# Extracting and Classifying Urdu Multiword Expressions

**Annette Hautli**
Department of Linguistics
University of Konstanz, Germany
annette.hautli@uni-konstanz.de

**Sebastian Sulger**
Department of Linguistics
University of Konstanz, Germany
sebastian.sulger@uni-konstanz.de

## Abstract

This paper describes a method for automatically extracting and classifying multiword expressions (MWEs) for Urdu on the basis of a relatively small unannotated corpus (around 8.12 million tokens). The MWEs are extracted by an unsupervised method and classified into two distinct classes, namely locations and person names. The classification is based on simple heuristics that take the co-occurrence of MWEs with distinct postpositions into account. The resulting classes are evaluated against a hand-annotated gold standard and achieve an f-score of 0.5 and 0.746 for locations and persons, respectively. A target application is the Urdu ParGram grammar, where MWEs are needed to generate a more precise syntactic and semantic analysis.

## 1 Introduction

Multiword expressions (MWEs) are expressions which can be semantically and syntactically idiosyncratic in nature; acting as a single unit, their meaning is not always predictable from their components. Their identification is therefore an important task for any Natural Language Processing (NLP) application that goes beyond the analysis of pure surface structure, in particular for languages with few other NLP tools available.

There is a vast amount of literature on extracting and classifying MWEs automatically; many approaches rely on already available resources that aid during the acquisition process. In the case of the Indo-Aryan language Urdu, a lack of linguistic resources such as annotated corpora or lexical knowledge bases impedes the task of detecting and classifying MWEs. Nevertheless, statistical measures and language-specific syntactic information can be employed to extract and classify MWEs.

Therefore, the method described in this paper can partly overcome the bottleneck of resource sparsity, despite the relatively small size of the available corpus and the simplistic approach taken. With the help of heuristics as to the occurrence of Urdu MWEs with characteristic postpositions and other cues, it is possible to cluster the MWEs into two groups: locations and person names. It is also possible to detect junk MWEs. The classification is then evaluated against a hand-annotated gold standard of Urdu MWEs.

An NLP tool where the MWEs can be employed is the Urdu ParGram grammar (Butt and King, 2007; Bögel et al., 2007; Bögel et al., 2009), which is based on the Lexical-Functional Grammar (LFG) formalism (Dalrymple, 2001). For this task, different types of MWEs need to be distinguished as they are treated differently in the syntactic analysis.

The paper is structured as follows: Section 2 provides a brief review of related work, in particular on MWE extraction in Indo-Aryan languages. Section 3 describes our methodology, with the evaluation following in Section 4. Section 5 presents the Urdu ParGram Grammar and its treatment of MWEs, followed by the discussion and the summary of the paper in Section 6.

## 2 Related Work

MWE extraction and classification has been the focus of a large amount of research. However, much work



*Proceedings of the ACL-HLT 2011 Student Session*, pages 24–29,
Portland, OR, USA 19-24 June 2011. 

has been conducted for well-resourced languages such as English, benefiting from large enough corpora (Attia et al., 2010), parallel data (Zarrieß and Kuhn, 2009) and NLP tools such as taggers or dependency parsers (Martens and Vandeghinste (2010), among others) and lexical resources (Pearce, 2001).

Related work on Indo-Aryan languages has mostly focused on the extraction of complex predicates, with the focus on Hindi (Mukerjee et al., 2006; Chakrabarti et al., 2008; Sinha, 2009) and Bengali (Das et al., 2010; Chakraborty and Bandyopadhyay, 2010). While complex predicates also make up a large part of the verbal inventory in Urdu (Butt, 1993), for the scope of this paper, we restrict ourselves to classifying MWEs as locations or person names and filter out junk bigrams.

Our approach deviates in several aspects to the related work in Indo-Aryan: First, we do not concentrate on specific POS constructions or dependency relations, but use an unannotated middle-sized corpus. For classification, we use simple heuristics by taking the postpositions of the MWEs into account. These can provide hints as to the nature of the MWE.

## 3 Methodology

### 3.1 Extraction and Identification of MWE Candidates

The bigram extraction was carried out on a corpus of around 8.12 million tokens of Urdu newspaper text, collected by the Center for Research in Urdu Language Processing (CRULP) (Hussain, 2008). We did not perform any pre-processing such as POS tagging or stop word removal.

Due to the relatively small size of our corpus, the frequency cut-off for bigrams was set to 5, i.e. all bigrams that occurred five times or more in the corpus were considered. This rendered a list of 172,847 bigrams which were then ranked with the $X^2$ association measure, using the UCS toolkit.[1]

The reasons for employing the $X^2$ association measure are twofold. First, papers using comparatively sized corpora reported encouraging results for similar experiments (Ramisch et al., 2008; Kizito et al., 2009). Second, initial manual comparison between MWE lists ranked according to all measures implemented in the UCS toolkit revealed the most convincing results for the $X^2$ test.

For the time being, we focus on bigram MWE extraction. While the UCS toolkit readily supports work on Unicode-based languages such as Urdu, it does not support trigram extraction; other freely available tools such as TEXT-NSP[2] do come with trigram support, but cannot handle Unicode script. As a consequence, we currently implement our own scripts to overcome these limitations.

### 3.2 Syntactic Cues

The clustering approach taken in this paper is based on Urdu-specific syntactic information that can be gathered straightforwardly from the corpus. Urdu has a number of postpositions that can be used to identify the nature of an MWE. Typographical cues such as initial capital letters do not exist in the Urdu script.

**Locative postpositions** The postposition پر (*par)* either expresses location on something which has a surface or that an object is next to something.[3] In addition, it expresses movement to a destination.

(1) نادیه تل ابیب پر گاي
nAdiyah t3ul AbEb par gAyI
Nadya Tel Aviv to go.Perf.Fem.Sg
'Nadya went to Tel Aviv.'

میں (*mEN*) expresses location in or at a point in space or time, whereas تك (*tak*) denotes that something extends to a specific point in space. سي (*sE*) shows movement away from a certain point in space.

These postpositions mostly occur with locations and are thus syntactic indicators for this type of MWE. However, in special cases, they can also occur with other nouns, in which case we predict wrong results during classification.

**Person-indicating syntactic cues** To classify an MWE as a person, we consider syntactic cues that usually occur after such MWEs. The ergative marker نے (*nE*) describes an agentive subject in transitive

[1] Available at `http://www.collocations.de`. See Evert (2004) for documentation.

[2] Available at `http://search.cpan.org/dist/Text-NSP`. See Banerjee and Pedersen (2003) for documentation.

[3] The employed transliteration scheme is explained in Malik et al. (2010).

| | Locative | | | Instr. | Ergative | Possessive | | | Acc./Dat. |
|---|---|---|---|---|---|---|---|---|---|
| | پر (*par*) | میں (*mEN*) | تك (*tak*) | سي (*sE*) | ني (*nE*) | كا (*kA*) | كي (*kE*) | كي (*kI*) | كو (*kO*) |
| LOC | √ | √ | √ | √ | — | — | — | — | — |
| PERS | — | — | — | √ | √ | √ | √ | √ | √ |
| JUNK | — | — | — | — | — | — | — | — | — |

Table 1: Heuristics for clustering Urdu MWEs by different postpositions

sentences; therefore, it forms part of our heuristic for finding person MWEs.

(2) نادیه نی یاسین کو مارا

nAdiyah nE yAsIn kO mArA
Nadya Erg Yasin Acc hit.Perf.Masc.Sg
'Nadya hit Yasin.'

The same holds for the possessive markers كا (*kA*), كي (*kE*) and كي (*kI*).

The accusative and dative case marker كو (*kO*) is also a possible indicator that the preceding MWE is a person.

These cues can also appear with common nouns, but the combination of MWE and syntactic cue hints to a person MWE. However, consider cases such as *New Delhi said that the taxes will rise.*, where *New Delhi* is treated as an agent with *nE* attached to it, providing a wrong clue as to the nature of the MWE.

### 3.3 Classifying Urdu MWEs

The classification of the extracted bigrams is solely based on syntactic information as described in the previous section. For every bigram, the postpositions that it occurs with are extracted from the corpus, together with the frequency of the co-occurrence.

Table 1 shows which postpositions are expected to occur with which type of MWE. The first stipulation is that only bigrams that occur with one of the locative postpositions plus the ablative/instrumental marker سي (*sE*) one or more times are considered to be locative MWEs (LOC). In contrast, bigrams are judged as persons (PERS) when they co-occur with all postpositions apart from the locative postpositions one or more times. If a bigram occurs with none of the postpositions, it is judged as being junk (JUNK). As a consequence this means that theoretically valid MWEs such as complex predicates, which never occur with a postposition, are misclassified as being JUNK.

Without any further processing, the resulting clusters are then evaluated against a hand-annotated gold standard, as described in the following section.

## 4 Evaluation

### 4.1 Gold Standard

Our gold standard comprises the 1300 highest ranked Urdu multiword candidates extracted from the CRULP corpus, using the $X^2$ association measure. The bigrams are then hand-annotated by a native speaker of Urdu and clustered into the following classes: locations, person names, companies, miscellaneous MWEs and junk. For the scope of this paper, we restrict ourselves to classifying MWEs as either locations or person names,. This also lies in the nature of the corpus: companies can usually be detected by endings such as "Corp." or "Ltd.", as is the case in English. However, these markers are often left out and are not present in the corpus at hand. Therefore, they cannot be used for our clustering. The class of miscellaneous MWEs contains complex predicates that we do not attempt to deal with here.

In total, the gold standard comprises 30 companies, 95 locations, 411 person names, 512 miscellaneous MWEs (mostly complex predicates) and 252 junk bigrams. We have not analyzed the gold standard any further, and restricting it to $n < 1300$ might improve the evaluation results.

### 4.2 Results

The bigrams are classified according to the heuristics outlined in Section 3.3. Evaluating against the hand-annotated gold standard yields the results in Table 2.

While the results are encouraging for persons with an f-score of 0.746, there is still room for improvement for locative MWEs. Part of the problem for per-

| | Precision | Recall | F-Score | #total | #found |
|---|---|---|---|---|---|
| LOC | 0.453 | 0.558 | **0.5** | 95 | 43 |
| PERS | 0.727 | 0.765 | **0.746** | 411 | 298 |
| JUNK | 0.472 | 0.317 | **0.379** | 252 | 119 |

Table 2: Results for MWE clustering

son names is that Urdu names are generally longer than two words, and as we have not considered trigrams yet, it is impossible to find a postposition after an incomplete though generally valid name. Locations tend to have the same problem, however the reasons for missing out on a large part of the locative MWEs are not quite clear and are currently being investigated.

Junk bigrams can be detected with an f-score of 0.379. Due to the heterogeneous nature of the miscellaneous MWEs (e.g., complex predicates), many of them are judged as being junk because they never occur with a postposition. If one could detect complex predicate and, possibly, other subgroups from the miscellaneous class, then classifying the junk MWEs would become easier.

## 5 Integration into the Urdu ParGram Grammar

The extracted MWEs are integrated into the Urdu ParGram grammar (Butt and King, 2007; Bögel et al., 2007; Bögel et al., 2009), a computational grammar for Urdu running with XLE (Crouch et al., 2010) and based on the syntax formalism of LFG (Dalrymple, 2001). XLE grammars are generally handwritten and not acquired a machine learning process or the like. This makes grammar development a very conscious task and it is imperative to deal with MWEs in order to achieve a linguistically valid and deep syntactic analysis that can be used for an additional semantic analysis.

MWEs that are correctly classified according to the gold standard are automatically integrated into the multiword lexicon of the grammar, accompanied by information about their nature (see example (3)).

In general, grammar input is first tokenized by a standard tokenizer that separates the input string into single tokens and replaces the white spaces with a special token boundary symbol. Each token is then passed through a cascade of finite-state morphological analyzers (Beesley and Karttunen, 2003). For MWEs, the matter is different as they are treated as a single unit to preserve the semantic information they carry. Apart from the meaning preservation, integrating MWEs into the grammar reduces parsing ambiguity and parsing time, while the perspicuity of the syntactic analyses is increased (Butt et al., 1999).

In order to prevent the MWEs from being independently analyzed by the finite-state morphology, a look-up is performed in a transducer which only contains MWEs with their morphological information. So instead of analyzing *t3ul* and *AbEb* separately, for example, they are analyzed as a single item carrying the morphological information `+Noun+Location`.[4]

(3) `t3ul` AbEb: /`t3ul` AbEb/ +Noun +Location

The resulting stem and tag sequence is then passed on to the grammar. See (4) for an example and Figures 1 and 2 for the corresponding c- and f-structure; the `+Location` tag in (3) is used to produce the location analysis in the f-structure. Note also that `t3ul AbEb` is displayed as a multiword under the `N` node in the c-structure.

(4) نادیه تل ابیب پر گاي
nAdiyah t3ul AbEb par gAyI
Nadya Tel Aviv to go.Perf.Fem.Sg
'Nadya went to Tel Aviv.'

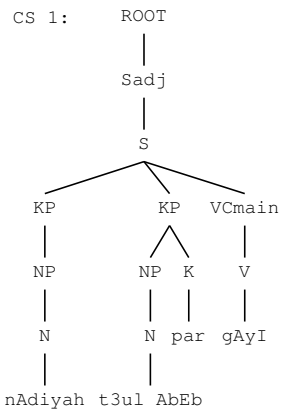


Figure 1: C-structure for (4)

[4]The ` symbol is an escape character, yielding a literal white space.

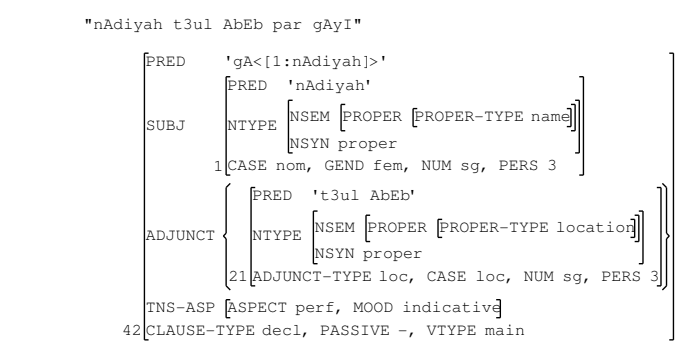


Figure 2: F-structure for (4)

## 6 Discussion, Summary and Future Work

Despite the simplistic approach for extracting and clustering Urdu MWEs taken in this paper, the results are encouraging with f-scores of 0.5 and 0.746 for locations and person names, respectively. We are well aware that this paper does not present a complete approach to classifying Urdu multiwords, but considering the targeted tool, the Urdu ParGram grammar, this methodology provides us with a set of MWEs that can be implemented to improve the syntactic analyses.

The methodology provided here can also guide MWE work in other languages facing the same resource sparsity as Urdu, given that distinctive syntactic cues are available in the language.

For Urdu, the syntactic cues are good indications of the nature of the MWE; future work on this subtopic might prove beneficial to the clustering regarding companies, complex predicates and junk MWEs. Another area for future work is to extend the extraction and classification to trigrams to improve the results especially for locations and person names. We also consider harvesting data sources from the web such as lists of cities, common names and companies in Pakistan and India. Such lists are not numerous for Urdu, but they may nevertheless help to generate a larger MWE lexicon.

## Acknowledgments

We would like to thank Samreen Khan for annotating the gold standard, as well as the anonymous reviewers for their valuable comments. This research was in part supported by the Deutsche Forschungsgemeinschaft (DFG).

## References

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef van Genabith. 2010. Automatic Extraction of Arabic Multiword Expressions. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*.

Satanjeev Banerjee and Ted Pedersen. 2003. The Design, Implementation and Use of the Ngram Statistics Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*.

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

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

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

Miriam Butt and Tracy Holloway King. 2007. Urdu in a Parallel Grammar Development Environment. *Language Resources and Evaluation*, 41(2):191–207.

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

Miriam Butt. 1993. *The Structure of Complex Predicates in Urdu*. Ph.D. thesis, Stanford University.

Debasri Chakrabarti, Vaijayanthi M. Sarma, and Pushpak Bhattacharyya. 2008. Hindi Compound Verbs and their Automatic Extraction. In *Proceedings of COLING 2008*, pages 27–30.

Tanmoy Chakraborty and Sivaji Bandyopadhyay. 2010. Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule-Based Approach. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 72–75.

Dick Crouch, Mary Dalrymple, Ronald M. Kaplan, Tracy Holloway King, John T. Maxwell III, and Paula Newman, 2010. *XLE Documentation*. Palo Alto Research Center.

Mary Dalrymple. 2001. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press.

Dipankar Das, Santanu Pal, Tapabrata Mondal, Tanmoy Chakraborty, and Sivaji Bandyopadhyay. 2010. Automatic Extraction of Complex Predicates in Bengali. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 37–45.

Stefan Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, IMS, University of Stuttgart.

Sarmad Hussain. 2008. Resources for Urdu Language Processing. In *Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08*.

John Kizito, Ismail Fahmi, Erik Tjong Kim Sang, Gosse Bouma, and John Nerbonne. 2009. Computational Linguistics and the History of Science. In Liborio Dibattista, editor, *Storia della Scienza e Linguistica Computazionale*. FrancoAngeli.

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)*.

Scott Martens and Vincent Vandeghinste. 2010. An Efficient, Generic Approach to Extracting Multi-Word Expressions from Dependency Trees. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*, pages 84–87.

Amitabha Mukerjee, Ankit Soni, and Achla M. Raina. 2006. Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties (MWE '06)*, pages 28–35.

David Pearce. 2001. Synonymy in Collocation Extraction. In *WordNet and Other Lexical Resources: Applications, Extensions & Customizations*, pages 41–46.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An Evaluation of Methods for the Extraction of Multiword Expressions. In *Proceedings of the Workshop on Multiword Expressions: Towards a Shared Task for Multiword Expressions (MWE 2008)*.

R. Mahesh K. Sinha. 2009. Mining Complex Predicates in Hindi Using a Parallel Hindi-English Corpus. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pages 40–46.

Sina Zarrieß and Jonas Kuhn. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pages 23–30.