

# Grammar Checking of Urdu and Sindhi Sentences by Using W3C XML Schema

Mutee u Rahman<sup>1</sup> and Asadullah Shah<sup>2</sup>

[muteerahman@gmail.com](mailto:muteerahman@gmail.com) and [dr\\_asadullahshah@hotmail.com](mailto:dr_asadullahshah@hotmail.com)

<sup>1</sup> Institute of Mathematics and Computer Science University of Sindh Jamshoro Pakistan

<sup>2</sup> Department of Computer Science Isra University Hala Road Hyderabad Sindh 71000 Pakistan

**Abstract:** Controlled versions of Urdu and Sindhi language sentences are representable in different mathematical models like Context Free Grammars (CFGs). A CFG is a finite collection of rules which tells that a given sentence (of formal or natural language) is grammatically correct or incorrect. Sentence structure of various natural languages is describable by using CFG and sentences can be parsed efficiently by using different CFG parsing techniques.

One of the applications of CFG is the description of XML document format via the notion of document type definitions (DTDs) and Schema. CFG rules are describable by using W3C XML Schema (standard recommendation of XML Schema by w3.org). It means that if CFG rules are represented in XML Schema then well formed XML documents of natural languages like Sindhi and Urdu can be validated by using those Schemas.

When CFG rules of Urdu and Sindhi sentences are represented by W3C XML Schema, text documents of these languages if represented in XML can be validated by using XML parsers. Issues related with the representation and validations of local language sentences in XML are discussed in this paper. CFG rules for samples of Sindhi and Urdu sentences are represented in W3C XML Schema. It is shown that W3C XML Schema can be used to validate the sentences of Urdu and Sindhi languages.

**Keywords:** Context Free Grammar, Local Language Processing, NLP, XML Schema

## 1. INTRODUCTION

Different models from formal language theory are being used to process natural languages on computers. Context free grammar (CFG) is one such model that is most commonly used to represent language syntax (grammar) rules for English and other natural languages [1]. Natural Languages like Urdu and Sindhi possess complex sentence structures and formal language models like CFG can not handle all the aspects of syntax for these languages. Despite of these limitations CFG can be defined for different sentence types of these natural languages [2]. CFG rules for different types of sentences can be used for grammar checking by the help of parsers.

### 1.1 Urdu Sentence Structure

Every sentence (جمله) in Urdu is divided into two parts Masand Alya (مسند اليه) and Masand (مسند) [3][4]. Masand

Alya is the beginning of the sentence and is always a noun (اسم). Masand is information about Masand Alya and can be a noun or verb (فعل). Sentences in which masand is noun are known as Ismia sentences (جمله اسميه) and sentences in which Masand is verb are called Failia sentences (جمله فعليه).

Failia and Ismia sentences are again divided into following constituent parts:

- Ismia Sentence
  - Mubtada (مبتدا)
  - MutaliqueMubtada (متعلق مبتدا) (Optional)
  - Khabar (خبر)
  - MutaliqueKhabar (متعلق خبر) (Optional)
  - FaileNaqis (فعل ناقص)
- Failia Sentence
  - Subject (Fa'ail) (فاعل)
  - MutaliqueFa'ail (متعلق فاعل)
  - Object (Mafool) (مفعول)
  - Verb (Fail) (فعل)
  - MutaliqueFail (متعلق فعل)

Different constituent parts of a sentence (Failia or Ismia) are again divided into different subcategories [3][4]. Sentences are grammatically correct if all the constituent parts and their sub parts occur in proper order.

### 1.2 Sindhi Sentence Structure

Sindhi sentence (جمله) is also divided in two parts known as Mubtada (مبتدا) and Khabar (خبر) [5][6].

Mubtada and Khabar are again divided into following parts:

- Mubtada
  - Subject (Fa'ail) (فاعل)
  - Fa'ailJoLag (فاعل جو لڳ)
- Khabar
  - Object (Mafool) (مفعول)
  - MafoolJoLag (مفعول جو لڳ)
  - Verb (Fail) (فعل)
  - FailJoLag (فعل جو لڳ)

Again different subparts are divisible into more parts. Urdu and Sindhi sentences share many constituent parts like Mubtada, Khabar and their subparts. CFG rules can be defined for different types of Urdu and Sindhi sentences and can be applied in different language processing systems.



```

    <xs:element ref="Faail"/>
  </xs:element ref="MutaliqeFaail"/>
</xs:sequence>
<xs:sequence>
  <xs:element ref="MutaliqeFaail"/>
  <xs:element ref="Faail"/>
</xs:sequence>
</xs:choice>
</xs:complexType>
</xs:element>
<!-- Masand Definition (مسند) -->
<xs:element name="Masand">
  <xs:complexType>
    <xs:choice>
      <xs:sequence>
        <xs:element ref="Mafool" minOccurs="0"/>
        <xs:element ref="MutaliqeMafooll"
          minOccurs="0"/>
        <xs:element ref="Fail"/>
        <xs:element ref="MutaliqeFail"
          minOccurs="0"/>
      </xs:sequence>
      <xs:sequence>
        <xs:element ref="MutaliqeFail"/>
        <xs:element ref="Fail"/>
      </xs:sequence>
      <xs:choice>
        <xs:sequence>
          <xs:element ref="Khabar"/>
          <xs:element ref="MutaliqeKhabar"/>
        </xs:sequence>
        <xs:sequence>
          <xs:element ref="MutaliqeKhabar"/>
          <xs:element ref="Khabar"/>
        </xs:sequence>
        <xs:element ref="Isam"/>
      </xs:choice>
    </xs:choice>
  </xs:complexType>
</xs:element>
<!-- FaileNaqis (فعل ناقص) Definition -->
<xs:element name="FaileNaqis">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="ہے"/>
      <xs:enumeration value="تھا"/>
      <xs:enumeration value="ہوگا"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<!-- Mubtada (مبتدا) Definition -->
<xs:element name="Mubtada">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Isam"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<!-- MutaliqeMubtada (متعلق مبتدا) Definition -->
<xs:element name="MutaliqeMubtada">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="میرا"/>
        <xs:enumeration value="تمہارا"/>
        <xs:enumeration value="اسکا"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <!-- Isam (اسم) Definition -->
  <xs:element name="Isam">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="الکبر"/>
        <xs:enumeration value="کراچی"/>
        <xs:enumeration value="حیدرآباد"/>
        <xs:enumeration value="بازار"/>
        <xs:enumeration value="بندرگاہ"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <!-- MutaliqeFail (متعلق فعل) Definition -->
  <xs:element name="MutaliqeFail">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="بازار سے"/>
        <xs:enumeration value="گھر سے"/>
        <xs:enumeration value="گھر میں"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <!-- Fail (فعل) Definition -->
  <xs:element name="Fail">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="روٹا ہے"/>
        <xs:enumeration value="کہتا ہے"/>
        <xs:enumeration value="کہتا ہے"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <!-- Mafool (مفعول) Definition -->
  <xs:element name="Mafool">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Isam"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <!-- MutaliqeMafool (متعلق مفعول) Definition -->
  <xs:element name="MutaliqeMafooll">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="کے گھر"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <!-- Khabar (خبر) Definition -->
  <xs:element name="Khabar">
    <xs:complexType>
      <xs:sequence>

```

```

    <xs:element ref="Isam"/>
  </xs:sequence>
</xs:complexType>
</xs:element>
<!--MutaliqeKhabar (متعلق خبر) Definition -->
<xs:element name="MutaliqeKhabar">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="سارے"/>
      <xs:enumeration value="تھوڑے سے"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<xs:element name="Faail">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="Isam"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="MutaliqeFaail">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="ايگ"/>
      <xs:enumeration value="مير"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
</xs:schema>

```

### 3. W3C XML SCHEMA FOR CFG RULES OF SINDHI SENTENCE

CFG rules for Sindhi sentence and its constituent parts are given below:

جملو ← ميندا خبر  
 ميندا ← فاعل جو لڳ (Optional) فاعل  
 ميندا ← فاعل (فاعل جو لڳ (Optional)  
 خبر ← فعل جو لڳ (Optional) مفعول (Optional)  
 مفعول جو لڳ (Optional) فعل  
 فاعل ← اسم  
 فعل ← لڪي ٿو  
 فعل ← ويو آهي  
 مفعول ← اسم  
 اسم ← علي  
 اسم ← سليم  
 فعل جو لڳ ← اڃ  
 مفعول جو لڳ ← جي گهر  
 فاعل جو لڳ ← منهنجو

W3C XML Schema for above CFG rules is as follows:

```

<?xml version="1.0" encoding="UTF-8"?>
<xs:schema
  xmlns:xs="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="qualified"
  attributeFormDefault="qualified">
  <!-- Definition for Sindhi Sentence (جملو) -->
  <xs:element name="Jumlo">

```

```

    <xs:complexType>
      <xs:choice>
        <xs:sequence>
          <xs:element ref="Mubtada"/>
          <xs:element ref="Khabar"/>
        </xs:sequence>
      </xs:choice>
    </xs:complexType>
  </xs:element>
  <!-- Faail (فاعل) Definition -->
  <xs:element name="Faail">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Isam"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
  <!-- Isam (اسم) Definition -->
  <xs:element name="Isam">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="علي"/>
        <xs:enumeration value="سليم"/>
        <xs:enumeration value="ڪراچي"/>
        <xs:enumeration value="حيدرآباد"/>
        <xs:enumeration value="دوست"/>
        <xs:enumeration value="دشمن"/>
        <xs:enumeration value="ساٿي"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <!-- Definition for Fa'ail Jo Lag (فاعل جو لڳ) -->
  <xs:element name="FaailJoLag">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="منهنجو"/>
        <xs:enumeration value="تنهنجو"/>
        <xs:enumeration value="هنجو"/>
        <xs:enumeration value="ڪنهنجو"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <!-- Definition for Fail Jo Lag (فعل جو لڳ) -->
  <xs:element name="FailJoLag">
    <xs:simpleType>
      <xs:restriction base="xs:string">
        <xs:enumeration value="اڃ"/>
        <xs:enumeration value="ڪلهه"/>
        <xs:enumeration value="سڀاڻي"/>
      </xs:restriction>
    </xs:simpleType>
  </xs:element>
  <!-- Mafool (مفعول) Definition -->
  <xs:element name="Mafool">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="Isam"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

```

```

<!-- Definition for Mafool Jo Lag (مفعول جو لڳ) -->
<xs:element name="MafoolJoLag">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="گهر جي"/>
      <xs:enumeration value="شهر جي"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<!-- Fail (فعل) Definition -->
<xs:element name="Fail">
  <xs:simpleType>
    <xs:restriction base="xs:string">
      <xs:enumeration value="ويو آهي"/>
      <xs:enumeration value="ايو آهي"/>
      <xs:enumeration value="ويندو"/>
      <xs:enumeration value="ٿو لڪي"/>
      <xs:enumeration value="پڙهندو"/>
    </xs:restriction>
  </xs:simpleType>
</xs:element>
<!-- Mubtada (مبتدا) Definition -->
<xs:element name="Mubtada">
  <xs:complexType>
    <xs:choice>
      <xs:sequence>
        <xs:element ref="FaailJoLag"/>
        <xs:element ref="Faail"/>
      </xs:sequence>
      <xs:sequence>
        <xs:element ref="Faail"/>
        <xs:element ref="FaailJoLag"/>
      </xs:sequence>
      <xs:sequence>
        <xs:element ref="Isam"/>
      </xs:sequence>
    </xs:choice>
  </xs:complexType>
</xs:element>
<!-- Definition for Khabar (خير) of a Sentence -->
<xs:element name="Khabar">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="FailJoLag" minOccurs="0"/>
      <xs:element ref="Mafool" minOccurs="0"/>
      <xs:element ref="MafoolJoLag" minOccurs="0"/>
      <xs:element ref="Fail"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
</xs:schema>

```

#### 4. RESULTS

Altova Style Vision 2005 and Microsoft Excel 2003 Professional Edition are used here to demonstrate the use of W3C XML Schema for Sindhi sentence validation (see Figure 1). Figure 2 shows the demonstration of invalid XML sentences in AltovaStyle Vision 2005.

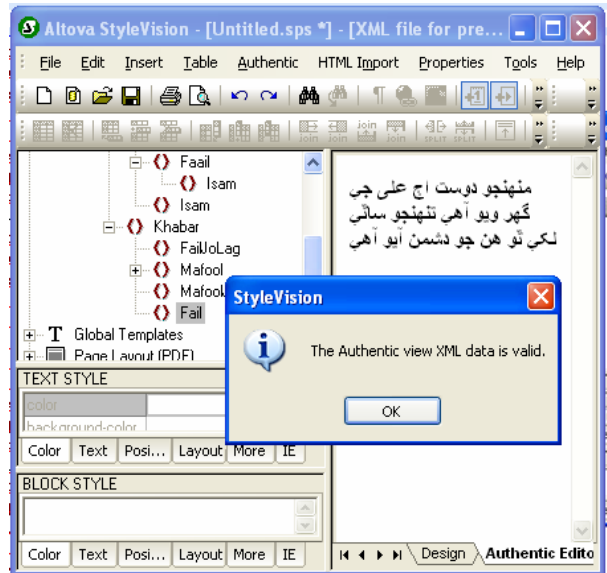


Figure 1. Validation of Sindhi XML Sentences in Altova StyleVision 2005 by using Sindhi sentence XML schema.

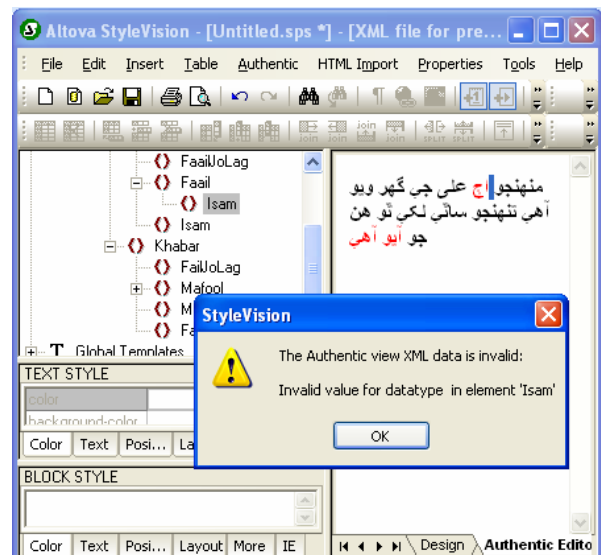


Figure 2. Error message for invalid Sindhi XML Sentences in AltovaStyleVision 2005.

Microsoft Excel 2003 Professional Edition also supports schema validated data by applying XML maps. New XML map can be created by using following steps:

1. On the **Data** menu, point to **XML**, and then click **XML Source** to open the **XML Source** task pane.
2. To map one or more elements to your worksheet, select the elements in the **XML Source** task pane. To select nonadjacent elements, click one element, and then hold down CTRL and click each element.
3. Drag the selected elements to the worksheet location where you want them to appear.

Once XML map is created one can import valid Sindhi XML files into Excel sheet. XML map for Sindhi Sentence by using W3C XML Schema with valid XML document data import is shown in Figure 3.

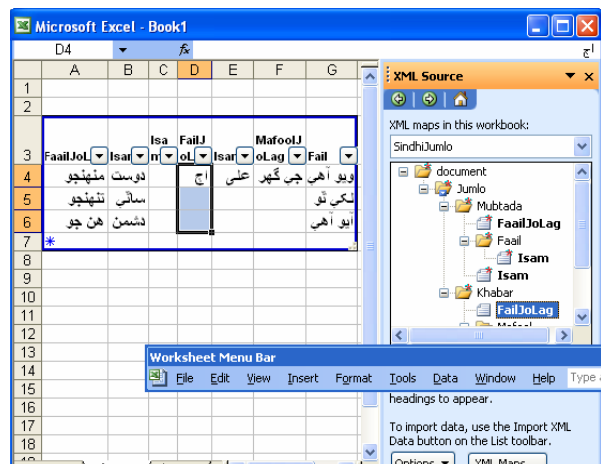


Figure 3. Valid XML data import by using W3C XML Schema map in Microsoft Excel 2003.

## 5. CONCLUSION

CFGs can handle many aspects of natural language syntax. Languages like Urdu and Sindhi have complex sentence structure which is not completely describable by CFG. Even though CFGs are considerably useful for the description of syntax rules of Urdu and Sindhi sentences. CFGs are also used for the description of data validation rules in XML documents. W3C XML schema is standard formalism for the description of data validation rules. CFGs are completely describable by XML Schemas, it means CFGs of Urdu and Sindhi sentences can be converted into XML Schemas and those Schemas can be used for the validation of Urdu and Sindhi XML documents.

Because schemas are XML documents therefore they are flexible, reprogrammable and easily deployable in any type of application. Schemas can be embedded in different applications without any special consideration. If valid XML Schemas are created for different types of Urdu and Sindhi sentences then it is possible to check the grammar of those sentences in any type of application with XML support. XML Schemas given in above sections can be used very easily and effectively in applications like

Microsoft Word 2003 Professional Edition, Microsoft Excel 2003 Professional Edition and Altova Style Vision. W3C XML Schema for Sindhi and Urdu can also be used for automated sentence generation systems, natural language interfaces to computers, machine translation and for knowledge representation of AI systems.

## REFERENCES

- [1] Daniel Jurafsky & James H. Martin. *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Pearson Education Series 2002.
- [2] Mutee-u-Rahman, Asadullah Shah. Grammar Checking Model for Local Languages. In *proceedings to the SCONEST (Student Conference on Engineering Sciences and Technology) 2003*. SCONEST-15, Hamdard & Bahria University Karachi Pakistan, October 2003.
- [3] Bashir Ahmed Siddiqui. *Jadid Urdu Composition*. Kitabistan Publishing Company 2000.
- [4] Ghulam Jilani Makhdoom. *Darsi Urdu Composition*. Darsi Idara Limited Educational Publishers 1992.
- [5] Hassan Ali Thaeem. *Sindhi Grammar and Composition*. Rehbar Publishers 2001.
- [6] Ali Muhammad Baloch. *Rahber-e-Sindhi Composition and Grammar*. Gaba Educational Books 2000.
- [7] Elliott Rusty Harold. *XML Bible*. Hungry Minds Inc 2001.
- [8] K. M. Goeschka, Helmut Reis, Robert Smeikal. XML Based Robust Client-Server Communication for a Distributed Telecommunication Management System. *36th Annual Hawaii International Conference on System Sciences (HICSS'03)*. pp. 122c - Track 4. January 2003 IEEE.
- [9] Petr Kroha, Lars Gemeinhardt. Using XML in a Web-Oriented Information System. *12th International Workshop on Database and Expert Systems Applications*. September 2001 IEEE.