Sindh Univ. Res. Jour. (Sci. Ser.) Vol.49 (004) 733-738 (2017)

http://doi.org/10.26692/sujo/2017.12.0049

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Developing a Computational Syntax of Sindhi Language in Lexical Functional Grammar Framework

M.U. RAHMAN[++], H.U. KAZI

Department of Computer Science, Isra University, Hyderabad Sindh 71000 Pakistan

Received 12th April 2017 and Revised 10th November 2017

Abstract: Sindhi language lacks computational linguistics resources for deep syntactic analysis. This paper presents a work on computational morphology and grammar development of Sindhi Language. An LFG (Lexical Functional Grammar) based model for Sindhi grammar is developed where morphological constructions are modeled in Xerox Lexicon Compiler (LEXC), and syntactic constructions are modeled in LFG by using Xerox Linguistic Environment (XLE). While developing morphology and syntax of Sindhi, different part of speech classes, phrase structures, tense, aspect, mood and agreement are considered wherever applicable. The developed computational grammar is tested against two different test suites. First test suite contains 617 handcrafted sentences in 10 different test files containing sentences with different syntactic features. Second test suite contains real time corpus of two text books of Sindhi class one with 258 sentences. Results show 98.05% and 96.5% parsing percentage of test suite 1 and test suite 2 respectively.

Keywords: Syntax, Computational Morphology, Sindhi LFG.

1. INTRODUCTION

Computational grammar development and deep linguistic analysis provide structural details for natural language understanding by machines. Modern multilingual information processing systems use these details to understand and process information in different languages. Sindhi lacks resources like computational grammars and deep linguistic analysis systems. Development of such resources for Sindhi is open research area in computational linguistics and natural language processing domains.

This research proposes a computational grammar of Sindhi developed and evaluated in lexical functional grammar (LFG) (Dalrymple, 2001) framework. Various grammatical constructs of Sindhi language are analyzed and implemented. Morphological analysis as required by syntax modeling is implemented in finite state morphology (FSM) and integrated with LFG. Various morphological constructions of Sindhi including number, gender, case, tense, aspect and mood are considered during implementation. Xerox Linguistic Environment (XLE) (Dick, *et. al.*, 2008) is used to implement Sindhi LFG. Xerox Finite State Technology (XFST) tools (Kenneth and Lauri, 2002) are used to implement FSM of Sindhi which is then integrated with LFG within XLE environment. Roman transliteration is used in this study on ParGram guidelines (Kamran, *et al.,* 2010). A transliteration system is separately developed and used to convert Sindhi sentences in roman script. Capital letters in transliteration scheme represent long vowels of Sindhi, for example “A”(آ), “O” (او), “I” (اِي), and “U” (أُو). Small letters are used for consonants and short vowels.

1.1. Finite State Morphology

Two level finite state morphology (Roche and Shabes, 1997) plays essential role in implementation of morphological analyzers for natural languages. Fig. 1. shows the process of two level morphology modeling using FSTs. **(Fig.1. (a)** shows the finite state transducer where either upper or lower layer is used as input and the other one as output. A sample orthography FST rule can be “y→ie / ^____s#” which says that “y” will be replaced with “ie” whenever it is between morpheme boundary “^” and ending “s” (“^” and “#” represent morpheme boundary and word boundary respectively). This rule simply converts intermediate plural forms with “-ys” ending into “-ies” as shown Fig.1. Overall conversion process can be seen in **(Fig.1. (b). Fig.1. (c))** shows the block diagram of this process.

1.2. Lexical Functional Grammar

Lexical Functional Grammar (LFG) is a natural language syntax representation formalism based on generative grammars. LFG defines the structure of language and relationship among different aspects of linguistic structure. Various relations are defined at lexicon level as LFG has a rich lexical structure. LFG represents linguistic structure at different levels which include lexicon, constituency structure (c-structure) and functional structure (f-structure) levels. A lexical entry in LFG may include part of speech, number, gender, case, and argument structure in case of verbs and some postpositions and adjectives.

[++]Corresponding author: Email: muteerurahman@gmail.com

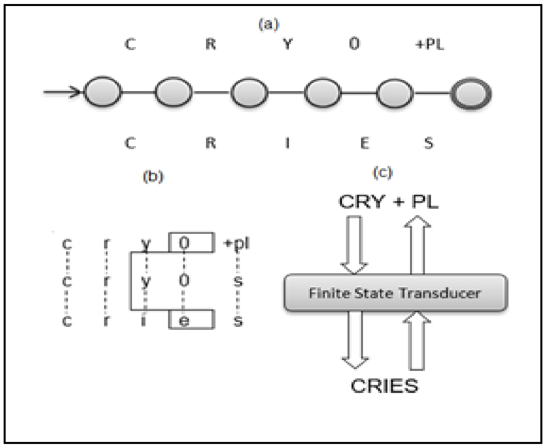
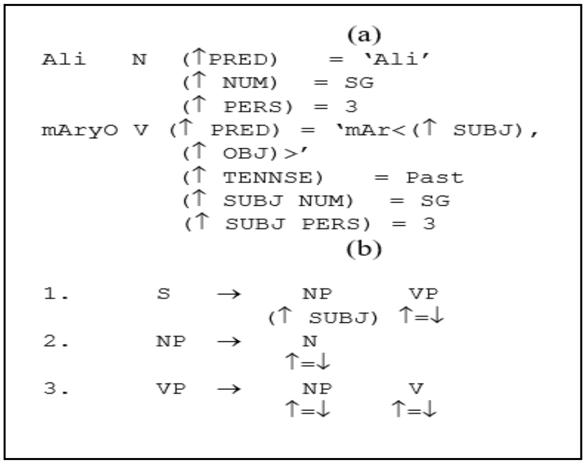


Fig. 1. Two Level Morphology Process.



```
                         (a)
Ali    N    (↑PRED)    = 'Ali'
            (↑ NUM)    = SG
            (↑ PERS)   = 3
mAryO V  (↑ PRED)  =  'mAr<(↑ SUBJ),
            (↑ OBJ)>'
            (↑ TENNSE)    = Past
            (↑ SUBJ NUM)   = SG
            (↑ SUBJ PERS)  = 3
                         (b)

1.     S     →     NP        VP
                (↑ SUBJ)   ↑=↓
2.     NP    →     N
                  ↑=↓
3.     VP    →     NP        V
                  ↑=↓       ↑=↓
```

Fig. 2. Lexicon and C-Structure Rules.

Sample proper noun and verb entries are shown in **(Fig.2. a).** C-structure representation is first level of syntax in LFG and handles word or phrase grouping and their precedence in a phrase structure tree along-with some grouping and order constraints (C-structure rules can be seen in **(Fig.2. b).** F-structure is another level which represents more abstract relations between different functional constructs like subject, object, secondary object. A parse tree and f-structure generated by Fig. 2. rules and lexicon entries is shown in **(Fig. 3)**.

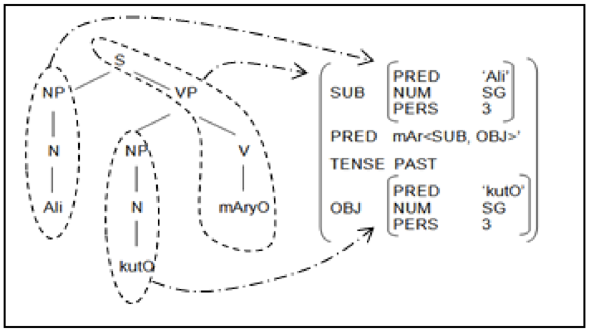


Fig. 3. Parse Tree and F-Structure of a Sample Sentence

Parsing of a sample sentence (Ali killed the dog) is shown with syntax analysis with subject and object grammatical functions. Subsequent sections discuss related work, implementation details, coverage of developed grammar, results and conclusions.

2. RELATED WORK

To the best of our knowledge literature about Sindhi syntax analysis in modern linguistic frameworks like LFG is not available; however, studies in Context Free Grammars and Linear Specification Language can be found in (Rahman and Shah, 2003) and (Rahman, *et. al*, 2007). First study has over generation problems and second study lacks the agreement problem solution and feature representations. Another study is Grammatical Framework Resource Grammar for Sindhi (Oad, 2012). This study includes syntax coverage along-with morphology where a preliminary framework for morphology and syntax of Sindhi is presented; however complex morpho-syntactic features of Sindhi are still subject to research. Few computational linguistics resources are also available which include an online dictionary (CLE, 2016), and a POS tagset (Mahar and Memon, 2010a). Some preliminary NLP research studies for Sindhi are also in place which include part of speech tagging (Mahar and Memon, 2011), (Mahar *et al.*, 2011), and text to speech modeling (Mahar *et al.*, 2010). Recently various online dictionaries are made available by Sindhi Language Authority (SLA, 2016). Among south Asian languages Urdu is extensively studied with LFG perspective. Urdu became part of parallel grammar project (ParGram) (Butt and King, 2002) and was analyzed with large scale grammar development perspective. Jafar Rizvi in his PhD thesis (Rizvi, 2007) also presented Urdu syntax analysis in LFG.

3. IMPLEMENTATION

Overall implementation model is shown in **(Fig. 4).** Based on identified morphology and syntax patterns Sindhi grammar is analyzed and studied with LFG perspective. Sindhi morphological constructions are implemented in finite state morphology. XFST Lexicon Compiler and XLE are used to develop Sindhi morphology and Syntax respectively. Different components are interfaced with each other in XLE to parse and analyze Sindhi sentences. LFG grammar is integrated with developed FSM and sentences are transliterated into roman script. Developed LFG grammar which generates parse trees and functional structures with deep syntactic analysis for these sentences.

3.1 Implementing Morphology

Different morphological paradigms of nouns, pronouns, adjectives, adverbs and verbs are represented in finite state transducers in LEXC (Lexicon Compiler)

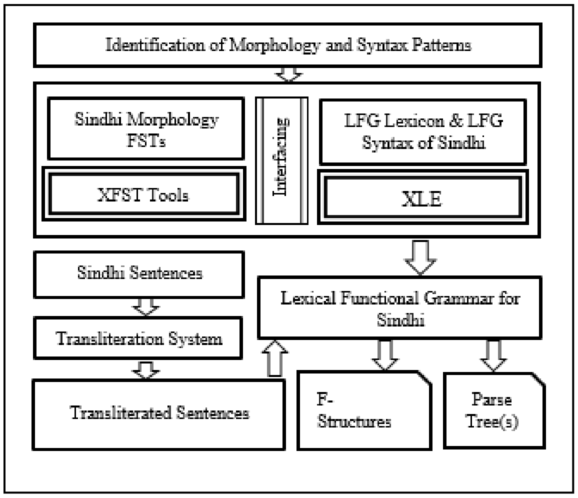


Fig. 4. Grammar Development Model.

(Karttunen, 1993) and are compiled to generate finite state machines which represent Sindhi lexicon. These state machines act as function machines where either upper side represents the input and lower side represents the output or vice versa. Due to this reversible property when lower side becomes input these FSTs function as morphological analyzers and when upper side is input these will function as surface form generators. An example entry of noun in LEXC script is given below:

```
CHOkir+Noun+Common+Count+Animate
```

This will produce intermediate animate common count noun form "CHOkir", this transducer is followed by another transducer in series (via sub-lexicon link) which takes further input tags as shown below:

```
+Sg+Masc+Nominative
```

This tag sequence produces the singular masculine nominative morpheme "O". The overall concatenated tag sequence preceded by stem (upper side) and concatenated output (lower side) are given below:

Upper: `CHOkir+Noun+Common+Count+`
`Animate+Sg+Masc+Nominative`
Intermediate: `CHOkir           O`
Lower: `CHOkirO`

While going from upper to lower side, surface form "CHOkirO" of stem "CHOkir" with features specified in tag sequence is generated; going from lower to upper will give following morphological analysis of noun "CHOkirO".

```
CHOkir {"+Noun" "+Common" "+Count"
"+Animate" "+Sg" "+Masc" "+Nominative"}
```

Above morphological analysis says that "CHOkirO" is a morphological form of stem "CHOkir" which is a common animate count noun in singular masculine form with nominative case. In the same way oblique morphological form (used as base form for various syntactic cases of nouns) "CHOkirE" is generated by producing and concatenating the oblique morpheme "E" by input tag sequence given below and output sequence "CHOkir" and "E". Total twelve (12) different inflections of stem "CHOkir" are taken care of. A total of 21 different common noun categories are identified according to their inflectional properties. For every category, a different sub-lexicon is defined. Usually proper nouns are not inflected therefore their entries only contain the feature tags. However, in Sindhi there are exceptional cases of proper noun inflections. For example, a person name "dOdO" can have number, and case inflections "dOdA" (plural or singular vocative) and "dOdE" (oblique form). A sub-lexicon is defined to handle these inflections. Verb in Sindhi is a morphologically complex word class. Verbs are marked by number, gender, case, tense, aspect and mood. Various categories of auxiliary verbs are also inflected by number, gender, and case; auxiliaries may also be used as tense and aspect markers with inflections. Copula verbs also undergo morphological changes. Verb lexicon covers auxiliary verb, copula verb and main verb morphology. Analyses show that a verb in Sindhi can have up to 75 different morphological forms. Pronoun, Adjective, and adverb morphology is also modeled on same lines like noun and verb morphology.

3.2. Implementing Syntax

Different syntactic constructions of Sindhi are implemented in XLE by defining Sindhi LFG rules. Morphology defined in LEXC scripts is compiled to finite state transducers (discussed above) and integrated to LFG grammar via morphology syntax interface in XLE environment.

Nominal Elements: Nominal elements include nouns, pronouns, adjectives, adverbs and phrases constituted by these elements. Different NP constructions implemented include: pronoun-noun, adjective-noun, and pronoun-adjective-noun combinations. These noun phrase combinations are further complicated by coordination, postpositional phrases and relative clauses. Different cases of nominal elements including nominative, accusative, dative, ablative, locative, instrumental, participant, genitive/possessive, agentive and vocative are taken care of. Different complications of syntactic case marking are handled by defining a special case phrase KP (Bögel, *et. al.,* 2009) which represents case marked noun phrase constructions. For genitive case, separate phrase KPPoss (possessive case phrase) is defined which handles special agreement features required for agreement by different constituents of a sentence. LFG definition of KPPoss in XLE format is given below:

```
   KPPoss -->    NP: {(! N-FORM)=c obl |
        (! NTYPE NSYN)= proper} ^=!;
                 KPoss: ^=!.
```

LFG lexicon entry of KPoss (possessive case marker) "jO" showing extra attributes is as follows.

```
jO     KPoss * (^ PP-FORM)=of
                (^ K-NUM)=sg
                (^ K-GEND)=masc
                (^ K-FORM)=nom
                (^ CASE)=gen.
```

Extra attributes K-NUM, K-GEND, and K-FORM (K represents case) are introduced here to reflect the possessive case marker attributes to be agreed with possessed noun attributes.

Verbal Elements: Verbal elements include verbs which subcategorize (require arguments) for different grammatical functions. These grammatical functions include subject (SUBJ), object (OBJ), secondary object (OBJ2), oblique (OBL), PREDLINK, complement (COMP) and open complement XCOMP. Noun phrases (including all nominal elements) either define these functions or play essential role in their definition within a sentence. Sentence constituents therefore include verbs, their arguments and adjunct (ADJUNCT) elements which do not subcategorize for verbs. Different Verb categories include predicative verbs (main verbs and copula verbs), modal verbs and auxiliary verbs. Main, auxiliary and modal verbs are combined to make verbal complex. Auxiliaries are also used to mark tense, aspect and mood. Implementation includes verbal subcategorization for different grammatical functions listed above, verbal complex, and tense-aspect-mood analysis. Tense coverage include aorist formations, present, past and future tenses. Aspectual formations including perfective, imperfective-habitual and imperfective-continuous are analyzed by implemented LFG rules. Verb mood is also analyzed, coverage of different mood constructions includes: subjunctive, presumptive, imperative, declarative or indicative, permissive, prohibitive, capacitive, suggestive, and compulsive moods. A short version of sentence definition in LFG format is given below:

```
  S-->   NP:(^SUBJ)=! (! GEND)=(^ GEND);)
      (KP: (^ OBJ2)=! (! CASE)=c dat)
    (KP: (^ OBL)=! {(! CASE)=c inst | (!
               CASE)=c agent})
     (KP: (^ OBJ)=! {(! CASE)=c acc | (!
               CASE)=c nom})
VC: (! NUM)=(^NUM) (! GEND)=(^ GEND) ^=!.
```

Above rules define sentence S as a sequence of noun phrase (NP) which is a subject, followed by optional case phrases (KPs) which include indirect object (OBJ2), oblique (OBL) and direct object (OBJ) followed by verb complex which may include combinations of different verb types. Above given rule defines the general structure of Sindhi sentence. Different constraints like (! GEND) = (^ GEND) and (! CASE=c dat) are placed to ensure gender case and number agreement. Consider following sentence:

Ali	**CHOkirE-khE**	**KHatu**
Ali. Nom.M	**boy. Obl.Sg.M-Dat**	**letter. Nom.M.Sg**

likhE	**payO**
write. Aorist. Sg	**Aux.Cont**

Ali is writing a letter to the boy.

In above sentence there are three verbal arguments, a subject "Ali", an indirect object "CHOkirO" in oblique form and a direct object "KHatu" in nominative case. **(Fig. 5)** shows parse tree of the sentence and F-structure with syntactic details is shown in **(Fig. 6).**

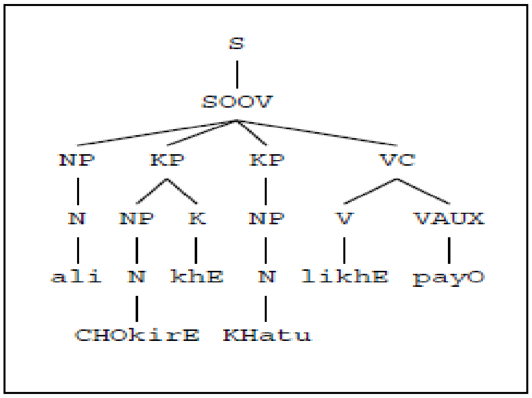
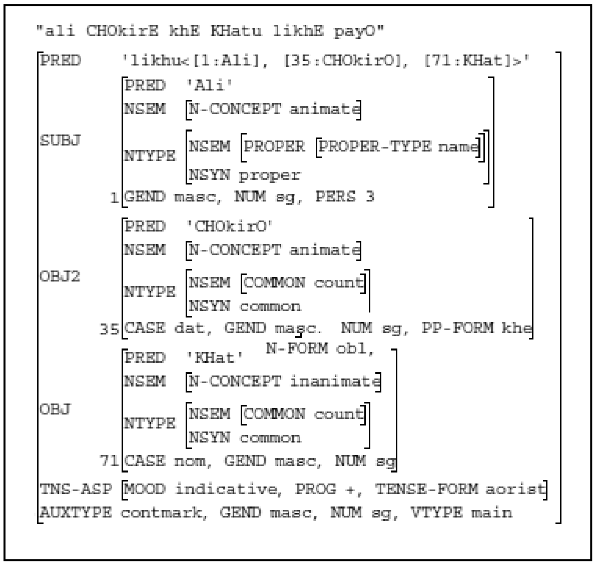


Fig. 5. Sample sentence with imperfective continuous aspect



```
"ali CHOkirE khE KHatu likhE payO"
[PRED    'likhu<[1:Ali], [35:CHOkirO], [71:KHat]>'                  ]
[        [PRED  'Ali'                                    ]          ]
[        [NSEM  [N-CONCEPT animate]                      ]          ]
[SUBJ    [NTYPE [NSEM [PROPER [PROPER-TYPE name]]]        ]          ]
[        [      [NSYN proper                    ]        ]          ]
[       1[GEND masc, NUM sg, PERS 3                      ]          ]
[        [PRED  'CHOkirO'                                      ]    ]
[        [NSEM  [N-CONCEPT animate]                            ]    ]
[OBJ2    [NTYPE [NSEM [COMMON count]]                          ]    ]
[        [      [NSYN common        ]                          ]    ]
[      35[CASE dat, GEND masc.  NUM sg, PP-FORM khe            ]    ]
[        [PRED  'KHat'  N-FORM obl,  ]                              ]
[        [NSEM  [N-CONCEPT inanimate]]                              ]
[OBJ     [NTYPE [NSEM [COMMON count]]]                              ]
[        [      [NSYN common        ]]                              ]
[      71[CASE nom, GEND masc, NUM sg]                              ]
[TNS-ASP [MOOD indicative, PROG +, TENSE-FORM aorist]              ]
[AUXTYPE contmark, GEND masc, NUM sg, VTYPE main                    ]
```

Fig. 6. LFG Analysis of a sentence with SUBJ, OBJ and OBJ2 sub categorization in aorist tense form with imperfective continuous aspect.

Sindhi pronominal suffixes may appear with nouns, verbs, postpositions, and adverbs of place. Pronominal suffixes are treated as special lexical entries in lexicon. For example, consider transitive verb "likhu" (write); when appears with 1st person pronominal suffix "-iyami" becomes "likh-iyami" (I wrote). Morphological analysis of "likhiyami" is given below:

```
{likhiyami "+Token" | likhu "+Verb"
   "+Psx"  "+SSg"  "+S1P"  "+SMF"
    "+SObl"  "+Sg"  "+PastPart"}
```

Here, "+Psx" attribute says that this is a pronominal suffixed form. The tag pattern "+Sxxx" represent different attributes of subject reflected by pronominal suffix. "+PastPart" tag says that verb form is past participle.

4. COVERAGE

Morphological coverage includes: finite state models of nouns, pronouns, adjectives, adverbs and verbs, postpositions, conjunctions and adverbs. Case, mood, tense and aspect morphology of nominal and verbal elements is also implemented. **(Table 1)** shows some figures about morphology coverage. Interestingly adjectives have more average inflections per stem as compared to nouns. This is due to degree change inflections of native Sindhi adjectives where inflections are doubled as compared to nouns. Pronoun inflections per stem is also 3.58 due to number gender and case inflections (mostly in wh-pronons). Syntax coverage include noun phrase constructions with all nominal elements and verbal elements. Verb subcategorization with subject, object, oblique, secondary object, complement, open complement, adjunct, open adjunct, and predicate link (predlink), coordination, subordination, mood, case, aspect, tense, and agreement is also implemented. Coverage of LFG rules is shown in **(Table - 2**) Total 24 rules are implemented and are used to parse the sentences in test suites. Most of rules are completely used along-with their sub-rules / choices. However, few rules are partially used as their sub-rules or choices are not used completely.

5. RESULTS

The developed grammar is evaluated against two different test suites. Test suite 1 contains 10 different test files with a total of 617 sentences covering various linguistic features. These sentences were given as input to the developed grammar. Thus, total 605 sentences were parsed successfully with deep linguistic analysis. A bar chart showing results of test suite 1 is given in **(Fig. 7-8).** In two test files of Test suite 2 total 258 sentences selected from Sindhi class one books were there and 249 were successfully parsed. Results show 98.05% and 96.5% parsing percentage of test suite 1 and test suite 2 respectively.

Table 1. Morphology Coverage

Word Class	Stems	Inflections	Average Inflections / Stem
Verbs	100	5013	50.13
Nouns	323	1729	5.35
Pronouns	79	283	3.58
Adjectives	71	394	5.55
Adverbs	38	38	1.00
Total	**611**	**7457**	**12.20**

Table 2. Grammar Coverage

Total Number of LFG Rules	24
Coverage by test Corpus	24
Partially Un-Used Rules	6
Unused choices in 6 Partially Unused rules	87

Parsing results of individual files of test suite 2 are shown in bar chart of Fig. 8. Sentences not parsed in test Suite 1 and 2 were either bad sentences (ungrammatical) or having unhandled phenomenon.

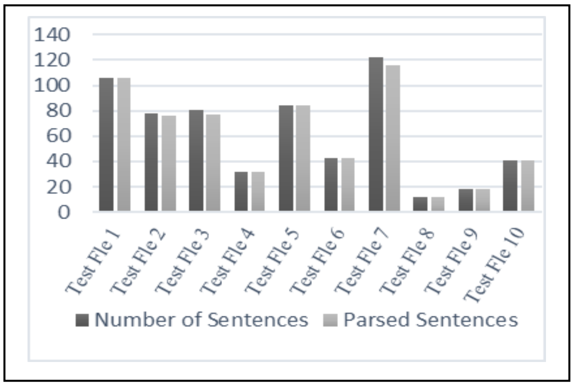


Fig.7. Parsing Results of Individual Files of Test Suite 1

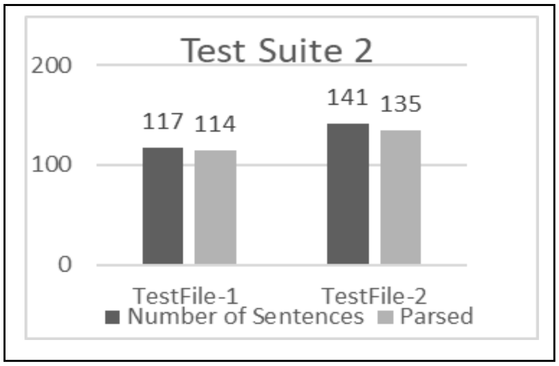


Fig. 8. Parsing Results of Individual Files of Test Suite 2

6. CONCLUSIONS

Developed grammar covers the morphological and syntactic constructions discussed above. Morphological analysis shows interesting results like adjectives have more average inflections than nouns, and pronouns have 3.58 average inflections per word. Also, verb can have up to 75 different morphological forms. Results of deep linguistic analysis of Sindhi sentences in LFG will provide basis for Sindhi language understanding by machines. These results are based on linguistic knowledge and generated results capture this knowledge at different levels including morphology, syntax and semantics. These linguistically rich structures can be given input to machine learning algorithms and this synthesis of deep linguistic analysis and machine learning can be used for more accurate feature extractions. Predicate argument structures generated by LFG can be used to extract semantic triples which are fundamental building blocks of knowledge representation in machine readable format. Use of semantic triples generated by predicate argument structures has applications in semantic web, knowledge extraction and information processing. Future research on developed grammar may also include work on incorporating optimality theory, and rewriting the grammar on ParGram guidelines.

REFERENCES:

Bögel, T., M. Butt, A. Hautli, and S. Sulger, (2009) Urdu and the modular architecture of ParGram, In proceedings of Conference on Language and Technology Lahore Pakistan.

Butt, M. and T. H. King, (2002). Urdu and the Parallel Grammar project. In Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12 (1-3). Association for Computational Linguistics.

CLE. (2016), Sindhi English Dictionary. http://www.clepk.org/sed1/. (Accessed December, 2016).

Dalrymple, M. (2001). Lexical-Functional Grammar. John Wiley and Sons, Ltd

Dick, C., M. Dalrymple, R. Kaplan, T. H. King, J. Maxwell, and P. Newman, (2008). XLE documentation. Palo Alto Research Center.

Kamran M., M. Ahmed, T. Sulger, S. Bögel, T. Gulzar, A. Raza, and M. Butt, (2010). Transliterating Urdu for a Broad-Coverage Urdu/Hindi LFG Grammar. In LREC 2010, Seventh International Conference on Language Resources and Evaluation (pp. 2921-2927). Malta.

Karttunen, L. (1993). Finite-state lexicon compiler. Xerox Corporation. Palo Alto Research Center.

Kenneth R., B. and L. Karttunen. (2002), Finite-State Morphology: Xerox Tools and Techniques. Studies in Natural Language Processing. Cambridge University Press.

Mahar J. A. and G. Q. Memon (2011), Probabilistic Analysis of Sindhi Word Prediction using N-Grams Australian Journal of Basic and Applied Sciences, 5(5): 1137-1143

Mahar J. A. and G. Q. Memon (2010a). Rule Based Part of Speech Tagging of Sindhi Language, icsap, International Conference on Signal Acquisition and Processing,.101-106.

Mahar, J. A., G. Q. Memon, and S. H. A. Shah, (2010), “WordNet Based Sindhi Text to Speech Synthesis System”, Proceedings of the 2010 Second International Conference on Computer Research and Development, .20-24, 07-10.

Mahar, J. A., H. Shaikh, A. R. Solangi. (2011). Comparative Analysis of Rule Based, Syntactic and Semantic Sindhi Parts of Speech Tagging Systems. International Journal of Academic Research. Vol. 3. No. 5. 44-49.

Oad, J. D. (2012). Implementing GF Resource Grammar for Sindhi. Master’s Thesis. Department of Applied Information Technology Chalmers University of Technology Gothenburg, Sweden.

Rahman, M. U., A. Shah, (2003). Grammar Checking Model for Local Languages., In Proceedings of SCONEST (Student Conference on Engineering Sciences and Technology). Hamdard Uni., Karachi.

Rahman, M. U., A. Shah, R. A. Memon, (2007) Partial Word Order Syntax of Urdu/Sindhi and Linear Specification Language. Journal of Independent Studies and Research (JISR) Vol. 5, 2, 13–18.

Rizvi, S. M. J. (2007). Development of algorithms and computational grammar for Urdu (PhD thesis, Pakistan Institute of Engineering and Applied Science, Islamabad, Pakistan. http://prr. hec. gov. pk/chapters/ 2072-0. pdf

Roche, E. and Yves Shabes. (1997), Finite-State Language Processing, MIT Press, Cambridge, MA.

SLA. (2016) Sindhi Language Authority. Official Website. http://www.sindhila.org. (Accessed December 2016).