Partial Word Order Syntax of Urdu/Sindhi and Linear Specification Language

Mutee-u-Rahman[1], Asadullah Shah[2] and Dr. Riaz A. Memon[3]
[1]Hamdard Institute of Information Technology Hamdard University, Karachi
[2]Isra University Hala Road Hyderabad Sindh 71000, Pakistan
[3] IMCS University of Sindh Jamshoro

Abstract: *Like most of the South-Asian languages Urdu and Sindhi are partial word order languages. Conventional syntax representation models like Context Free Grammars are not capable enough to cope with partial word order syntax. Linear Specification Language (LSL) is an extension of Context-Free Grammars (CFGs) which allows arbitrary partial order (free word order) on the right hand side of grammar rule. Partial word order in LSL is handled by using different types of linear precedence (LP) constraints. LSL by using LP constraints is capable enough to represent the syntax of partial word order sentence. Issues related to represent Urdu/Sindhi language sentences with their constituent parts in LSL are discussed. LSL versions for different types of Urdu and Sindhi sentences are presented.*

Keywords: *Natural Language Processing, Context Free Grammars, Linear Specification Language, Urdu/Sindhi Language Processing.*

1. INTRODUCTION

Urdu and Sindhi language sentences have complex syntax structure with partial word order. A partial word order sentence may have many proper orders of its constituent parts. By imposing few restrictions on the word order of local languages one can represent controlled word order sentences in conventional syntax representation models like Context Free Grammars (CFGs)[1][2][3]. CFGs are suitable only for controlled or fixed word order sentences and partial word order is not efficiently modeled by them [4].

Linear Specification Language (LSL)[5] is an extension of CFG with some changes on the right hand side of the grammar rule. These changes are in the shape of Linear Precedence (LP) constraints which are part of the grammar rule. The LP constraints define some precedence rules which specify the constituent ordering of the sentence. Local language sentences can have completely or partially free word orders [4]. For example the Urdu sentence **"زاهد بيمارہے"** can have different word orders like **"بيمارہے زاهد"** and **"بيمار زاهد ہے"**. In the same way Sindhi sentence **"علي خط لکيو"** can have the forms "**خط علي لکيو"** and **"علي خط لکيو"**. In subsequent sections formal definition of LSL is discussed in detail. LSL syntax representation of Urdu and Sindhi sentences is also presented with examples.

1.1. Formal Definition of LSL Grammar

A Linear Specification Language Grammar **G** = (V, T, S, P, L) is defined by [6] as given below:

V: Set of variables (Non-terminals)
T: Set of terminal symbols
S: Start symbol (***S*** ∈ ***V***)
L: Set of lexical entries (A lexical entry is a pair $\boldsymbol{Y} \rightarrow a$, where ***Y*** ∈ ***V*** and ***a*** ∈ ***T***)
P: Set of rules (productions)
Every LSL rule consists of following two parts:

i. A two-place relation between a variable and a set of variables
ii. Some linear precedence constraints (known as LP constraints) between variables

For instance

$$\left.\begin{array}{l} S \rightarrow \ V\ X\ Y\ Z; \text{ ---------- (i)} \\ V < X,\ X \ll Y,\ \langle\ V\ \rangle \text{ ---------- (ii)} \end{array}\right\} \text{------ 1.1}$$

Where V, X, Y and Z are variables, (i) is a two-place relationship between variables and (ii) shows LP constraints between variables.

Three different types of LP constraints are given below:

- < : Weak precedence written as V < X states that production of variable V is completely to the left of the terminal production of variable X.
- « : Immediate precedence written as X « Y states that the rightmost terminal derived from X stands immediately to the left of the leftmost terminal derived from Y.
- ⟨ ⟩: Isolation written as ⟨V⟩ states that terminal production of V is continuous. Sometimes LHS of the grammar rule can also be isolated which means that complete derivation of the rule should be continuous.

A rule without LP constraint (denoted by **ε**) states that any ordering of V, X, Y and Z can be there (free constituent ordering).

Now suppose that the terminal yield of V is *this is*, that of X is *a*, that of Y is *black* and that of Z is *book*.
For a sentence to be grammatical according to rule 1.1, it must hold that:

1. The terminal yield of V (*this is*) must occur to the left of the terminal yield of X (*a*).
2. The terminal yield of Y (*black*) must occur immediately to the left of the terminal yield of Z (*book*).
3. Yield of V (*this is*) must be continuous.

According to above three conditions which are specified in LP constraints of grammar 1.1 following sentences are grammatically correct.

- *this is a black book*
- *book this is a black*
- *this is book a black*

All the above sentences are grammatically correct according to the LSL rule given in 1.1 (but not according to English grammar).
Now consider another sentence:

- *a book this is black*

Above sentence is invalid because it violates condition (1) which says that V < X (*this is* should be before *a*).

The example shows that LSL formalism has ability to control partially free word orders in sentences which is the key idea behind the use of LSL grammar for Urdu and Sindhi sentences with relaxed word order.

2. LSL FOR URDU SENTENCES

In the example discussed below the grammar of *Ismia*[7] Urdu sentence with relaxed word order is considered. Grammar is an extension of Urdu CFG presented in [2].

Consider an LSL grammar for *Ismia* Urdu sentence
G = (V, T, S, P, L)
Where
V = { جملہ متعلق مبتدا، اسم ، مسند الیہ ، مسند ، خبر، متعلق خبر، فاعل، فعل ناقص، مبتدا، اسمیہ، }
T = {، دیکھا ،ہے، گیا ہے،آج ، بندرگاہ ،اکبر، کراچی، سارے، بازار سے میرا }
S = { جملہ اسمیہ }

Productions rules (P) with LP constraints and Lexical entries (L) are given in Figure 1.

The LP constraints in rule 1 state that constituent parts of *Ismia* sentence can have any partial order provided that *Fail-e-Naqis* will always come after *Masand Alya*. Rules 2 and 3 are without LP constraints which mean that their constituent parts can have any order. Rule 4 and 5 describe that *Isam* is isolated and continuous (which may not be the case for many other sentence types).

Following sentences are derivable by using above LSL grammar rules and therefore are grammatically correct sentences.

- کراچی بندرگاہ ہے
- کراچی ہے بندرگاہ
- بندرگاہ کراچی ہے

The constituent ordering according to LP constraints is shown in Table 1, 2 and 3 respectively.

Table 1

LP Constraint	مسند الیہ< فعل ناقص		
Constituents	**فعل ناقص**	**مسند**	**مسند الیہ**
Sentence	ہے	بندرگاہ	کراچی

1. جملہ اسمیہ← مسند الیہ مسند فعل ناقص
   ؛ مسند الیہ< فعل ناقص
2. مسند الیہ ← مبتدا متعلق مبتدا(*Optional*) ؛ ε
3. مسند ← خبر متعلق خبر (*Optional*) ؛ ε
4. خبر ← اسم ; ⟨ اسم ⟩
5. مبتدا ← اسم ; ⟨ اسم ⟩
6. اسم ← کراچی
7. اسم ← اکبر
8. اسم ← بندرگاہ
9. اسم ← دوست
10. متعلق مبتدا ← میرا
11. متعلق خبر ← سارے
12. فعل ناقص ← ہے

Figure 1. LSL rules for *Ismia* Urdu sentence

Table 2

LP Constraint	مسند الیہ< فعل ناقص		
Constituents	**مسند**	**فعل ناقص**	**مسند الیہ**
Sentence	بندرگاہ	ہے	کراچی

Table 3

LP Constraint	مسند الیہ< فعل ناقص		
Constituents	**فعل ناقص**	**مسند الیہ**	**مسند**
Sentence	ہے	کراچی	بندرگاہ

All the LP constraints of LSL grammar of Figure 1 are satisfied in above sentences. In all above sentences the order of *Masand* is completely free while the order of *Masand Alya* and *Fail-e-Naqis* is partially free. Because of partial word order LP constraint the sentence **"کراچی ہے بندرگاہ"** is not grammatically correct according to above LSL grammar it violates the condition **مسند الیہ < فعل ناقص**. Derivation tree for all correct sentences is given in Figure 2. Because of partial word order LP constraints only one derivation tree will be formed for all possible orders.

Now consider another set of productions for *Failia*[8] Urdu sentence. The sets **V** and **T** are same as discussed above. Production rules P with LP constraints and lexical entries given in Figure 3.

In rule 1 of figure 3 *Masand Alya* is isolated and continuous in the definition of *Jumla-e-Failia* which may not be the case for its own definition. While *Masand* has no LP constraint which means that *Masand* can have free order of its derived (with their LP constraints) constituent parts in *Jumla-e-Falia*.

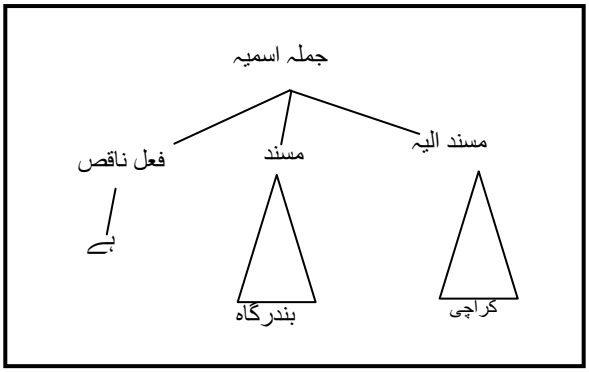

Figure 2. Derivation Tree for Sample *Ismia* (اسمیہ) Sentences (کراچی بندرگاہ ہے), (بندرگاہ کراچی ہے) and (بندرگاہ کراچی ہے).

1. جملہ فعلیہ← مسند الیہ مسند؛ ⟨ مسند الیہ ⟩
2. مسند الیہ← متعلق فاعل فاعل؛ ⟨ متعلق فاعل⟩ ⟨ فاعل⟩
3. مسند← متعلق فعل مفعول متعلق مفعول فعل؛
   مفعول« متعلق مفعول ⟨ فعل⟩ ⟨ متعلق فعل⟩
4. مفعول ← اسم؛ ⟨ اسم⟩
5. فاعل ← اسم ؛ ⟨ اسم⟩
6. متعلق فاعل← میرا
7. متعلق فعل ← آج
8. فعل ← گیا ہے
9. متعلق مفعول←کے گھر
10. اسم ← اکبر
11. اسم ← دوست

Figure 3. LSL rules for *Failia* Urdu sentence.

Rule 2 specifies that *Masand Alya* can have isolated continuous constituent parts with any order. In rule 3 *Mafool* has immediate precedence over *Mutaliq-e-Mafool* which means that *Mafool* should come before *Mutaliq-e-Mafool* and the rightmost terminal derived from *Mafool* stands immediately to the left of the leftmost terminal derived from *Mutaliq-e-Mafool*. All other parts of *Masand* are isolated and continuous with any order. Remaining rules include lexical entries and simple rules with isolation and continuation.

Following sentences are grammatically correct according to the LSL grammar given in Figure 3.

- میرا دوست آج اکبر کے گھر گیا ہے
- دوست میرا آج اکبر کے گھر گیا ہے
- آج میرا دوست اکبر کے گھر گیا ہے
- آج دوست میرا اکبر کے گھر گیا ہے
- اکبر کے گھر آج میرا دوست گیا ہے
- اکبر کے گھر میرا دوست آج گیا ہے
- اکبر کے گھر دوست میرا آج گیا ہے
- ا کبر کے گھر میرا دوست گیا ہے آج
- ا کبر کے گھر دوست میرا گیا ہےا آج
- میرا دوست آج گیا ہے اکبر کے گھر
- میرا دوست اکبر کے گھر آج گیا ہے
- آج ا کبر کے گھر دوست میرا گیا ہے
- آج ا کبر کے گھر میرا دوست گیا ہے
- آج ا کبر کے گھر گیا ہے دوست میرا
- آج ا کبر کے گھر گیا ہے میرا دوست
- آج گیا ہے دوست میرا اکبر کے گھر

All the above word orders (and many others) are represented by the LSL rules of Figure 3 and are therefore grammatically correct. The constituent ordering of two of the above sentences according to LP constraints is given below in Table 4 and 5.

Table 4

LP Constraints	مسند				⟨ مسند الیہ ⟩	
	⟨ فعل⟩ ⟨ متعلق فعل⟩ مفعول« متعلق مفعول				⟨ فاعل⟩ ⟨ متعلق فاعل⟩	
Constituents	فعل	متعلق مفعول	مفعول	متعلق فعل	متعلق فاعل	فاعل
Sentence	گیا ہے آ	کے گھر	اکبر	آج	میرا	دوست

Table 5

LP Constraints	مسند				⟨ مسند الیہ ⟩	
	مفعول« ⟨ فعل⟩ ⟨ متعلق فعل⟩ متعلق مفعول				⟨ فاعل⟩ ⟨ متعلق فاعل⟩	
Constituents	فعل	متعلق مفعول	مفعول	متعلق فاعل	فاعل	متعلق فعل
Sentence	گیا ہے آ	کے گھر	اکبر	میرا	دوست	آج

Again there will be only one derivation tree as shown in Figure 4 for all possible word orders given above, because order information is encapsulated within LP constraints of LSL grammar.

Following sentence is not grammatically correct according to LSL rules of Figure 3 because it is stated in rule 1 that *Masand Alya* must be continuous and it is violated in the sentence given below by discontinuation of **"میرا دوست"**.

- دوست آج گیا ہے اکبر کے گھر میرا

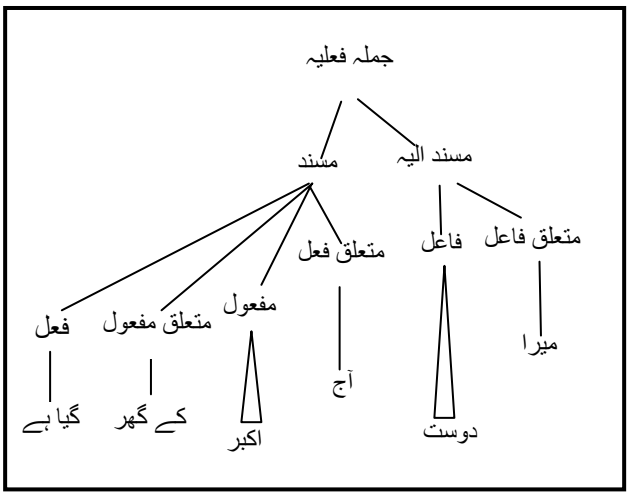

Figure 4. Derivation tree for all possible word orders of the sentence (میرا دوست آج اکبر کے گھر گیا ہے) according to LSL grammar of Figure 3.

It should be noted that writing CF grammar for all above word orders results in excessive number of rules that is for every word order there will be a different rule.

3. LSL FOR SINDHI SENTENCES

Now consider simple Sindhi sentence (علي لکي ٿو) and its LSL grammar given below:

G = (V, T, S, P, L)

Where

V = {جملو، مبتدا، خبر، فاعل، فاعل جو لڳ، مفعول، مفعول جو لڳ، فعل، فعل جو لڳ، اسم}

T = {علي، سليم، لکي ٿو، ويوآهي، دوست}

S = {جملو}

Production rules P with LP constraints and lexical entries L are given in Figure 5.

Rule 1 of Figure 5 specifies that constituent parts of sentence can have any order with independence and continuation of *Mubtada*[9][10]. In the same way LP constraints of rule 2, 3, 4 and 5 specify the partial or free word order of sentence constituents with necessary independence and continuation.

Consider the sentence (علي لکي ٿو). Due to partial word order definition of sentence (that is continuity of *Mubtada*) only two possible proper word orders of above sentence are:

- علي لکي ٿو
- لکي ٿو علي

Derivation tree for above sentences is given in Figure 6. Constituent ordering of the all proper word orders is given below in Table 6 and 7.

1. جملو← مبتدا خبر ؛ ‹ مبتدا ›
2. مبتدا← فاعل جو لڳ(*Optional*) فاعل ؛‹ فاعل جو لڳ› ‹ فاعل›
3. خبر← فعل جو لڳ(*Optional*) مفعول(*Optional*) مفعول جو لڳ(*Optional*) فعل ؛ مفعول جو لڳ » مفعول ‹ فعل › ‹ فعل جو لڳ›
4. فاعل← اسم ؛‹اسم›
5. مفعول← اسم؛ ‹اسم›
6. فعل← لکي ٿو
7. فعل← ويوآهي
8. اسم ← علي
9. اسم ← سليم
10. فعل جو لڳ← اڇ
11. مفعول جو لڳ← جي گهر
12. فاعل جولڳ← منهنجو

Figure 5. LSL rules for Sindhi Sentence.

Table 6

LP Constraints	‹ فاعل› ‹ فعل › ‹ مبتدا › خبر	
Constituents	**فعل**	**فاعل**
Sentence	لکي ٿو	علي

Table 7

LP Constraints	‹ فاعل› ‹ فعل › ‹ مبتدا › خبر	
Constituents	**فاعل**	**فعل**
Sentence	علي	لکي ٿو

The sentence **"لکي علي ٿو"** is not grammatically correct according to above LSL grammar because it violates the LP constraint of rule 1 which says that *Mubtada* must be continuous.

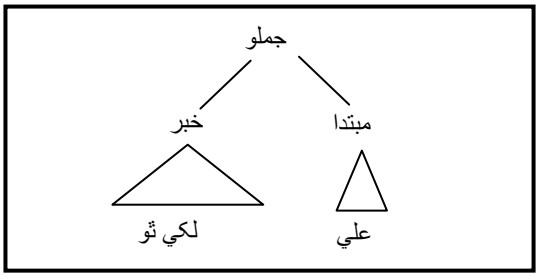

Figure 6. Derivation tree for sentences (علي لکي ٿو) and (لکي ٿو علي)

Now consider another sentence with more constituent parts **"منهنجو دوست اچ سليم جي گهر ويوآهي"**. All the following word orders are correct according to LSL grammar of Figure 5.

- منهنجو دوست اچ سليم جي گهر ويوآهي
- دوست منهنجو اچ سليم جي گهر ويوآهي
- اچ منهنجو دوست سليم جي گهر ويوآهي
- اچ دوست منهنجو سليم جي گهر ويوآهي
- منهنجو دوست سليم جي گهر ويوآهي اچ
- دوست منهنجو سليم جي گهر ويوآهي اچ
- سليم جي گهر ويوآهي اچ دوست منهنجو
- سليم جي گهر ويوآهي اچ منهنجو دوست
- اچ سليم جي گهر ويوآهي منهنجو دوست
- اچ سليم جي گهر ويوآهي دوست منهنجو
- سليم جي گهر اچ دوست منهنجو ويوآهي
- سليم جي گهر اچ دوست منهنجو ويوآهي
- سليم جي گهر اچ ويوآهي منهنجو دوست
- سليم جي گهر اچ ويوآهي دوست منهنجو

The constituent ordering of two of the above sentences according to LP constraints is given below in Table 8 and 9.

Table 8

LP Constraints	خبر				‹ مبتدا ›	
	‹ فعل › ‹ فعل جو لگ› مفعول جو لگ » مفعول				‹ فاعل› فاعل جو › لگ	
Constituents	فعل	فاعل	فاعل جو لگ	فعل جو لگ	مفعول جو لگ	مفعول
Sentence	ويوآهي	دوست	منهنجو	اچ	جي گهر	سليم

Table 9

LP Constraints	خبر				‹ مبتدا ›	
	‹ فعل › ‹ فعل جو لگ› مفعول جو لگ » مفعول				‹ فاعل› فاعل جو › لگ	
Constituents	فاعل جو لگ	فاعل	فعل	فعل جو لگ	مفعول جو لگ	مفعول
Sentence	منهنجو	دوست	ويوآهي	اچ	جي گهر	سليم

Word order is not completely free in above sentences. For example the following sentences are grammatically incorrect according to LSL grammar of Figure 5.

- دوست اچ سليم جي گهر ويوآهيمنهنجو
- منهنجو اچ دوست سليم جي گهر ويوآهي
- منهنجو اچ سليم جي گهر ويوآهي دوست

In all the sentences given above the continuation of *Mubtada* is violated as is specified by rule 1. Derivation tree for all constituent orders is shown in Figure 7.

LSL rules of Figure 5 are identical to the Urdu rules given in Figure 3; because of the similarities of both Urdu and Sindhi language sentence structures.

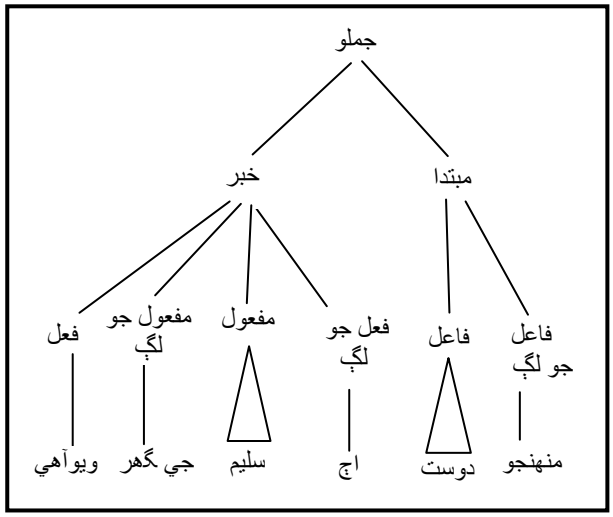

Figure 7. Derivation Tree of a SampleSentence
(منهنجو دوست اچ سليم جي گهر ويوآهي)

4. CONCLUSION

Context Free Grammars are not suitable for computational treatment of free/partial word order syntax. Linear Specification Language grammar is an extension of CFGs that can be used to handle the partial or free word order of natural language text. LP constraints are used to define the word orders with or without restrictions in sentences. LSL is quite capable of defining partial word order of Urdu/Sindhi sentences. One LSL rule can describe so many proper word orders of a sentence which is not possible by CFG in which we need to write different CFG rule for every proper order of words/constituents in a sentence. Once LSL is defined it can be used to validate all different word/constituent orders of a sentence (which is actually checking syntax of many sentences with same rule). One interesting fact is that we need to construct only one derivation tree (also known as parse tree) for all proper word orders of a sentence. This is because word order information is embedded inside LSL rules in the shape of LP constraints. Another interesting fact that is shown by LSL examples is that Urdu and Sindhi LSL grammars seem to be identical which strengthens the possibility of grammar parallelism in these languages.

REFERENCES

[1] Johen E. Hopcroft, Rajeev Motwani, Jeffrey D. Ullman (2001) *Introduction to Automata Theory, Languages*, and Computation. Pearson Education Inc 2001.

[2] Mutee-u-Rahman, Asadullah Shah (2003), "Grammar Checking Model for Local Languages" . In *proceedings of SCONEST (Student Conference on Engineering Sciences and Technology) 2003*. SCON-S15, Hamdard & Bahria University Karachi Pakistan, October 2003.

[3] Mutee-u-Rahman and Asadullah Shah (2004), "Grammar Checking of Urdu and Sindhi Sentences by Using W3C XML Schema." In *Proceedings of IEEE, ACM NCET (National Conference on Emerging Technologies)* 2004 Karachi. pp 120-125.

[4] Akshar Bharati Vineet Chaitanya and Rajeev Sangal (200). *Natural Language Processing A Paninian Perspective*. Prentice Hall of India Pvt Limited 2000.

[5] Thilo Goetz and Gerald Penn (1997). A Proposed Linear Specification Language. Volume 34. Arbeitspapiere des SFB 340. Universität Tübinge

[6] Suhre, O. (1999). Computational aspects of a grammar for-malism for languages with freer word order. Diplomarbeit.(Volume 154 in Arbeitspapiere des SFB 340, 2000).

[7] Bashir Ahmed Siddiqui (2000), *Jadid Urdu Composition*. Kitabistan Publishing Company.

[8] Ghulam Jilani Makhdoom(1992), *Darsi Urdu composition*. Darsi Idara Limited Educational Publishers.

[9] Hassan Ali Thaeem (2001), *Sindhi Grammar and Composition*. Rehbar Publishers.

[10] Ali Muhammad Baloch(200), *Rahber-e-Sindhi Composition and Grammar*. Gaba Educational Books.