Towards a Computational Semantic Analyzer for Urdu

Annette Hautli **Miriam Butt**
Department of Linguistics
University of Konstanz
{annette.hautli|miriam.butt}@uni-konstanz.de

Abstract

This paper describes a first approach to a computational semantic analyzer for Urdu on the basis of the deep syntactic analysis done by the Urdu grammar ParGram. Apart from the semantic construction, external lexical resources such as an Urdu WordNet and a preliminary VerbNet style resource for Urdu are developed and connected to the semantic analyzer. These resources allow for a deeper level of representation by providing real-word knowledge such as hypernyms of lexical entities and information on thematic roles. We therefore contribute to the overall goal of providing more insights into the computationally efficient analysis of Urdu, in particular to computational semantic analysis.

1 Introduction

The state of the art in wide-coverage deep syntactic parsing has allowed semantic processing to come within reach of applications in computational linguistics (Bos et al., 2004). This new possibility for wide-coverage computational semantic analysis however also raises questions about appropriate meaning representations as well as engineering issues. We address some of these here.

In achieving the goal of producing a deep, broad-coverage semantic analysis for text, much effort has been put into the development of robust and broad-coverage syntactic and semantic parsers as well as lexical resources. However, the focus has mostly been on European languages.

For Urdu, neither a wide-coverage computational semantic analyzer nor a wealth of lexical resources exist to date; however, efforts have been put into the development of a syntactic parser within the framework of Lexical-Functional Grammar (LFG) (Bresnan and Kaplan, 1982; Dalrymple, 2001), namely the Urdu ParGram Grammar (Butt and King, 2002; Bögel et al., 2007; Bögel et al., 2009). As to the development of lexical resources, Ahmed and Hautli (2010) have generated a preliminary Urdu WordNet on the basis of Hindi WordNet (Bhattacharyya, 2010). A lexical resource for Urdu verbs following the methodology of the English VerbNet (Kipper-Schuler, 2005) is currently under construction, some of its content has already been hooked into the our Urdu semantic analyzer Urdu (section 2.2.2).

The computationally efficient semantic analysis of Urdu is a completely new area of research and it is not immediately clear what a cross-linguistically motivated representation and analysis should look like. Therefore, the aim of this paper is to present a first approach to a computational semantic representation of Urdu and to discuss some of the challenges that have to be dealt with. In addition we show how external lexical resources can be linked to the system and discuss what information these lexical resources contribute to the overall semantic analysis.

The paper is structured as follows: Section 2 elaborates on some of the resources available for Urdu, followed by a detailed description of the semantic analyzer in Section 3. Section 4 elaborates on some of the issues involved in building the system, followed by the conclusion in Section 5.

2 Concepts

2.1 The Urdu ParGram Grammar

The Urdu LFG grammar (Butt and King, 2002; Bögel et al., 2007; Bögel et al., 2009) is part of an international research program called ParGram (Parallel Grammars) (Butt et al., 2002), aiming at developing parallel syntactic analyses for different languages within the LFG framework (Butt et al., 1999). The underlying platform that is used to

Proceedings of the 9th Workshop on Asian Language Resources, pages 71–78,
Chiang Mai, Thailand, November 12 and 13, 2011.

develop parallel LFG grammars is XLE (Crouch et al., 2011), developed at Palo Alto Research Center (PARC) and consisting of cutting-edge algorithms for parsing and generating LFG grammars along with a user interface for writing and debugging.

LFG postulates two basic levels of syntactic description for natural language utterances. Phrase structure configurations (linear order, constituency and hierarchical relations) are represented in a *constituent structure* (c-structure), whereas grammatical functions are explicitly represented at the other level of description, the *functional structure* (f-structure), an attribute value matrix (AVM).

Building XLE grammars involves the manual writing of syntactic rules that are annotated with f-structure information. It is possible to incorporate a stochastic disambiguation module into the grammar (Riezler et al., 2002), but this still needs to be done for the Urdu grammar. The amount of manual work makes grammar development a higher-level task, whose positive side is the integration of theoretically well informed analyses that hold generally across languages.

However, grammar rules are not the only component of an XLE grammar. Figure 1 provides an overview of the complete processing pipeline.

tokenizer & morphology (FST)
↓
transliteration (FST)
↓
syntax (XLE LFG)
↓
semantics (XFR ORDERED REWRITING)

Figure 1: Urdu XLE pipeline

At first, sentences are tokenized into words, these words are transliterated into a Roman version of the Arabic script (Malik et al., 2010) and then morphologically analyzed by a finite-state morphological analyzer (Bögel et al., 2007). The transliteration allows us to abstract away from some of the vagaries of the Urdu script as well as open up our grammar for the processing and generation of Hindi (cf. section 2.2.1).

The information gained from the morphological analyzer is passed on to the XLE syntax component, where the grammar rules generate c- and f-structure. The semantic XFR system will be presented in full detail in Section 3. Note that it is possible to reverse the pipeline and generate back out from an f-structure analysis (but not as yet from a semantic representation).

This syntactically deep approach is particularly well suited for languages with fairly free word order, such as Urdu, as it looks beyond the surface arrangement of words in a sentence and provides a deep functional and semantic analysis. As an example, see Figure 2 for a c- and f-structure of (1).

(1) اس نے تل آبیب میں سیب کھایا
us nE t3ul AbEb mEN sEb kHAyA
he Erg Tel Aviv in apple eat.Perf.F.Sg
'He ate an apple in Tel Aviv.'

The level we are most concerned with is the f-structure, as it is a first step towards a semantic analysis (f-structures have been shown to be equivalent to quasi logical forms; (van Genabith and Crouch, 1996)). In cases where parts of constituents are scattered across the sentence, e.g., as in discontinuous parts of an NP (Raza and Ahmed, 2011), the f-structure collects these pieces in the one grammatical function representation they belong to. This greatly facilitates the automatic semantic analysis because we can build on a deep and very detailed syntactic analysis that already abstracts from the surface sentential order.

Looking at Figure 2, the c-structure is shown on the left and models the linear order and hierarchical relationshiop of the constituents. In the AVM on the right, the f-structure, the main predicate of the sentence is *kHA* 'to eat', the subject (`SUBJ`) of the sentence is the pronoun *us* 'he/she', with the object (`OBJ`) *sEb* 'apple'. The location is analysed as an adjunct, an optional element in the sentence. Information on tense and aspect is captured in the `TNS-ASP` f-structure at the bottom. There is also some lexical semantic information contained in the analysis under `LEX-SEM`, namely that it is an agentive, ingestive verb.[1]

In addition to the f-structure, a computational semantic analysis abstracts even further away from the syntax and is able to provide information on the lexical semantics of the words involved by supplementing the analysis with information from external lexical resources, see section 3.

2.2 Lexical Resources for Urdu

2.2.1 Urdu WordNet

Due to the resource sparseness in Indo-Aryan languages, there are only a few lexical resources

[1] The `CHECK` feature collects grammar internal features for well-formedness checking and can be filtered out.

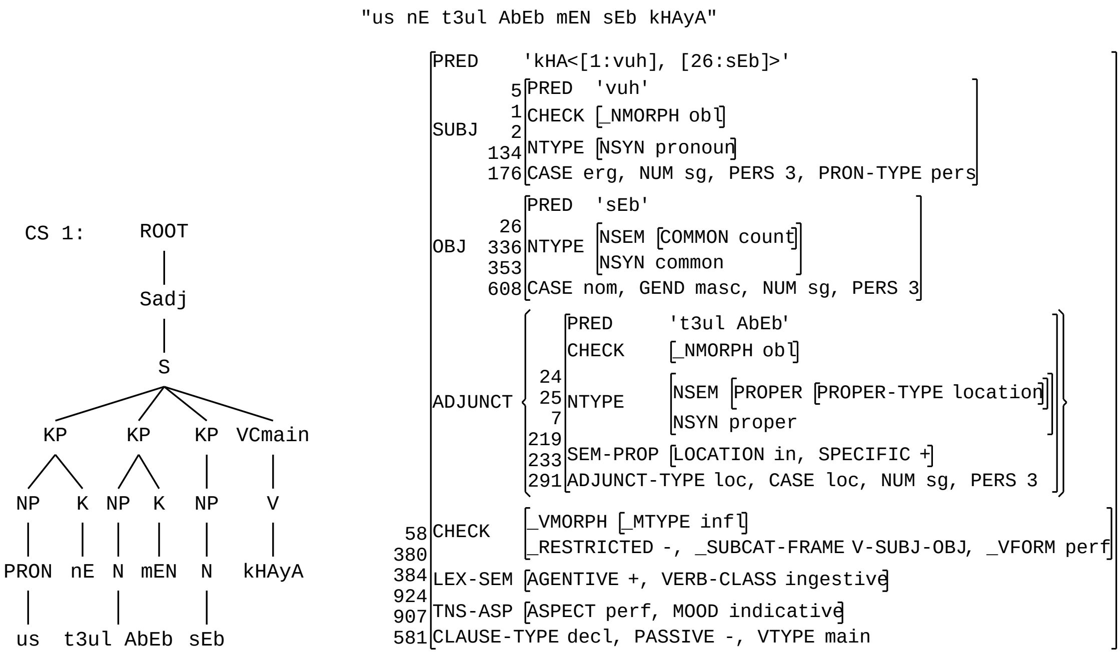

Figure 2: C- and f-structure for *us nE t3ul AbEb mEN sEb kHAyA* 'He ate an apple in Tel Aviv.'

already available, one of them Hindi Wordnet (Bhattacharyya et al., 2008; Bhattacharyya, 2010) which is inspired in methodology and architecture by the English WordNet (Fellbaum, 1998). Fortunately, Urdu and Hindi are structurally almost identical, although the two writing systems (a version of Arabic and Devanagari, respectively) differ markedly. This difference can be overcome by employing a transliterator from Arabic to Roman script and vice versa (Malik et al., 2010), combining it with a transliterator that maps Roman to Devanagari script (also vice versa), using XFST by (Beesley and Karttunen, 2003). Figure 3 sketches the pipeline of how we arrive at a preliminary Urdu WordNet (Ahmed and Hautli, 2010).

Urdu input
↓
transliteration to Hindi
↓
lookup in Hindi WordNet
↓
extraction of Hindi WordNet information
↓
removing synset description and example
↓
basic lexical resource for Urdu

Figure 3: Hindi/Urdu WordNet pipeline

By using this methodology it is possible to generate a preliminary Urdu WordNet (Ahmed and Hautli, 2010) that can be employed in various NLP applications, among them the semantic representation presented in this paper. A sample output for the noun *sEb* 'apple' is shown in Figure 4.

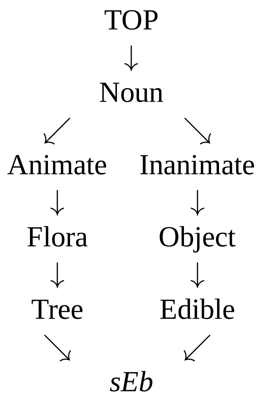

Figure 4: Sample Urdu WordNet output

First experiments have shown that this approach is a promising first step towards creating a basic lexical ontology for Urdu, however a thoroughly worked out Urdu WordNet will require additional work. For one, lexical items that are Hindi-specific will need to be flagged and words that are Urdu-specific will need to be introduced. In particular, the ezafe construction (Butt et al., 2008), illustrated in (2), will need to be dealt with. However, this is not trivial, as the ezafe *e* is often not written in Urdu, as is indeed the case in (2).

(2) وذیر اعظم
vazir-e azAm
minister.M.Sg-Ezafe great
'the prime minister' (lit. the great minister)

2.2.2 Towards a resource for Urdu verbs

One lexical resource used by the English LFG grammar is the English VerbNet (Kipper-Schuler, 2005), which categorizes English verbs according to Levin's verb classes (Levin, 1993). On the one hand, verbs are grouped according to their semantic relatedness, e.g. ingestive verbs or verbs of motion. Moreover, these related verbs are grouped into further subclasses according to their syntactic behavior. In addition to this semantic and syntactic classification, VerbNet also encodes information on the event structure and the thematic roles (Fillmore, 1985) of a verb. A similar resource is being developed for Hindi (Begum et al., 2008), however instead of Western style thematic roles, Panini's *karaka* relations are used.

For Urdu, we are currently working on creating a VerbNet style resource, carefully taking into account the characteristics of Urdu verbs regarding their syntactic behavior and including sufficient lexical semantic information so that the resource can be used in NLP tools. At the moment, most of the classification work is done by hand, because we also want to capture the very subtle variations which are likely to be lost in an automatic approach. For the case at hand, we are particularly interested in getting information on the thematic roles of a verb.

As an example, we consider the *rakH* 'put' class. A subgroup of these verbs allow for a locative alternation illustrated in (3)–(4). For the automatic semantic analysis it is solely important that the correct thematic roles are assigned to the arguments of the verb. Therefore we have to include this information in the verb resource for Urdu. By combining the information coming from the f-structure, where, for example, a locative adjunct is marked as such, with the information coming from the lexical resource, we can arrive at a semantic analysis that represents concepts rather than the actual sentence.

(3) میں نے گلاس میں پانی بھرا
mEN=nE gilAs mEN pAnI bHarA
I=Erg glas.M.Sg in water.M.Sg fill.Perf.M.Sg
'I filled water in the glass.'

⟨`Agent, Theme, Location`⟩

(4) میں نے پانی سے گلاس بھرا
mEN=nE pAnI=sE gilAs bHarA
I=Erg water.M.Sg=Instr glass.Nom fill.Perf.M.Sg
I filled the glass with water.

⟨`Agent, Location, Theme`⟩

The group of verbs in this class are: ڈھانپ DHANp 'to cover', بھر bHar 'to fill', سج saj 'to get decorated' and ڈھک DHak 'to cover'.

3 Urdu computational semantic analysis

3.1 General methodology

The primary aim of the semantic analyzer is to provide a more abstract level of linguistic representation, building on the information that is coming from the syntax, particularly from the f-structure. The Prolog-based XFR rewrite rules offer a suitable method for XLE grammars to arrive at a semantic representation (Crouch and King, 2006). Although they operate mainly on f-structures, c-structure information can also be used, e.g. to investigate scope issues further.

The XFR system is a language-independent component of XLE that can be used for various tasks, e.g. machine translation or the mapping of f-structures to semantic representations. The XFR semantic representation is driven neither by a specific semantic theory about meaning representation, nor by a theoretically motivated apparatus of meaning construction. It is a computational solution, which is why it is seen as a "semantic conversion" rathern than a "semantic construction".

XFR comprises a set of rewrite rules, the facts on the left hand side of a rule are rewritten to the facts on the right hand side. In addition, the rewrite rules are ordered, i.e. the first rule applies to the original input, the second rule takes as input the output of the first rule and so on.

For a concrete example of an XFR rewrite rule, we consider the f-structure in Figure 2. Given the case that we would want to systematically replace the subject and the object of the sentence with the right corresponding thematic roles, we could employ the rule in Figure 5.

```
PRED(%1,kHA), SUBJ(%1,%2), OBJ(%1,%3)
==>
context_head(%1,kHA),
role(Agent,kHA,%2),
role(Patient,kHA,%3).
```

Figure 5: Example of an XFR rewrite rule

The matrix f-structure is represented by the variable `%1`, its `SUBJ` f-structure is stored under variable `%2`, the `OBJ` under variable `%3`. If the facts do not match correctly, the rule does not apply. If the rule applies, the facts on the left hand side are consumed and rewritten to the facts on the right hand side of the rule. The representation that is generated is a flat representation of the predicate argument structure of the clause, i.e. it is not distributed across f-structures, . Despite the oversimplifying nature of the rule in Figure 5, the methodology remains the same for more complex rule constructions. In the following we present a more complex XFR rule.

3.2 The Urdu XFR system

Figure 6 presents a schematic view of the Urdu semantics pipeline.

syntax (XLE LFG)
↓
semantics (XFR ORDERED REWRITING)
↑
inclusion of lexical resources
(XFR ORDERED REWRITING)

Figure 6: Urdu Semantics pipeline

Input to the semantic representation is the syntactic XLE analysis as shown in Figure 2, which is stored in a Prolog format and can then be further processed by the XFR system. The output of the XFR semantic rules is shown in Figure 7. This representation does not yet contain information coming from lexical resources yet; its inclusion will be discussed in greater detail in Section 3.3.

In the semantic representation at hand, the sentential predicate *kHA* 'eat' is the `context_head` of the semantic representation, a term which is equivalent to the notion of the main predication in the formal semantics literature. The subcategorized arguments in the sentence are rewritten to `role` facts, the default roles of `sem_subj` and `sem_obj` will later be replaced by the thematic roles coming from the verb lexical resource.

Another main factor of the semantic analysis in that one should be able to see the domain of predication, i.e. the contexts in which the predications of the sentence hold. In the case at hand, there is only one context where predications can be true, namely context `t` (`in_context(t, ...)`) with its head *kHA* 'eat' (`context_head(t,kHA:87)`).[2]

Main clauses as well as relative clauses and other subordinate clauses open up new contexts in which predications are true or false. These clauses can be identified due to the syntactic analysis at f-structure level, where they are analyzed as COMPs or XCOMPs (complementizers). Lexical items such as negation markers also open up new contexts, e.g. the negation *nahIN* 'not'. By checking which predications hold in which contexts, sophisticated analyses of facts vs. beliefs and modal contexts can be achieved.

Another very important component of the syntactic as well as the semantic analysis is the inclusion of named entities in the lexicon. Hautli and Sulger (2011) have used automatic methods on a raw Urdu corpus to detect these so-called multiwords and also to classify them. They are very important for the system, because the components have a non-compositional meaning and should be treated as one unit.

This becomes apparent when looking at the predicate of the `ADJUNCT`, *t3ul AbEb* 'Tel Aviv'. It is analyzed as one unit and is the bare modifier of the verb phrase (`bare_mod(kHA:87,'t3ul AbEb':41)`). Due to the f-structure information [`PROPER-TYPE location`], it is clear that the modifier is locative, which is captured by the fact (`proper_name't3ul AbEb', location`).

The `skolem_info` facts store the part-of-speech information for each lexical item in the sentence and are the prerequisite for looking up words in lexical resources. The `original_fsattr` facts provide information according to which ambiguous information from the lexical resources can be disambiguated. The information about the subcategorization frame (`subcat`) is kept for the same reason, in case where a verb has multiple frames in the lexical resource, the system can choose the appropriate one according the subcategorization information coming from the syntax.

In cases where multiple valid semantic representations are generated, all analyses are displayed. Disambiguation on that level would require a more discourse-oriented analysis, which we do not provide at the moment.

[2]The numeral after each lexical item is simply a feature of bookkeeping, so that lexical items occurring twice in a sentence can be distinguished.

```
cf(1, context_head(t,kHA:87)),
cf(1, in_context(t,perf(kHA:87))),
cf(1, in_context(t,cardinality(sEb:70,sg))),
cf(1, in_context(t,cardinality('t3ul AbEb':41,sg))),
cf(1, in_context(t,cardinality(vuh:0,sg))),
cf(1, in_context(t,proper_name('t3ul AbEb':41,location,'t3ul AbEb')),
cf(1, in_context(t,role('sem_subj',kHA:87,vuh:0))),
cf(1, in_context(t,role('sem_obj',kHA:87,sEb:70))),
cf(1, in_context(t,role(bare_mod,kHA:87,'t3ul AbEb':41))),
cf(1, name_source('t3ul AbEb':41,lex)),
cf(1, name_type('t3ul AbEb':41,location)),
cf(1, original_fsattr('ADJUNCT',kHA:87,'t3ul AbEb':41)),
cf(1, original_fsattr('OBJ',kHA:87,sEb:70)),
cf(1, original_fsattr('SUBJ',kHA:87,vuh:0)),
cf(1, original_fsattr(gender,'t3ul AbEb':41,'-')),
cf(1, original_fsattr(human,'t3ul AbEb':41,'-')),
cf(1, original_fsattr(subcat,kHA:87,'V-SUBJ-OBJ')),
cf(1, skolem_info(kHA:87,kHA,verb,verb,t)),
cf(1, skolem_info(sEb:70,sEb,noun,common,t)),
cf(1, skolem_info('t3ul AbEb':41,'t3ul AbEb',name,location,t)),
cf(1, skolem_info(vuh:0,vuh,noun,pronoun,t)),
cf(1, subcat(kHA:87,'V-SUBJ-OBJ'))
```

Figure 7: Semantic representation for *us nE t3ul AbEb meN sEb kHAyA* 'He ate an apple in Tel Aviv.'

3.3 The inclusion of lexical resources

This section deals more closely with the inclusion of external lexical semantic information, making the general XFR methodology quite a powerful one because knowledge from various sources can be combined in one system.

By including knowledge contained in an Urdu WordNet and the Urdu verb resource, we can include hypernym relations such as that an apple is a fruit or the thematic roles of the arguments in a clause. This abstraction is not on the syntactic level any more, but is now at the level of lexical semantics. The benefit of such a representation is that we arrive at a meta-level of analysis where concepts are represented rather than linguistic structure.

For Urdu WordNet, we consider all senses that are produced for one item by the resource (i.e. see the two different senses for *sEb* 'apple' in Figure 4) and we take the direct hypernym of the lexical item. For the verb resource, we are mainly concerned with the thematic roles that are assigned to the arguments of the sentence.

In order to include external resources, they are reformatted as non-resourced Prolog facts (template name `uwn` for information from Urdu WordNet and `verbs` for thematic role information in Figure 8) that can be picked up by the XFR rewrite rules. For that, the templates are called on the right side of an XFR rule and compared with the information from the semantic representation. If the information matches, the rule applies and the lexical items are rewritten to include the conceptual information from the lexical resources.

See Figure 8 for an example of non-resourced facts that are being called by XFR rules and that rewrite information coming from the semantic representation in Figure 7. If the context head of the sentence is *kHA* 'eat' in a context with a variable `%Ctx`[3], and if within the same context `%Ctx` there is a semantic subject with variable `%S` and a semantic object with variable `%O` that are also captured in the `original_fsattr` and the `skolem_info` facts and given that the verb is found in the non-resourced facts, then rewrite the arguments to their thematic roles.

The second rewrite rule includes information from Urdu WordNet and inserts the hypernym of the verb *kHA* 'eat', namely that it is a verb of consumption.

The resulting semantic representation is presented in a less formal way in Figure 9. The Agent of the sentence, *vuh* 'he/she' performs a consumptive action, *kHA* 'eat' towards the Patient, *sEb* 'apple' and this act is performed at the location *t3ul AbEb* 'Tel Aviv'.

[3]The '+' in front of the first fact keeps the fact from being rewritten and can be called in later rule sequences.

```
|-uwn(kHA,Consumption);
|-verbs(kHA,Agent,Patient);

+context_head(%Ctx,kHA),
in_context(%Ctx,role(sem_subj,kHA,%S),
in_context(%Ctx,role(sem_obj,kHA,%O),
original_fsattr(SUBJ,%Pred,%S),
original_fsattr(OBJ,%Pred,%O),
skolem_info(kHA,kHA,verb,verb,%Ctx),
skolem_info(%O,%O,noun,common,%Ctx)),
skolem_info(%S,%S,noun,pronoun,%Ctx)),
verbs(kHA,%TRole1,%TRole2)
==>
in_context(%Ctx,role(%TRole1,kHA,%S),
in_context(%Ctx,role(%TRole2,kHA,%O).


context_head(%Ctx,%Pred),
uwn(%Pred,%Hyper)
==>
context_head(%Ctx,%Hyper).
```

Figure 8: Including lexical semantic information via XFR rules

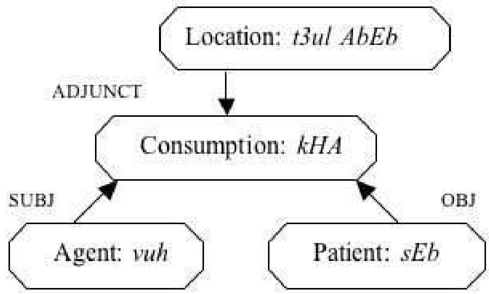

Figure 9: Final representation for *us nE t3ul AbEb meN sEb kHAyA* 'He ate an apple in Tel Aviv.'

4 Discussion

As for every computational grammar or semantic or analyzer, one wishes to have a thorough quantitative and qualitative evaluation justifying its robustness, coverage and accuracy, however we cannot provide such a justification in this paper. Although efforts are underway to create an independent gold standard for Urdu in the form of dependency triples as has been done for English (King et al., 2003), no such standard exists for Urdu to date. On the computational semantics side, this is also due to the fact that there has not been very much research in that direction.

The computational semantic analysis presented in this paper draws heavily on the syntactic analysis performed by the Urdu ParGram grammar. The more expressive the generated f-structures, the more detailed the semantic representations are. However, one could also use f-structures coming from other parsers whose output is reformatted according to the XLE standard and one could run the XFR system on these, potentially stochastic, f-structures as well.

As to adequate meaning representation, the overall move towards developing parallel semantics on top of parallel grammars within the ParGram community has only just started investigating appropriate representations of semantic concepts. The expressive power of a system also depends heavily on the external lexical resources that are available, in comparison to English, Urdu is far behind. However, with efforts like this computational semantic analyzer and the implementation of basic lexical resources we can contribute to the variety of tools available for Urdu.

5 Summary

In this paper we have presented a first approach to an automatic semantic analysis for Urdu, building on a deep and very detailed syntactic analysis by the Urdu ParGram Grammar. We have given a brief overview of the methodology of XFR rewrite rules, how they operate on top of LFG f-structures and what kind of semantic analysis they can provide. The inclusion of available lexical resources facilitates the generation of an abstract level of the representation of concepts rather than surface syntactic structure. We have also discussed some of the issues involved in building a semantic analyzer for a language where few other resources exist and where a lot of work has to go into the theoretical as well as the computationally efficient analysis of the language itself.

References

Tafseer Ahmed and Annette Hautli. 2010. An Experiment for a basic lexical resource for Urdu on the basis of Hindi WordNet. In *Proceedings of CLT 2010*, Islamabad, Pakistan.

Kenneth Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford.

Rafiya Begum, Samar Husain, Lakshmi Bai, and Dipti Misra Sharma. 2008. Developing Verb Frames for Hindi. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Pushpak Bhattacharyya, Prabhakar Pande, and Laxmi Lupu. 2008. *Hindi WordNet*. Linguistic Data Consortium, Philadelphia.

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Malta*.

Tina Bögel, Miriram Butt, Annette Hautli, and Sebastian Sulger. 2007. Developing a Finite-State Morphological Analyzer for Urdu and Hindi: Some Issues. In *Proceedings of FSMNLP07*. Postdam, Germany.

Tina Bögel, Miriram Butt, Annette Hautli, and Sebastian Sulger. 2009. Urdu and the Modular Architecture of ParGram. In *Proceedings of the Conference on Language and Technology (CLT09)*. CRULP, Lahore, Pakistan.

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representations from a CCG parser. In *COLING '04: Proceedings of the 20th International Conference on Computational Linguistics*, page 1240.

Joan Bresnan and Ronald M. Kaplan, 1982. *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge.

Miriam Butt and Tracy Holloway King. 2002. Urdu and The Parallel Grammar Project. In *Proceedings of COLING2002, 3rd workshop on Asian language resources and international standardization*, pages 39–45, Taipei, Taiwan.

Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook.* CSLI Publications, Stanford.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of COLING2002, Workshop on Grammar Engineering and Evaluation*, pages 1–7, Taipei, Taiwan.

Miriam Butt, Tina Bögel, and Sebastian Sulger. 2008. Urdu Ezafe and the Morphology-Syntax Interface. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of LFG08*. CSLI Publications, Stanford.

Dick Crouch and Tracy Holloway King. 2006. Semantics via F-structure Rewriting. In *LFG06 Proceedings*. CSLI Publications, Stanford.

Dick Crouch, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2011. XLE Documentation. http://www2.parc.com/isl/groups/nltt/xle/doc/.

Mary Dalrymple, 2001. *Lexical Functional Grammar*, volume 34. Academic Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press.

Charles J. Fillmore. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica*, VI(2):222–254.

Annette Hautli and Sebastian Sulger. 2011. Extracting and Classifying Urdu Multiword Expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Langauge Technologies (ACL-HLT '11): Student Session*, Portland, Oregon.

Tracy Holloway King, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ron Kaplan. 2003. The PARC700 Dependency Bank. In *Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*.

Karin Kipper-Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Beth Levin. 1993. *English Verb Classes and Alternations*. Chicago: The University of Chicago Press.

Muhammad Kamran Malik, Tafseer Ahmed, Sebastian Sulger, Tina Bögel, Atif Gulzar, Ghulam Raza, Sarmad Hussain, and Miriam Butt. 2010. Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Malta*.

Ghulam Raza and Tafseer Ahmed. 2011. Argument Scrambling within Urdu NPs. In *Proceedings of LFG11*, Hong Kong.

Stefan Riezler, Tracy Holloway King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadephia, PA.

Josef van Genabith and Dick Crouch. 1996. Direct and underspecified interpretations of LFG f-structures. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, volume 1, pages 262–267, Copenhagen, Denmark.