URDU TREEBANK

Muddassira Arshad, Aasim Ali
Punjab University College of Information Technology (PUCIT), University of the Punjab, Lahore
muddassira@pucit.edu.pk, aasim.ali@pucit.edu.pk
(Presented at the 5th International. Multidisciplinary Conference, 29-31 Oct., at, ICBS, Lahore)

***ABSTRACT**: Treebank is a parsed corpus of text annotated with the syntactic information in the form of tags that yield the phrasal information within the corpus. The first outcome of this study is to design a phrasal and functional tag set for Urdu by minimal changes in the tag set of Penn Treebank, that has become a de-facto standard. Our tag set comprises of 22 phrasal tags and 20 functional tags. Second outcome of this work is the development of initial Treebank for Urdu, which remains publically available to the research and development community. There are 500 sentences of Urdu translation of religious text with an average length of 12 words. Context free grammar containing 109 rules and 313 lexical entries, for Urdu has been extracted from this Treebank. An online parser is used to verify the coverage of grammar on 50 test sentences with 80% recall. However, multiple parse trees are generated against each test sentence.*

1 INTRODUCTION

The term "Treebank" was introduced in 2003. It is defined as "linguistically annotated corpus that includes some grammatical analysis beyond the part of speech level" [1]. Therefore, treebank are employed to represent annotated corpus yielding syntactic or semantic information, and are useful in parser analysis, extracting the accurate information from the text, finding the word senses etc.

The paper focuses on development of a fundamental resource for Urdu language - an Urdu Language Treebank, using the existing POS tagged data. PUCIT already developed POS tagged corpus based on selected ahadith translated in Urdu.[2] The Urdu Treebank uses Penn treebank guidelines [3] with some extensions adapted to incorporate Urdu Language constructs and classifies sentences, and various forms of clauses and phrases. These results were then used for generation of CFGs, followed by testing the unseen data after providing the test data and the extracted CFG using a third party online utility.

The paper is organized in a systematic way. Section 2 covers the background of treebank development, the design issues to be catered while developing a treebank. Section 3 encompasses the methodology used for Design of Urdu Treebank, proposed design decisions and the tag set design to yield the treebank. Section 4 entails the treebank development by manually annotating the POS tagged data in accordance with our design decisions. Examples of the tagged data using Phrase Structure representation are presented using bracket notation. Finally, Section 5 entails the conclusions and recommendations for future.

2 LITERATURE REVIEW

This section reviews the literature related to Treebank, Urdu language, and the efforts to develop Treebank for Urdu.

2.1 Treebank

Initially the term “treebank” was used to represent the manual annotation of text corpus for grammatical analysis to specify the various phrases like verb phrases, noun phrases, to entail grammar within the sentence. This manual annotation is different from the "parsed corpus", where the "parsed corpus" is used when the text is automatically analyzed with the help of some tool. Now a days, these terms are used interchangeably [4].

In Machine Learning, treebank serves a valuable resource for the development and analysis of Linguistics Analyzers [5]. Treebank also have application in induction of probabilistic grammars for parsing[5], where statistical models are used for broad-coverage parsing [6].

In literature, various techniques have been applied for treebank annotations. Constituency annotation bracketing have been used in projects like (Penn) English tree bank [3], and Lancaster Parsed Corpus [7]. In this scheme, nested level of tags are applied starting from the words annotated with Parts of Speech (POS) tags, followed by phrasal tags for phrases, and then clause and sentence level annotations. Constituency annotation has drawbacks of increased count of same phrase category’s distinct expansions [5].

In literature, functional annotation scheme is also applied in projects like Penn treebank II [8] considering the approach more useful for shallow semantic analysis like. predicate argument structure [9]. Lexical Functional Grammar focuses on both Functional as well as constituency structure considering both as primitive ones, where the functional structure is more prevalent for Treebank annotations [5].

Treebank development focuses to make considerations to ensure consistency by using similar language phenomenon throughout the available text by incorporating manual and automatic annotation processes. This approach is applicable for the languages where the parsers are available. In that case, the text is automatically analyzed and later experts make the correction in case there are any errors/ambiguities in the parsed results. An alternate of post correction of automatically annotated text is the "human disambiguation" where the available choices are shown to the users, who make their preferences to make corrections in case the automated generation had problems/multiple phrases annotation. "Human disambiguation" process increases the performance as the corpus size increases, but has disadvantages as it varies from individual to individual, and each individual has his own preferences.

In literature, the preferred approach in this regard is; allowing different annotators to work independently and then should compare their work. However, this process is quite expensive. In practice, automated analysis techniques are used for possible errors detection [5].

Marcus used Fidditch-deterministic parser [3] [10] to provide an initial parse, followed by hand corrections of by the data annotators. In addition, Penn treebank annotates missing entries as null elements for understood subject of infinitive/imperative parser, or missing variants.(e.g missing subordinate conjunction word like “that” in subordinate clause, or for the Traces indicating the misplaced constituents

are annotated with T indicating traces). Moreover, the proposed position of wh-constituent was also suggested [3].

2.2 Urdu Language

Urdu language is continuously evolving by incorporating words from languages like Persian, Arabic, English, Snaskirit. In addition, Urdu language is a free order language; means the order of subject (S), object (O), and verb (V) is free, though it usually maintains its natural order that is SOV. Urdu language follows the Persian-Arabic script. A very obvious feature of this script is that sentence writing direction is bidirectional, mostly from right to left and it is always cursive. Urdu has various morphological features like gender, number, cases, and honor. Urdu also deals with various types of cases like nominative, oblique, vocative, ergative, dative, accusative, instrumental, ablative, locative [6] [11,12].

Urdu frequently uses postposition in addition to the preposition. Also, it has a distinct Case Phrase (KP) which is a noun phrase followed by a case marker, thus identifying the purpose of noun used e.g. direct object, indirect object, instrument etc. These case phrases appear in sentence (S), verb phrases (VP) [13]. An Urdu NP may comprise of clauses, case phrases, noun phrases, or primitive constituents like nouns, verbs etc.

There exist quite variation in representing the Urdu Noun phrases where nouns may have 2-5 subcategories [14,15]. T. In order to incorporate simplicity, more applicability of the treebank, we opted for theory neutral approach.

Similarly pronouns, adverbs and adjectives are also discussed with their further subtypes. [14,15,16,17]. Term **case** [6] refers to an inflectional category-system [18]. Certain postpositions like ne, ko, se, ka, kay, kei, me~, par, tak plays special role while analyzing the phrases, clauses and the sentencesFor Example, whenever the postposition *ne(نے)* occurs after the noun phrase, it indicates that the noun phrase contains the subject information. This helps in identifying the subjects, objects, instruments, possessive nouns (of genitives) [18].

2.3 Treebank Efforts for Urdu

Some efforts have been made for the development of Urdu treebank so far. In one of the efforts, a treebank comprising of 1011 sentences was developed using hybrid dependency approach [19] [20]. Particles were given much importance as they communicate special semantic meaning. After CFG extraction, probabilities were calculated, which were then converted to the Definite Clause Grammar (DCG) for rules execution. The case information was delegated to the POS level. Genitives are differently annotated i.e. instead of following NP followed by case marker (Ka, ke, kai) and then the NP to show the possessive case, NP followed by the case indicated (ka, ke, kai) is used. [23] proposed following NP Case NP annotation for genitive case phrase.

Primarily designed for Hindi Urdu support uses, Dependency Structure (DS) [21] motivated from Panini Grammar and, Phrase Structure (PS) motivated from Chomsky's work. In addition, the treebank will also use Proposition Bank PropBank (PB) to annotate corpus. PropBank is useful in specifying verb-meaning specific verbal roles by specifying predicate node, the agent and the patient of the predicates [22]. The PropBank guidelines are still being reviewed whereas Phrase Structure and Dependency Structure annotations are already developed. Bhatt [22] also presented automatic conversion of PS to DS and PropBank and vice versa. The pre-release Anacora's document does not contain the Urdu support [24]. So far the resource comprising of 353k Hindi and almost 60k of Urdu words are annotated.

3 METHODOLOGY

Treebank design encompasses linguistic research, development of language technology, availability of analysis tool, corpus availability etc.

For the development of the treebank resource, following design choices were considered:

3.1 Corpus Selection

Mostly a choice is made to select between Corpus of written language or the spoken one. Similarly, considerations are made to finalize the Text genres i.e. opting a particular genre or a balanced collection of genre. In practice, most of the English treebanks uses the corpus of existing treebanks (e.g. Brown's), but the corpus may also be based on contemporary text selected from newspaper (e.g. Penn) [3].

We have selected the Urdu translation of ahadith data which is already POS tagged. [2]

3.2 Corpus Size

The choice of corpus size varies from building a large corpus with less detailed annotation e.g. Penn treebank [3], to a small corpus with much detailed annotation e.g. SUSANNE Corpus [35].

In our project, corpus size is 500 sentences which further contain sub-sentences, phrases and clauses. They comprise more than 6000 tokens of text.

3.3 Annotation Scheme

Fundamental question while specifying the scheme of annotation is the choice between linguistic analysis or the representation analysis, where the linguistic analysis focuses on the nature of syntactic representation, linguistic categories choices, annotation guidelines, whereas, representation analysis encompasses the annotation representation. The alternatives available are a particular markup language/ plain text, storage is one file/several files etc.

In practice, layered annotation schemes are preferred. The base layer is usually the word-level layer (e.g. Parts of Speech), which is quite similar across the available treebanks. On top of the base layer, syntax (sometimes semantics) is encoded which makes it syntactic layer. This layer is different across the treebanks [5].

For an annotation scheme, design consideration is made to make it a particular "theory specific" by selecting a particular theoretical framework, or "theory neutral" for which achieve the broader consensus of the various available theories is focused. The decision of annotation scheme selection requires the identification of target application areas e.g. linguistic researches, etc.

In our context we have used word level POS tagged Ahadith data, and on top of this layer phrase structure annotation is applied which is theory neutral so that the broader consensus could be established for acceptability of the treebank. Moreover, this would yield a simplified solution.

3.4 Design Decisions

The following design decisions were taken before the treebank was developed

1. Treebank tags should be theory neutral, thus tags are assigned to the broader categories e.g. instead of using separate phrasal tags for phrases with proper noun or common nouns [14] [15], just noun phrase tag is used.
2. Phrase structure (PS) representation is used to annotate text, therefore adding simplicity, as well as a basis for annotation using predicate argument structure. Although Dependency Structure (DS) annotation is frequently used for the free order languages, however use of case phrases (KP) will be used to determine the order/role of the noun phrases. Moreover, use of PS annotation will simplify the task.
3. Annotation scheme will be similar to the Penn treebank annotation i.e.) (Sentence [Subordinate clause (S [*] (Phrasal tag[-functional tag] [Phrasal Tags]*)*) -) where – is the sentence marker.
4. Every sentence will be enclosed in Sentence tag S.
5. In Urdu, Nouns are followed by the case markers, thus resulting in case phrases [6]. KP is used to identify roles of noun phrases, which could be subject, object, instrument, location etc. However, like Penn Treebank [3] direct & indirect objects are not explicitly labeled.
6. Compound Pre-postposition tag **CMP** contains other Pre-postpositions
7. CMP is introduced to ensure compound postpositions Unknown phrases will be marked with **X**
8. Null subjects are marked with (KP-NOM (**NP-SBJ** *)), where * is a null marker. Misplaced elements, topicalized words are annotated with T to specify the trace information.
9. Pronouns will be tagged using Noun phrases
10. For the sake of simplicity, all forms of Verbs annotated by VBL, VBD, VBZ, VBN, and VBP are annotated with Verb Phrase VP.
11. Words annotated with POS tag KER are the light verbs, and are therefore tagged within Verb Phrases.
12. POS tagged VALA words representing the noun or adjectival information is handled in their respective noun or adjectival phrase. Moreover, VALA word when showing work to be started is handled within VP
13. Foreign words are handled w.r.t the context they appear. The phrases are marked with X if they cannot be understood.
14. Titles like حضرت or praises like صلی <FW>اللہ <FW>علیہ<FW> و<FW>سلم <FW> are labeled with Honor Phrase (HP) tags
15. Subordinate clauses indicated by simple one words like if (agar), or the compound words like (yahan tek kai - یہاں تک کہ) will also contain sentences annotated by S after the SBAR identified words.
16. Quantifier phrases (QP) were used for single word as well as multiword quantifiers. In addition units were also annotated with the quantities (if any) in QP.
17. Defined levels of attachment are:
    a. Sentence indicated by **S** could be attached to case phrase (**KP**), Verb Phrase (**VP**), fronted constituents like topicalized phrases, conjunctions like and (اور) /or (یا), punctuations etc.
    b. Every valid sentence must have a **KP** and subject Noun Phrase (**NP**) denoted as **NP-SBJ**
    c. Pre-postpositional Phrases (**PP**) may be attached within NP or outside NP/ KP depending upon the meaning
    d. Honor Phrase (**HP**) representing phrase to give respect/honor is enclosed within NP
    e. NP may contains other NP
    f. VP may contain other VPs or NPs

Keeping in view the design decisions, tagset of Penn treebank and the corpus available following tag set was finalized:

Table1: Tags which are same as in Penn Treebank

Tag used	Tag Description
ADJP	Adjectival Phrase
ADVP	Adverbial Phrase
CONJP	Conjunction phrase
FRAG	Fragment
INTJP	Interjection Phrase
NP	Noun Phrase
PRN	Parenthetical Phrase
QP	Quantifier Phrases
S	Sentence
SBAR	Subordinate Conjunction
SBARQ	Question subordinate Conjunction
SQ	Yes/No question phrase
VP	Verb Phrase
WHADJP	WH Adjectival
WHADVP	WH Adverbial Phrase
WHNP	WH Noun Phrase
X	Unknown

3.5 Tags added for Urdu data set

Keeping in view the difference of Urdu with the English language, following tags are suggested to be added in our tagset.

Table2: tags incorporated to the existing treebanks for ahadith specific work

Tag	Description
CMP	Compound pre-postposition phrase. Not included in Penn treebank, added for use for Urdu Corpus especially where more than one case markers are specified after noun phrase.
KP	Case Phrase, Not included in Penn treebank, added for use for Urdu Corpus
HP	Honor Phrase, Not included in Penn treebank, added for use for Urdu Corpus
PP	Pre-postposition phrase. Use of PP is extended to incorporate postpositions too in addition to postpositions as well.
WHPP	WH Pre-postpositional. WHPP here annotates WH phrases where prepositions or postpositions could also be covered.

3.6 Functional Tagset

In order to add semantic meanings, some of the functional tags are used. These annotations help to increase specifications of the functional and grammatical roles. They are appended with – sign after the treebank phrase/clause level tags.

Table3: Function tags used in our work

Func-tional Tag	Description
ACC	Marks that the case phrase is Accaustive. Used

	when the direct object is followed by case marker "ko".
ADV	When any constituent other than the adverbs or pre-postposition are used as adverbs it marks it adverbial
BNF	Marks the beneficiary of the action performed
DAT	Annotated that the case phrase is Dative. Used when the indirect object or the subject is followed by case marker "ko".
DIR	Conveys the directional information, e.g. ki taraf (towards)
ERG	Marks that the case phrase is ergative. Used when the Subject noun phrase is followed by "ne"
EXT	Annotates that the adverb represents extent information e.g. us say ziada (more than that)
GEN	Labels the genitive phrase comprising of noun phrase followed by genitive case marker (ka, kai, ki) and a noun prase.
INST	Instruments used by the subject, oblique or adjunct. It is used when the noun phrase is followed by the case marker "se"
LOC	Represent that the adverbs represent a location information
MNR	Specifies that adverb represents manner information e.g. ahesta sai (slowly)
NOM	Specifies that the case phrase is a nominative one as no case marker follows the noun phrase
PRD	Marks the non VP predicate. Will be used in future for predicate argument representation
PRP	Specifies that the clause represent is the explanation or entails purpose of the preceding clause/sentence
SBJ	Used to mark a nominal noun as a subject
TMP	Identifies that adverbs represent time based information e.g. jab, tab, kitni dair
TPC	When the fronted elements is dislocated and appear before the NP-SBJ, then it is marked as a topicalized element
VOC	pecifies that the noun is used as in vocation form, e.g. when a person is called.

4 DATA AND TAGGING

We have selected the corpus of 500 ahadith translated in Urdu that has been already POS tagged. The data was initially tagged with the phrase tags of our designed tagset. It was then romanized for CFG extraction.

Following are the examples of the phrase tag application to yield the Urdu treebank.

Grammar is extracted in CFG format using our own grammar extractor from treebank. In order to provide computational support, the tagged corpus was first Romanized. Later the CFG rules were extracted.

Using the CFG rules few unseen sentences were automatically annotated, that yield multiple parse trees for each sentence.

Sample annotated text is represented:

```
(S
 (SBAR پھر<SC>
  (S
   (KP-DAT (NP-SBJ اس <PR><OBL>)کو<CM> )
   (PP-TMP (NP اس<DM>)(PP کے<CM>  بعد <NNCM> ))
   (VP  (NP کنکریاں<NN><F><P>)
        (VP پھینکتے<VB> ہوئے <VBL>دیکھا<VB>)))) .<SM> )
```

Where the angular brackets represent POS tags, and parenthesis represent treebank tagging using phrase structure annotations.

Sample CFG is of the form:

S->VP | SM
SM->-
KP-ERG->NP-SBJ | CM
CM->کو | نے
VB->اٹھا
VP->KP-ERG | KP-ACC | VB | AUXA | VP

While testing with the unseen data, the parse trees generated showed 80% recall.

5 CONCLUSION AND FUTURE WORK

Our proposed work has generated a development output of a fundamental resource for Urdu Language comprising of over 5000 words, and associated set of CFG rules. In addition, we have contributed with phrase level rules for Urdu text as well as strategies of rule applications for Urdu language.

We hope to extend this work by increasing the Treebank size to get the more comprehensive resource. Use of predicate argument structure approach to get the semantics of the phrases, clauses as well as sentences is a valuable upgradation of our work [10].

REFERENCES

[1] Sampson, G.(2003), *Thoughts on Two Decades of Drawing Trees'*, 23-41.

[2] Asif, T., (2012), *Developing a POS tagged Resource in Urdu*, Punjab University College of Information Technology, University of the Punjab.

[3] Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., ... & Schasberger, B. (1994, March). The Penn Treebank: annotating predicate argument structure. In Proceedings of the workshop on Human Language Technology (pp. 114-119). Association for Computational Linguistics.

[4] Abeillé, A (ed.)(2003a), T*reebanks: Building and using Parsed Corpora*, Dordrecht: Kluwer.

[5] Atwell, E.S., (2007), HSK-Corpus Linguistics, MILES Release 18.02

[6] Butt, M. (2006), *Theories of case*, Cambridge University Press

[7] Garside, R. et al (1992), *Lancaster Parsed Corpus* by ICAME, Bergen: The Norwegian Computing center for Humanities.

[8] Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1995). Treebank 2. Linguistic Data Consortium, Philadelphia.

[9] Kingbury, P., Palmer, M., (2003), *PropBank, the Next Level of Treebank*, Nivre/Hinrichs.

[10] Hindle (1983), *User Mannual for fidditch, a deterministric parser*.

[11] Ali, A. (2010), *Study of Morphology of Urdu Language, for its Computational Modeling.* Pub: VDM.

[12] Ali, A. (2011), *Syntax of Urdu Language (A survey of Urdu Language syntax).* LAP, Lambert Academic Publishing.

[13] Ali A., Hussain S., et al (2007), *Study of Noun Phrase in Urdu*,

[14] Haq, M.A. (1987), اردو صرف و نحو, Anjuman-e-Taraqqi Urdu(Hind)

[15] Siddiqui, A. (1971), جامع القواعد , Markazi Urdu Board

[16] Javed, I. (1981),نیئ اردو قواعد , Taraqqi Urdu Bureau, New Delhi.

[17] Platts, J. T. (1909), *A Grammar of the Hindustani or Urdu Language"*, London.

[18] Schmidt, R L. (1999), *Urdu: An Essential Grammar"*, Routledge, London and New York.

[19]Abbas, Q., (2014), *Building Computational Resources: The URDU.KON-TB Treebank and the Urdu Parser"*, *Konstanzer Online-Publikations-System (KOPS).*

[20] Abbas, Q. et. al. (2009), *Development of Tree-bank Based Probabilistic Grammar for Urdu Language*, International Journal of Electrical & Computer Sciences IJECS-IJENS Vol: 09 No:09

[21] Mel'cuk, I., (1988), *Dependency Structure: Theory and Practice*, State University of New York Press.

[22] Bhatt R., et al., (2009), *A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu, In the Proceedings of the Third Linguistic Annotation Workshop, held in conjunction with ACL-IJCNLP*, Singapore.

[23] Raza, G., (2010), *Analyzing the Structure of Urdu NPs with Multiple Genitives*, Proceedings of the Conference on Language & Technology 2010, Islamabad.

[24] Barati, A.,(2012), *Ancorra: Treebanks for Indian Languages*.