version of the parser was trained on the Penn treebank. The

[5] http://maltparser.org

German version of the parser was trained on the Tiger treebank. In future work, we might use a more recent model of the German parser which was trained on the Tiger treebank enriched with features from deep LFG parses (Øvrelid et al., 2009).

Technically, the resulting resource of parallel, word-aligned dependency parses is stored as a relational database where, for each token, its monolingual properties, like lemma, POS, and syntactic head, as well as its crosslingual relations, i.e. the aligned tokens of the target language, can be efficiently represented. The extraction procedures described in the following section are basically implemented as a cascade of queries on the database.

3.2 Syntax-based Paraphrase Extraction

Given the parallel dependency trees obtained from EuroParl, we now want to extract German-English paraphrasing translations that involve the correspondence between a simplex lexical item on the source (German) side and a complex phrasal expression on the target (English) side. As an example, consider the sentence pair given in (5)-(6) where the German verb *behindern* corresponds to the English expressions *constitute an obstacle to*. The extraction of such complex translational correspondences involves the major challenge that, typically, only certain parts of the target phrase can be reliably aligned to the source item due to the low occurrence correlation of the other parts. For instance, in sentences (5)-(6), GIZA++ is not likely to be able to capture the correspondence between the German main verb and the whole English phrase, but instead to find only an alignment link between the German main verb and the noun *obstacle*. For further detail on this alignment problem see Zarrieß and Kuhn (2009).

The general intuition is that the alignment of phrasal correspondences somehow needs to relax the requirement of high cooccurrence correlation while still detecting reliable translation instances. The main idea of our paraphrase detection approach is to relax the cooccurrence correlation based on leveraging syntactic information. For instance, consider the pair of parallel configurations in figure 3 for the sentence pair given in (5) and (6). Although there is no strict one-to-one alignment for the German verb, the basic predicate-argument structure is parallel: The verb's arguments directly correspond to each other and are all dominated by a verbal root node.

We propose a generate-and-filter strategy for our translation detection which extracts partial, largely parallel dependency configurations. The input to the candidate generation is a source lexical item in a predefined syntactic configuration that exhibits two or more argument slots, e.g. a verb with its subject and object argument dependency relations. The output of the candidate generation is a set of translation instances where the German verb occurs in the predefined argument frame and the English translation exhibits argument slots that can be consistently aligned with the source slots. To filter noise due to parsing or alignment errors, we further introduce a filter on the length of the path that connects the target root and its dependents and a filter that excludes paths crossing sentence boundaries.

(5) Die Korruption **behindert** die Entwicklung.
The corruption impedes the development.

(6) Corruption **constitutes an obstacle to** development.

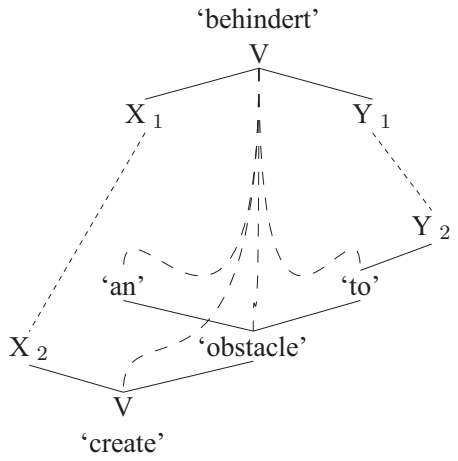

Figure 3: Example of a phrasal translational correspondence configuration

By allowing target dependency paths to be aligned to source single dependency relations, we admit configurations where the source item is translated by more than one word. Thus, we propose to use the aligned arguments as anchors of the configuration. Then, the search basically selects all the items lying on a path between the target root item and the target arguments, very similar to Lin and Pantel (2001) who pursue a similar approach for paraphrase extraction on monolingual corpora. For instance, given the configuration in figure 3, we allow the German verb, which connects the argument X_1 and X_2, to be aligned to the path connecting Y_1 and Y_2.

We evaluate the accuracy of our translation detection approach, especially for the accuracy of phrasal translations, by manually classifying 300 random, exclusively phrasal translations that our system detects for 50 German verbs (all selecting for a nominative subject and accusative object). We extract 50 random, transitive verbs from the German LFG grammar lexicon. We supply these verbs in their desired syntactic configuration to the translation search method described in this section and recover the reliable alignments that the search detects. Out of the resulting correspondences, we select 300 random instances from the set of phrasal configurations. The results of the classification are displayed in table 1. We observe that 17.1% of the translations detected by the system do not correspond to semantic correspondences, whereas more than 80% show different patterns of complex translation patterns. For a more detailed discussion of the translation search and an analysis of the patterns see Zarrieß and Kuhn (2009).

Trans. type	Proportion		
		MWE type	Proportion
MWEs	57.5%	V Part	8.2%
		V Prep	51.8%
		LVC[6]	32.4%
		Idiom	10.6%
Paraphrases	24.4%		
Alternations	1.0%		
Noise	17.1%		

Table 1: Classification of 300 types sampled from the set of one-to-many translations for 50 verbs

3.3 From Crosslingual to Monolingual Correspondences

In section 2, we have explained the general method allowing for the deduction monolingual semantic correspondences from translation pairs. The paraphrase extraction produces pairs of given source-configurations s_i and their corresponding set of translations T_i, where for each argument slot on the source side $a_{s_n} \in s_i$, each target configuration $t_i \in T_i$ contains a corresponding argument slot $a_{t_n} \in t_i$. From a particular set of target configurations T_i, we would obtain the resulting set of target correspondences by taking the Cartesian product of this set $T_i \times T_i$. In practice, taking the product of the set of target configurations would result in a huge set of meaning postulates since the number of translations to be found in parallel corpora is usually very high (Dyvik, 2004). Moreover, the set product would replicate a major part of the already existing rules for word level correspondences since our target correspondence might not necessarily be phrasal expressions.

A theoretical issue brought up by paraphrase or entailment induction is the directionality of the relation. As Basili et al. (2007) point out, automatically assembled paraphrase resources usually lack a notion of directionality and capture entailment as a bidirectional equivalence relation. The authors also remark, however, that the problem of context dependence is usually more serious in practice than the problem of uni-directional entailment. As with the treatment of context-dependence, the work presented in this paper does not deal with directionaly either.

The current implementation of the paraphrase derivation proceeds as follows: the set of target translations T_i is separated into a set of configurations that exhibit a word-level correspondence to the source item, T_{w_i}, and a set of configurations that exhibit a phrasal expression corresponding to the source item, T_{p_i}. From the set of simplex translations T_{w_i}, we select the most frequent translation and relate it to all elements of complex translations T_{p_i}. Future work might implement more sophisticated methods for the selection of the actual monolingual correspondences.

[6]light verb construction

3.4 Discussion

First of all, this syntax-based extraction of translational correspondences has the advantage that the alignment is supplied with additional cues to coccurrence and can thus extract configurations that do not have to be very frequent. Moreover, we can control for the syntactic configuration on the source side of the translation such that we are likely to find, for a given pivot configuration, instances of that configuration that all share the same argument frame. In section 2, we have seen that consistent argument frames among the source language pivots are essential for establishing the correct correspondence between the argument slots. In this way, our syntax-based search inherently controls for the voice of the source verb which is also crucial for establishing the correct argument correspondences.

The syntax-based approach also has certain drawbacks, for instance, the fact that it relies, to a certain extent, on syntactic parallelism, i.e. on the parallel realization of predicate arguments. However, research in syntax projection has shown that the divergence of dependency structure across languages might be quite drastic such that, without additional information sources, straightforward projection of dependency relations between aligned pairs of words yields relatively poor results (Bouma et al., 2008; Hwa et al., 2005). Similarly, Pado (2007) observes that crosslingual parallelism on the level of predicate argument relations still shows considerable variation. For our method which presupposes the parallelism of the argument slots, this means that it cannot take account of the many translation instances that do not exhibit syntactic parallelism (e.g. target passive translations of a source active verb where the agent of the source verb is omitted).

A further limitation of the syntax-based approach lies in the fact that the expressions we want to extract need to be "syntactically anchored". In the case of transitive verbs where the pair of arguments can naturally serve as syntactic anchor, this does not pose a problem. But lexical items like adjectives or intransitive verbs which do not have more than one argument position do not offer the possibility of finding their translational equivalents by looking for the path that connects the translation of their arguments. Future work on the extraction method might investigate a more general way to take the syntactic configuration of a translation into account, adressing the partial parallelism as well as the anchoring problem.

4 Induction of Paraphrase Meanings from Syntactic Correspondences

The paraphrase extraction described in the previous section produces pairs of dependency graph configurations. In each of the configurations, a verb and its arguments on the source side correspond to a target phrasal expression that realizes the same argument slots somewhere in its dependency configuration. This section deals with the method we employ to map these parallel dependency configurations onto semantic correspondence rules that can be applied in an LFG-based meaning

construction component. In section 4.1, we will first describe the basic properties of the transfer semantic representation and the architecture performing the meaning construction. Section 4.2 then deals with the induction of the paraphrase lexicon in the context of the transfer semantics.

4.1 An LFG-based Transfer Component for Meaning Construction

The meaning construction described in Crouch and King (2006) converts LFG F-structures produced by the English ParGram grammar to flat representations in a Neo-Davidsonian style. Since the ParGram initiative (Butt et al., 2002) has particularly focussed on crosslingual parallelism on the level of syntactic analyses, this symbolic conversion routine can be easily ported to other LFG grammars, as has been done e.g. for German (Zarrieß, 2009).

The main idea of the system is to convert the surface-independent, syntactic relations and features encoded in an F-structure to normalized semantic relations. The semantic conversion was implemented by means of the XLE platform, used for grammar development in the ParGram project. It makes use of the built-in transfer module to convert LFG F-structures to semantic representations. The idea to use transfer ordered rewriting rules to model a semantic construction has also been pursued by Spreyer and Frank (2005) who use the transfer module to model a Minimal Recursion Semantics construction for the German treebank TIGER.

4.1.1 The Meaning Representation

To begin with an example, a simplified F-structure analysis for the following sentence and the corresponding semantic representation are given in figure 4.

(7) Where was Peter seen?

First, the interrogative pronoun induces a semantic context that embeds the proposition headed by the main verb. For the sake of readability, we visualize the semantic contexts as DRT-style boxes. The syntactic arguments and adjuncts of the main predicate are represented in terms of semantic roles of the context introduced by the main predicate or some higher semantic operator. Thus, the grammatical roles of the main verb in sentence (7) are semantically normalized such that the subject of the passive is assigned the Stimulus role and an implicit Experiencer is introduced; see figure 4. This type of semantic representation is inspired by Neo-Davidsonian event semantics (in the style of Parsons (1990)). Other semantic properties of the event introduced by the main verb such as tense or nominal properties such as quantification and cardinality are explicitly encoded as conventionalized predications.

The contexts can be thought of as propositions or possible worlds. They are headed by an operator that can recursively embed further contexts. Context embeddings can be induced by, e.g. negation, conditionals or clause-embeddings.

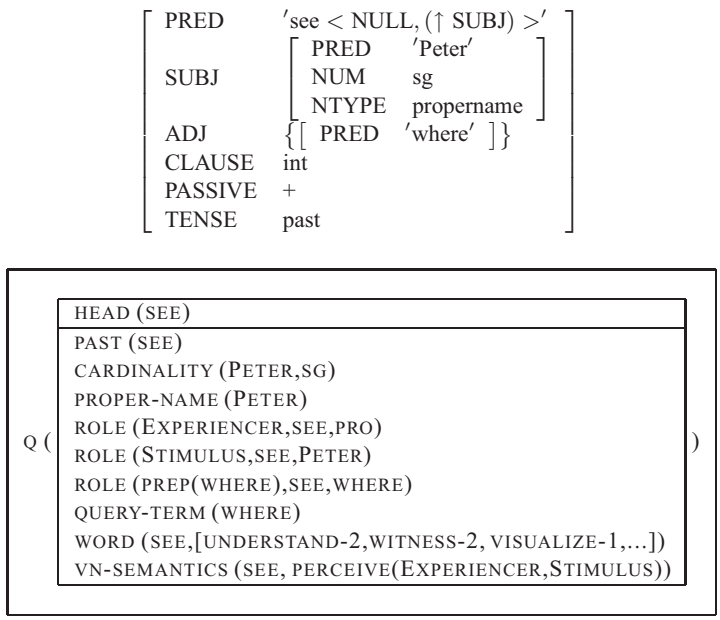

Figure 4: LFG F-structure analysis and corresponding semantic representation

4.1.2 Representation of Semantic Correspondences on the Word-Level

The semantic representation in figure 4 illustrates how the lexical meaning of the individual words that make up the sentence is represented. The predication WORD links the word `see` to a WordNet index that contains the corresponding synsets of the predicate. Thus, all synonyms of a word are enumerated in the representation and, by this means, directly available to entailment components. Therefore, if the WordNet entry for `see` contains the predicate `spot`, the correspondence between the sentence (7) and e.g. the sentence *Where was John spotted?* boils down to a matching of the semantic representation of these two sentences. Likewise, the predication VN-SEMANTICS states a VerbNet entry for the given verb. These entries capture verb-class equivalences and some deeper alternations. Generally, the meaning representation of a sentence explicitly and exhaustively asserts the lexical entries of its individual words, a strategy we referred to as "deductive closure" in section 1. For a more detailed description of the lexical resource interface implemented for the English transfer semantics system see Crouch and King (2005).

4.1.3 The Meaning Construction Component

The XLE transfer module, which is used for the implementation of the conversion of F-structures to semantic representations, is a term rewrite system that applies an ordered list of rewrite rules to a given F-structure input and yields an output transfer structure. Depending on the rewrite mode and on the definition of the rule, the output can be a fully-fledged F-structure again or else a set of (recursively embedded)

prolog clauses whose format is not further constrained. An example rewrite rule which yields F-structure output is given in figure 5. It applies to F-structures that have a passive and vtype feature as well as an oblique agent, mapping the oblique agent to a subject. The F-structure scope or embedding of the features (`%V` in this case) is given as the first variable of the fact.

```
VTYPE(%V, %%), PASSIVE(%V,+),
OBL-AG(%V, %LogicalSUBJ)
==>
SUBJ(%V, %LogicalSUBJ).
```

Figure 5: Example rewrite rule for passive normalization

The transfer system comes as a generic rewrite system and does not only apply to XLE F-structures. Therefore, it can be generally used to formulate mappings between clausal structures (given in the Prolog-format currently used by XLE). This flexible rewrite architecture makes it possible to organize the semantic construction or conversion in a modular way since rules can also apply to semantic transfer structures. This architecture substantially eases the integration of lexical knowledge. An example for an exemplary semantic lexicon and its integration in the semantic conversion is given in figure 6. The fact marked with `|-` first asserts that *aristocracy* is a collective noun. The following rule then matches all input meaning representations that contain a singular collective noun and rewrites their cardinality to plural.

```
|- collective_number(aristocracy).

collective_number(%NounForm),
in_context(%C, cardinality(%NounForm, sg))
==>
in_context(%C, cardinality(%NounForm, pl)).
```

Figure 6: Example rewrite rule for semantic rewrite

Essentially, the induction of the lexicon entries presented in section 4.2 makes use of the transfer system by automatically generating transfer rules that map partial semantic representations onto some semantically equivalent representation.

4.2 Deriving Transfer Rules for Semantic Correspondences

Our ultimate goal in this section is to define lexicon entries (or transfer rules) for paraphrases that match sentences that contain instances of these paraphrases. The lexicon entries will augment the transfer meaning representation with entailments which hold for larger units than single words. Thus, we extend the strategy of "deductive closure" from simplex lexical items to complex phrases.

The main requirement for the procedure of lexicon entry derivation is that it has to be independent of the semantic pattern of the paraphrase. As can be seen in the classification of the extracted paraphrase types in table 1, phrasal correspondences in parallel corpora yield very different types of semantic correspondences. As an example, consider the paraphrases in (9) and (10) and their semantics (as derived by the LFG-based transfer semantics) which have been found corresponding to the *hinder*-configuration in example (8). Item (9) exemplifies a light verb construction with the complex preposition *in the way of*. The paraphrase in (10) is even more complex because it exhibits a coordination, mapped to a complex event operator, where the argument slots of the dependency configuration have to be mapped to several semantic roles in the meaning representation.

These examples make clear that the definition of the lexicon entries cannot be done by handwritten templates. In order to match the paraphrase meaning representation with possible input sentences, the lexicon entries need to anticipate the analysis that is assigned to the paraphrase by the core meaning construction.

(8) x hinders y

HEAD (hinder)
ROLE (Agent,hinder,X)
ROLE (Patient,hinder,Y)

(9) x puts obstacles in the way of y

HEAD (put)
ROLE (Agent,put,X)
ROLE (Theme,put,obstacle)
ROLE (Destination,put,way)
ROLE (prep(of),way,Y)

(10) x obstructs or even fires on y

HEAD (Complex Event)		
	OR (	HEAD (obstruct)
		ROLE (Agent,obstruct,X)
		ROLE (Experiencer,obstruct,Y)
	,	
		HEAD (fire)
		ROLE (Agent,fire,X)
		ROLE (Patient,fire,pro)
		ROLE (prep(on),fire,Y)
	)	

Figure 7: Examples for complex semantic correspondences

4.2.1 The conversion routine

To convert the pairs of dependency configurations, produced by the paraphrase extraction, to pairs of meaning representations, we make use of the XLE analysis pipeline. The basic method of paraphrase conversion works as follows:

1. Map the set of dependency configurations, as well as the target verb configuration, to a set of surface sentences parsable by the XLE engine. Replace the argument slots of the dependency configurations by some dummy pronouns that will be uniquely identifiable later.

2. Parse all the paraphrase sentences, as well as the target meaning sentence, with XLE. Run the F-structure output through the transfer semantics construction.

3. Run the output semantic representations through a sequence of transfer rules that removes the context-specific clauses from the semantic representation.

4. In each of the stripped down semantic representations, replace the scope identifiers by prolog variables. Replace the dummy pronouns, corresponding to the aligned argument slots, by prolog variables such that one argument slot corresponds to a unique variable.

5. Generate a set of transfer rules that has a paraphrase meaning representation as its left-hand side and the meaning representation of the target simplex verb as its right-hand side.

The mapping of the extracted dependency configurations onto surface sentences (point 1) is simply done by linearizing the lexical items in their original word order. We normalize the inflection of the verbal root node to third person singular such that the subject dummy pronoun can be parsed as a third person pronoun. This first step of the routine is simplified by the fact that we only investigate verbs in a well-defined argument frame such that the surface sentences can receive a full syntactic (and semantic) analysis. In so doing, we avoid having to identify parts of the meaning representation since (almost) the entire semantic analysis of the paraphrase sentence actually corresponds to the paraphrase meaning.

Note, however, that the complete meaning representation of the surface paraphrase sentence obtained from the meaning construction is not yet exactly the lexical entry we want to include in the paraphrase lexicon. If we parse an example sentence like *x puts obstacles in the way of y*, we necessarily obtain a semantic analysis that includes non-general, context-specific features like tense for verbs or cardinality for nominals. Therefore, we define an additional set of transfer rules which is applied on top of the usual meaning construction and which deletes these non-general components of the paraphrase meaning. Besides tense and cardinality, this list of deletion rules contains the pronoun-specific facts, specifier predications and clauses that keep the original F-structure attributes.[7]

[7]The general deletion of these meaning components is a simplification in the case of more or less fixed idioms, e.g. *kick the bucket* where *bucket* has to be singular.

4.2.2 Example

The conversion routine described in the preceding section yields transfer rules that map an arbitrarily complex predication to a simplex semantic relation, assigning a non-compositional meaning to the paraphrase. Coming back to the example sentences (8) and (9), we obtain a lexical entry for the light verb construction *put obstacles in the way of* as given in figure 8. In contrast to the representations shown so far, the rule is given in its internal specification where the contexts do not correspond to boxes, but to context predicates.

The right hand side of the rule introduces a new word *hinder* and maps the subject of *put* to the subject of *hinder* and the object of the preposition to the object of *hinder*. Note that the rule does not delete the original analysis of the paraphrase, but just augments the representation with an additional relation that holds between the involved referents (the + in front of the left hand side facts tells the rewrite mechanism not to delete them from the set of input clauses). This is consistent with the strategy of deductive closure that is already implemented for semantic equivalences on the word-level. After the application of the lexical entry in figure 8, the respective meaning representation also matches all further lexical rules that hold for the *hinder*-relation. By this means, the paraphrase *put obstacles in the way of* can also be related to all synonyms of *hinder*.

```
+context_head(%ctx,put:%put),
+in_context(%ctx,role(Theme,put:%p,obstacle:%o)),
+in_context(%ctx,role(Agent,put:%p,%X)),
+in_context(%ctx,role(prep(of),way:%w,%Y)),
+in_context(%ctx,role(Destination,put:%p,way:%w))
==>
context_head(%ctx,hinder:%p),
skolem_info(hinder:%p,hinder,verb,verb,%p,%ctx),
in_context(%ctx,role(Agent,hinder:%p,%X)),
in_context(%ctx,role(Patient,hinder:%p,%Y)).
```

Figure 8: Example of a paraphrase representation as a transfer rule.

In contrast to the surface string representation of paraphrases we discussed in section 2, the lexical entry given in figure 8 subsumes a large number of possible surface realizations of the paraphrase. The following sequence of example sentences illustrates a number of surface phenomena that the lexical entry abstracts from.

1. X is putting obstacles in the way of Y.
2. X is putting some major obstacles in the way of Y.
3. A huge obstacle was put in the way of Y by X.
4. X puts an obstacles in Y's way.

For instance, the application of the lexical entry is independent of the tense of the construction and possible modifications of the nominal *obstacle* since the non-general clauses where deleted during the conversion routine. The entry is also independent of the voice that is instantiated by the paraphrases. Also, since the semantic constructions maps genitive and *of*-possessives onto the same representation, the lexical entry abstracts even further from specific surface realizations. This is a very desirable property of a paraphrase lexicon since, apart from syntactically rather fixed multiword expressions, paraphrases can occur in a wide range of syntactic specifications.

A further issue illustrated by the rule in figure 8 is the treatment of embeddings by means of variables. The new *hinder*-relation is not necessarily added to the main context of the input semantic representation, but it is made dependent on the embedding of the clauses which match the left hand side of the rule. If the left-hand side of the rule exhibits any context-embeddings (e.g. the representation in (10)), the context of the right-hand side will be the root context of the left-hand side. This treatment of context ensures that the lexical entry does not change the inferential properties of the input meaning, but applies to embedded paraphrases (e.g. through negation) as well.

5 Conclusion

In this paper, we have presented a way to augment a rule-based, hand-crafted meaning construction with semantic knowledge that has been automatically acquired from a large corpus. In particular, we have addressed the problem that surface string correspondences, as they are found by most corpus-based paraphrase extraction tools, cannot be easily integrated into deep meaning representations of sentences. The main reason for this problem is that the deep meaning representations have to be anchored via their semantic roles. However, surface-string correspondence does not allow us to induce this anchoring. We have outlined a way to approximate the correspondence of semantic roles via aligned syntactic arguments in a parallel corpus considered at the stage of paraphrase extraction.

Our syntax-based paraphrase extraction operates on wide-coverage, shallow dependency analyses. Technically, this involves the limitation that the method currently only works for expressions that have at least two argument positions which can serve as syntactic anchors. The syntax-based approach also suffers from the drawback of relying on the partially parallel realization of the predicate arguments in the target language. However, the translation search which is based on syntactic anchors performs better than raw word-alignment for transitive verbs.

The implementation of our XLE-based conversion routine produces lexical entries for paraphrases by deriving transfer rules, as defined in the XLE transfer module. These rules capture correspondences of complex phrasal expressions and contribute information about their inferential properties (e.g. their compositionality, implicit presuppositions). We have further shown that deep meaning representa-

tions of paraphrases has the practical advantage that the resulting lexical entries can capture a wide range of surface realizations of that paraphrase.

The main challenge for future work will be the treatment of the context dependence of paraphrases. It is crucial for the strategy of "deductive closure" of meaning representations that all asserted entailments hold in the given text or sentence pair. One possible way to assure the quality of the paraphrase lexicon would be a manual post-processing step that removes overly context-specific paraphrases from the list of transfer rules. But finally, it seems indispensable to have a statistical disambiguation component integrated in the process of meaning construction that discards invalid entailments from the representation, based on a context-sensitive model of meaning.

References

Bannard, Colin and Callison-Burch, Chris. 2005. Paraphrasing with Bilingual Parallel Corpora. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604, Morristown, NJ, USA: Association for Computational Linguistics.

Basili, Roberto, Cao, Diego De, Marocco, Paolo and Pennacchiotti, Marco. 2007. Learning Selectional Preferences for Entailment or Paraphrasing Rules. In *Proceedings of RANLP 2007*.

Bobrow, Daniel G., Cheslow, Bob, Condoravdi, Cleo, Karttunen, Lauri, King, Tracy Holloway, Nairn, Rowan, de Paiva, Valeria, Price, Charlotte and Zaenen, Annie. 2007. Precision-focused Textual Inference. In *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 28 – 29.

Bouma, Gerlof, Kuhn, Jonas, Schrader, Bettina and Spreyer, Kathrin. 2008. Parallel LFG Grammars on Parallel Corpora: A Base for Practical Triangulation. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG08 Conference*, pages 169–189, Sydney, Australia: CSLI Publications, Stanford.

Butt, Miriam, Dyvik, Helge, King, Tracy Holloway, Masuichi, Hiroshi and Rohrer, Christian. 2002. The Parallel Grammar Project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan*.

Crouch, Richard and King, Tracy Holloway. 2005. Unifying Lexical Resources. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 32–37.

Crouch, Richard and King, Tracy Holloway. 2006. Semantics via F-Structure Rewriting. In Miriam Butt and Tracy Holloway King (eds.), *Proceedings of the LFG06 Conference*.

Dyvik, Helge. 2004. Translations as Semantic Mirrors. From Parallel Corpus to WordNet. *Language and Computers* 1, 311 – 326.

Erk, Katrin and Pado, Sebastian. 2009. Paraphrase Assessment in Structured Vector Space: Exploring Parameters and Datasets. In *Proceedings of the EACL Workshop on Geometrical Methods for Natural Language Semantics (GEMS)*.

Hwa, Rebecca, Resnik, Philip, Weinberg, Amy, Cabezas, Clara and Kolak, Okan. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering* 11(3), 311–325.

Ide, Nancy, Erjavec, Tomaz and Tufis, Dan. 2002. Sense Discrimination with Parallel Corpora. In *Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. ACL2002, July Philadelphia 2002*, pages 56–60.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *MT Summit 2005*.

Lin, Dekang and Pantel, Patrick. 2001. DIRT - Discovery of Inference Rules from Text. In *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328.

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kübler, Sandra, Marinov, Svetoslav and Marsi, Erwin. 2007. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering* 13(2), 95–135.

Och, Franz Josef and Ney, Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1), 19–51.

Øvrelid, Lilja, Kuhn, Jonas and Spreyer, Kathrin. 2009. Improving data-driven dependency parsing using large-scale LFG grammars. In *Proceedings of the Annual Meeting for the Association for Computational Linguistics (ACL) (Short Paper)*.

Pado, Sebastian. 2007. Translational Equivalence and Cross-lingual Parallelism: The Case of FrameNet Frames. In *Proceedings of the NODALIDA Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages*, Tartu, Estonia.

Parsons, Terence. 1990. *Events in the Semantics of English*, volume 19 of *Current studies in linguistics series*. Cambridge, Mass.: MIT Press.

Resnik, Philip and Yarowsky, David. 1999. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Nat. Lang. Eng.* 5(2), 113–133.

Spreyer, Kathrin and Frank, Anette. 2005. The TIGER 700 RMRS Bank: RMRS Construction from Dependencies. In *Proceedings of LINC 2005*, pages 1–10.

Zarrieß, Sina. 2009. Developing German Semantics on the basis of Parallel LFG Grammars. In Tracy Holloway King and Marianne Santaholma (eds.), *Proceedings of the ACL Workshop on Grammar Engineering Across Frameworks (GEAF)*, Singapore.

Zarrieß, Sina and Kuhn, Jonas. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In *Proceedings of the ACL Workshop on Multiword Expressions*, Singapore.