Sentence Reduction for Syntactic Analysis of Compound Sentences in Punjabi Language

S.K Sharma

DAV University, Jalandhar Punjab (India)

Abstract

Objective: This research paper is an attempt to develop a syntactic analysis system for compound sentences of Punjabi language. Methods/Statistical Analysis: Sentence simplification approach has been used for splitting compound sentences into simple sentences and then analyzing these simple sentences for syntactical error. A full form lexicon based morph, HMM based POS tagger and set of rules have been used for identification of grammatical mistakes. Findings: On testing an overall precision as 93.30, recall as 97.32 and F-measure as 95.25 is reported by the system. Application/Improvements: the system shows better performance on comparing it with existing Punjabi grammar (Precision=76.79, Recall=87.08) and Myanmar grammar checker that works on compound sentence and shows precision 83.75%.

Keywords: Grammar Checker, Compound Sentences, Sentence simplification, Punjabi Sentences.

1. Introduction

Syntactic analyzer is an automated tool used in most of the natural language engineering resources. It is usually incorporated with most of the word processing systems. A lot of work has been done in grammatical error detection and correction, mainly in English language. Very little work has been done for Indian languages. Probably, Bangla grammar checker [Alam, M. Jahangir et.al (2006)] [1], Urdu [Kabir, H. et.al (2002)] [30] and Punjabi grammar checker [Singh M et.al (2008), Sharma S. K. et.al] [8], [29] are the only systems developed for Indian languages.

As the length of a sentence increases, it becomes difficult to syntactically analyze it. Therefore it becomes necessary to decrease the length of the sentence by splitting it into more than one sentences. Long sentence generally falls into category of compound sentences and complex sentences. Compound sentences are generally composed of two or more independent clauses and therefore can be split into more than one simple sentences depending upon the number of independent clauses present in the compound sentence. In most of the cases while performing the syntactic analysis on compound sentences, it becomes difficult to identify the boundary of independent clauses and hence syntactic analyzer may raise false alarm for such sentences. Therefore, it becomes essential to simplify the large sentence by reducing the length of the sentence.

1.1 Introduction to Punjabi language

Punjabi language belongs to Indo-Aryan family of languages (Indic languages). Other members that belong to this family are Hindi, Bengali, Gujarati, Marathi etc. Punjabi is spoken in India, Pakistan,
Canada, USA, UK, and other countries with Punjabi immigrants. Punjabi language is the 10th most widely spoken language in the world, 4th most spoken language in Canada (The Times of India, 14th February, 2008) and the 11th in India with more than 29 million speakers. It is the official language of Punjab state. Punjabi is written in ‘Gurmukhi’ script in eastern Punjab (India), and in ‘Shahmukhi’ script in western Punjab (Pakistan). As compared to other languages like English, Punjabi is a morphologically rich language and has relatively free word order. It follows a Subject-Object-Verb (S-O-V) pattern.

2. Introduction to Machine learning

Machine learning can be defined as the field that provide the computer the capability to learn without being programmed. Sometimes it is called branch of Artificial Intelligence and generally both terms are used interchangeably. Machine learning has application in many areas including medical science [59], finance [60], query optimization, pattern recognition, healthcare sectors [61] [62] etc. It also plays an important role in Natural language processing especially in text processing, for identification of Part of speech tags, sentiment analysis, identification of entities etc. In NLP the ML technique can be expressed as model that can be used to generate other text from a given text i.e. paraphrasing [63]. It can also be used as set of instructions (algorithm) that can be used to extract useful information from the text like summarization system. Syntactic analysis of text is also one of the application of the ML.

3. Approaches Used For Syntactic Analysis

There are basically three approached used for grammar checking. These includes rule based approach [Daniel Naber, 2003] [6], syntax based approach [Jensen et al, 1993] and statistics based approach [Attwell, 1997]. The rule based approach has been used for Dutch language [Vosse, 1992][17], Czech and Bulgarian language [Kubon and Plátek, 1994][18], English language [Adriaens, 1994][19], Swedish language[Hein, 1998][20], German language [Schmidt-Wigger, 1998][21], English language [Ravin, 1998][22], Korean language [Young-Soog, 1998][23], Danish language [Paggio, 2000][24], French, German, and Spanish languages [Helfrich and Music, 2000][25], French language [Vandeventer, 2001][26], Swedish language [Carlberger et al., 2002, 2004][27], German language [Fliedner, 2002][28], Swedish language [Kann 2002 and Bigert et al. 2004][13], Urdu language [Kabir et al., 2002][30], English language [Naber, 2003][6], Swedish language [Hashemi, 2003][31], English language [Moré et al. 2004], [Rider, 2005][32], Brazilian Portuguese language [Kinoshita et al., 2006], Nepali language [ Bal and Shrestha, 2007], Persian language [Ehsan and Faili, 2010][14], Afan Oromo (language widely spoken and used in Ethiopia) [Tesfaye, 2011][33], Chinese language [Jiang et al., 2011][34], Malay language [Kasbon et al., 2011][35], Tagalog Filipino (official language of the Philippines) [Oco and Borra , 2011]. Statistics based approach has been used for English language [Park et al., 1997][9], French writers writing in English [Tschichold et al., 1997][10], English language [Powers, 1997][11], Swedish [Arppe, 1999][12], Bangla and English languages [Alam et al., 2006][1], Swedish language [Sjobergh, 2006][13], Persian language [Ehsan and Faili, 2010][14], Amharic language [Temesgen and Assabie, 2012][15], A Language Independent Statistical Grammar (LISG) checking system [Verena Henrich and Timo Reuter, 2009][16]. All these three approaches have their own advantages and drawbacks.

4. Approaches Used For Sentence Simplifications

Various techniques have been used for simplification of large sentences by different researchers. These includes Lexical Approach used for development of PSET (The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers) [38], KURA (Text Simplification for Reading Assistance) [44], HAPPI (Helping Aphasic People Process Online Information) [45], SIMPLEX (Putting It Simply: a Context-aware Approach to Lexical Simplification) [39], LexSiS: Lexical Simplification for Spanish [46] etc. Generation approach used for automatic Induction of rules for text simplification [36], simplification of Newspaper text to assist Aphasic reader [47], text simplification for information seeking applications [37], maintaining discourse when performing syntactic simplification [40], splitting of long sentence after explanation.
5. Proposed model

Compound sentences are composed of at least two independent clauses joined by coordinate conjunctions, comma or semicolon and these exists in fixed patterns [2], [58]. For grammar checking of compound sentences, each clause is extracted from the sentence and grammar checking is performed on it. Since there may be two to any number of independent clauses present in compound sentences, therefore, researcher proposed divide and conquer model that can be used for grammar checking of compound sentences. In accordance with divide and conquer, the compound sentence is simplified by splitting it into individual clauses and then each individual clause undergoes error detection and correction mechanism. In this way, overall grammar checking process for compound sentences takes place in two steps; first step is to split the compound sentence into simple sentences and second step is to perform grammar checking on each simple sentence. For extracting the independent clauses from the compound sentence, the clause boundary of these clauses is identified in the similar way as in [56] and [57]. Researcher’s system checks the errors in the sentence at phrase level, clause level and then at sentence level. For phrase level and clause level, rule based approach has been followed. For sentence level, this rule based approach has been extended to all the clauses of the sentence. The syntactic analysis in compound sentences takes place in three phases:

**Figure 1:** Proposed model for syntactic analysis of compound sentences in Punjabi language

### 5.1 Phase I

In this phase, style error at the sentence level is identified and rectified. This is done by analyzing the complete structure of the input compound sentence. Compound sentences have fixed patterns of structure [2], [3]. An input compound sentence is matched against these fixed patterns. In order to match the input sentence against these patterns, a database
containing all the possible patterns of compound sentences has been developed and stored in the form of regular expression.

5.2 Phase II

In the second phase, internal structure of compound sentences is analyzed for detection of errors. In the internal structure, errors at phrase level and then at clause level are detected and rectified. In order to perform error detection and correction at phrase and clause levels, each input compound sentence is simplified by splitting the sentence into independent clauses (simple sentences). This simplification of sentences is performed on the basis of clause boundary mark information [56], [57].

5.3 Phase III

To check the grammatical errors in simple sentences or in independent clauses of Punjabi sentences, ‘government and binding’ prevalent in Punjabi sentences has been studied. As per ‘government and binding’ prevalent, there exists a grammatical agreement between various components of a sentence like modifier and noun agreement, subject/object and verb agreement, noun and adjective agreement, noun phrase in oblique form before postposition etc. Also, as per government and binding, all the words present in an independent clause must grammatically agree with the head word of that clause. The head word of the clause is the first phrase head that is present in the noun phrase of the clause. Various types of errors detected and corrected by the system developed by the researcher has been listed in table 1. For each error type mentioned in table 1, a separate module has been developed. Each module detects and rectifies a specific type of error. During detection of error, all these modules are executed in sequence.

The algorithm used in the grammar checking of compound sentences is as following:

**Algorithm: Error detection in compound sentences**

**Databases used:** Error type.

**Input:** Incorrect simple sentence (independent clause obtained after simplification)

**Output:** Corrected simple sentence.

1. Get all the error type that have OnOff value set to 1, from the error type database sorted by the Priority field
2. Repeat steps 3 to 5 for current clause.
3. Repeat steps 4 and 5 for all the error types.
4. Call the respective module to perform the correction on the current clause.
5. Output the corrected sentence.

Consider the following incorrect compound sentence:

**Incorrect sentence:**

Punjabi: ਦੋ ਮੋਟਾ ਮੁੱਢੇ ਭੱਜ ਰਿਹਾ ਸੀ ਅਤੇ ਪੁਲਿਸ ਓਹਨਾਂ ਦਾ ਰਪੱਛਾ ਕਿ ਰਿਹਾ ਸੀ।

Roman Transliteration: (dō mōṭā muṇḍē bhajj rihā sī ātē pulis ōhnāṃ dā picchā kar rihā sī.)

English translation: (Two fatty boys run+ing was and cop them chase + ing was)

Two fatty boys was running and cop was chasing them.

In above incorrect sentence, there are two clauses in the compound sentence and each clause contains errors. The first clause is “ਦੋ ਮੋਟਾ ਮੁੱਢੇ ਭੱਜ ਰਿਹਾ ਸੀ “ (dō mōṭā muṇḍē bhajj rihā sī) and it contains noun modifier agreement error as the modifier ਮੋਟਾ (mōṭā) (singular) does not grammatically agree with noun ਮੁੱਢੇ (muṇḍē) (plural) in terms of number. Second clause is ਪੁਲਿਸ ਓਹਨਾਂ ਦਾ ਰਪੱਛਾ ਕਿ ਰਿਹਾ ਸੀ (pulis ōhnāṃ dā picchā kar rihā sī.), and it contains subject verb agreement error as the object ਪੁਲਿਸ (pulis) (feminine) does not grammatically agree with verb ਰਿਹਾ (rihā) (masculine) in terms of gender. Both these errors are detected and rectified in two steps. In the first step, two clauses are separated from the sentence and in the second step, these clauses are detected for the presence of error. After applying detection and correction on individual clauses, the final updated output given by the researcher’s system is:
Corrected sentence:

Punjabi: ਦੋ ਮੋਟੇ ਮੰਡ ਭੱਜ ਰਿਹਾ ਸੀ ਅਤੇ ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ।

Roman Transliteration: (dō mōṭē muṇḍē bhaj rahi śi ate pūliś oṁnām dā pīchā kār raḥi śi.)

English Translation: (Two fatty boys run+ing were and cop them chase + ing was)

Two fatty boys were running and cop was chasing them.

The complete architecture of the above method with example has been shown in figure 2. It is clear from figure 2 that the compound sentence is first simplified by splitting it at the conjunction and separating each independent clause. Then each independent clause is passed through grammar checking system where error detection and correction algorithm mention above is applied.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Compound Sentence</td>
<td>ਦੋ ਮੋਟੇ ਮੰਡ ਭੱਜ ਰਿਹਾ ਸੀ ਅਤੇ ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ।</td>
</tr>
<tr>
<td>Extract independent clauses</td>
<td>ਦੋ ਮੋਟੇ ਮੰਡ ਭੱਜ ਰਿਹਾ ਸੀ</td>
</tr>
<tr>
<td>IDC 1</td>
<td>ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ</td>
</tr>
<tr>
<td>IDC 2</td>
<td>ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ</td>
</tr>
<tr>
<td>Conjunction</td>
<td>ਜੇ ਮੇਟੇ ਮੰਡ ਭੱਜ ਰਿਹਾ ਸੀ ਅਤੇ ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ।</td>
</tr>
<tr>
<td>Grammar Checking System</td>
<td>ਜੇ ਮੇਟੇ ਮੰਡ ਭੱਜ ਰਿਹਾ ਸੀ ਅਤੇ ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ।</td>
</tr>
<tr>
<td>Updated IDC 1</td>
<td>ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ</td>
</tr>
<tr>
<td>Updated IDC 2</td>
<td>ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ</td>
</tr>
<tr>
<td>Conjunction + Updated IDC1 + Updated IDC2 = Corrected sentence</td>
<td>ਦੋ ਮੋਟੇ ਮੰਡ ਭੱਜ ਰਿਹਾ ਸੀ ਅਤੇ ਪੰਛੀ ਦੇਖਣ ਲਈ ਕਿ ਰਿਹਾ ਸੀ।</td>
</tr>
</tbody>
</table>

Figure 2: Architecture of grammar checking of compound sentence

6. Error Covered

Various types of syntactic errors in an independent clause handled by researcher’s system are listed in table 1. These errors are basically categorized into three classes; first is agreement error; second is postposition related errors and third is error due to order of words in noun and verb phrases. First column of the table represents the error category, second column represents the description of the grammatical mistake and third column shows the example containing incorrect and correct sentences related with the corresponding error shown in second column.
Table 1: Various Error types handled by the system

<table>
<thead>
<tr>
<th>Error Category</th>
<th>Description of error</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Agreement error              | All the Noun phrases joined by conjunctions to form group must agree in case.          | Incorrect: ਮੁੰਡੇ ਅਤੇ ਕੋਈ ਦੀਪਪ੍ਰੋਕਾਰ ਦੇ ਮਾਤਮਾਤ ਹੋਇ ਅੰਦਰ ਵੇਲੇ ਦੁਆਰਾ ਦੇ ਵੋ ਪ੍ਰਯੋਗਾਂ। (ਮੁੰਡਾ ਅਤੇ ਕੁਦਰਾਤ ਨੇ ਸਮਾਗਮ ਵੀ ਸਾਦਾ ਕੇ ਹੀਸ਼ਾ ਲਾਈ।)  
Correct: ਮੁੰਡੇ ਅਤੇ ਕੋਈ ਦੀਪਪ੍ਰੋਕਾਰ ਦੇ ਮਾਤਮਾਤ ਹੋਇ ਅੰਦਰ ਵੇਲੇ ਦੁਆਰਾ ਦੇ ਵੋ ਪ੍ਰਯੋਗਾਂ। (ਮੁੰਡਾ ਅਤੇ ਕੁਦਰਾਤ ਨੇ ਸਮਾਗਮ ਵੀ ਸਾਦਾ ਕੇ ਹੀਸ਼ਾ ਲਾਈ।) |
| Noun Modifier Agreement      |                                                                                       | Incorrect: ਇੱਠ ਮੁੰਡੇ ਬੇਠ ਖੇਡ ਹੋ ਗਈ। (ਈਥ ਮੁੰਡੇ ਬੇਠ ਖੇਡ ਹੋ ਗਈ।)  
Correct: ਇੱਠ ਮੁੰਡੇ ਬੇਠ ਖੇਡ ਹੋ ਗਈ। (ਈਥ ਮੁੰਡੇ ਬੇਠ ਖੇਡ ਹੋ ਗਈ।) |
| Adjective Phrase Agreement    |                                                                                       | Incorrect: ਦੋ ਰਿਸਾਲਾ ਮੋਕਾਸ ਲਿਖਦੀ ਹੈ। (ਤੋਰ ਰਿਸ਼ ਮੋਕਾਸ ਲਿਖਦੀ ਹੈ।)  
Correct: ਦੋ ਰਿਸਾਲਾ ਮੋਕਾਸ ਲਿਖਦੀ ਹੈ। (ਤੋਰ ਰਿਸ਼ ਮੋਕਾਸ ਲਿਖਦੀ ਹੈ।) |
| Subject Verb Agreement        |                                                                                       | Incorrect: ਮੁੰਡੇ ਖੇਡ ਰਿਹਾ ਹੁੰਦਾ ਹੈ। (ਮੁੰਡਾ ਖੇਡ ਰਿਹਾ ਹੁੰਦਾ ਹੈ।)  
Correct: ਮੁੰਡੇ ਖੇਡ ਰਿਹਾ ਹੁੰਦਾ ਹੈ। (ਮੁੰਡਾ ਖੇਡ ਰਿਹਾ ਹੁੰਦਾ ਹੈ।) |
| In a phrase or clause, all the words in agreement with headword must have same gender and number. |                                                                                       | Incorrect: ਹਿੰਦ ਮੇਰੇ ਘਰ ਹੋ ਗਈ। (ਹਿੰਦ ਮੇਰੇ ਘਰ ਹੋ ਗਈ।)  
Correct: ਹਿੰਦ ਮੇਰੇ ਘਰ ਹੋ ਗਈ। (ਹਿੰਦ ਮੇਰੇ ਘਰ ਹੋ ਗਈ।) |
| Postposition related error    | Noun phrase must be in oblique form before postposition.                               | Incorrect: ਮੁੰਡੇ ਤੁਂ ਸ਼ਾਬ ਦੇ ਲਿਖਾ। (ਮੁੰਡਾ ਤੁਂ ਸ਼ਾਬ ਦੇ ਲਿਖਾ।)  
Correct: ਮੁੰਡੇ ਤੁਂ ਸ਼ਾਬ ਦੇ ਲਿਖਾ। (ਮੁੰਡਾ ਤੁਂ ਸ਼ਾਬ ਦੇ ਲਿਖਾ।) |
| DAA (ਦਾ) postposition should be in agreement with noun phrase in terms of number and gender |                                                                                       | Incorrect: ਮੁੰਡੇ ਦੀ ਖਾਂ ਖੁਦਾ ਦੇ ਵਾਲੀ। (ਮੁੰਡਾ ਦੀ ਖਾਂ ਖਾਂ ਦੇ ਵਾਲੀ।)  
Correct: ਮੁੰਡੇ ਦੀ ਖਾਂ ਖੁਦਾ ਦੇ ਵਾਲੀ। (ਮੁੰਡਾ ਦੀ ਖਾਂ ਖਾਂ ਦੇ ਵਾਲੀ।) |
| Order of words in a phrase related error | Order of modifier in Noun phrase                                                      | Incorrect: ਮੁੰਡੇ ਨੇ ਪ੍ਰਤੀਕ ਸੋਂਧ ਦੀ। (ਸੋਂਧ ਪਾਣਿ ਮੁੰਡਾ ਨੇ ਸਿੱਖੀ।)  
Correct: ਮੁੰਡੇ ਨੇ ਪ੍ਰਤੀਕ ਸੋਂਧ ਦੀ। (ਸੋਂਧ ਪਾਣਿ ਮੁੰਡਾ ਨੇ ਸਿੱਖੀ।) |
| Order of words in a verb phrase                                         |                                                                                       | Incorrect: ਮੁੰਡਾ ਸ਼ਾਨਦਰ ਨਾਕੀ ਦੇਖਦੀ। (ਮੁੰਡਾ ਸ਼ਾਨਦਰ ਨਾਕੀ ਦੇਖਦੀ।)  
Correct: ਮੁੰਡਾ ਸ਼ਾਨਦਰ ਨਾਕੀ ਦੇਖਦੀ। (ਮੁੰਡਾ ਸ਼ਾਨਦਰ ਨਾਕੀ ਦੇਖਦੀ।) |

Each independent clause is checked for various grammatical errors. All these errors are detected in sequence as mention in table 1.

7. Result and discussion:

For testing the final module of syntactic analyzer for compound sentences, researcher first input the dummy test data. This dummy data contains the sentences having the errors for which the system is developed. After testing with dummy test data, real test data (data collected from hand written test papers by 6th to 10th standard students studying Punjabi as second language) and Hindi to Punjabi machine translated test data (http://h2p.learnpunjabi.org/) has been used to test the system. Number of sentences used from each type of data is mentioned in table 2. The results obtained have been tabulated in table 3.1, 3.2, 3.3, and figure 3.1, 3.2, 3.3.
Table 2: Test data used to test various types of error

<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Main error</th>
<th>Sub-category of main error</th>
<th>Number of sentences from dummy test data</th>
<th>Number of sentences from real data</th>
<th>Number of sentences from Hindi to Punjabi Machine translation system</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Modifier and Noun agreement</td>
<td>1.1 Agreement in terms of Number</td>
<td>220</td>
<td>18</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.2 Agreement in terms of Gender</td>
<td>200</td>
<td>23</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.3 Agreement in terms of Case</td>
<td>45</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>2.</td>
<td>Subject Verb agreement</td>
<td>2.1 Agreement in terms of Number</td>
<td>200</td>
<td>13</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.2 Agreement in terms of Gender</td>
<td>156</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.3 Agreement in terms of Person</td>
<td>223</td>
<td>13</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.4 Use of postposition with subject if the verb is transitive in perfect form</td>
<td>96</td>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>3.</td>
<td>Noun and Adjective agreement</td>
<td>3.1 Agreement in terms of Number</td>
<td>269</td>
<td>13</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.2 Agreement in terms of Gender</td>
<td>210</td>
<td>23</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3 Agreement in terms of Case</td>
<td>222</td>
<td>14</td>
<td>35</td>
</tr>
<tr>
<td>4.</td>
<td>Order of modifier of Noun phrase</td>
<td>4.1 Pronoun precedes all other modifiers</td>
<td>104</td>
<td>23</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.2 Numeral precedes adjective</td>
<td>86</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5.</td>
<td>Order of word in word phrase</td>
<td></td>
<td>537</td>
<td>49</td>
<td>121</td>
</tr>
<tr>
<td>6.</td>
<td>Agreement of noun phrase with DA postposition</td>
<td></td>
<td>29</td>
<td>32</td>
<td>80</td>
</tr>
<tr>
<td>7.</td>
<td>Oblique case of noun before Postposition</td>
<td></td>
<td>132</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Total Number of sentences</strong></td>
<td><strong>2729</strong></td>
<td><strong>269</strong></td>
<td><strong>666</strong></td>
</tr>
</tbody>
</table>
### Table 3.1: Experimental evaluation of grammar checking of compound sentences (Dummy test data)

<table>
<thead>
<tr>
<th>Error type</th>
<th>Number of Errors in input sentences [A]</th>
<th>No. of errors corrected by the system [B]</th>
<th>No. of errors wrongly corrected by the system [C]</th>
<th>Recall $\frac{B+C}{A} \times 100$</th>
<th>Precision $\frac{B}{A} \times 100$</th>
<th>F score $\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modifier and Noun Agreement (MNA)</td>
<td>465</td>
<td>463</td>
<td>0</td>
<td>99.56989</td>
<td>99.56989</td>
<td>99.56989</td>
</tr>
<tr>
<td>Subject Verb Agreement (SVA)</td>
<td>675</td>
<td>674</td>
<td>1</td>
<td>100</td>
<td>99.85185</td>
<td>99.92587</td>
</tr>
<tr>
<td>Noun and Adjective Agreement (NAA)</td>
<td>701</td>
<td>701</td>
<td>0</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Order of Modifier of a Noun Phrase (OMNP)</td>
<td>190</td>
<td>187</td>
<td>2</td>
<td>99.47368</td>
<td>98.42105</td>
<td>98.94457</td>
</tr>
<tr>
<td>Order of Words in a Verb Phrase (OWVP)</td>
<td>537</td>
<td>531</td>
<td>5</td>
<td>99.81378</td>
<td>98.88268</td>
<td>99.34605</td>
</tr>
<tr>
<td>Due to Postposition DA (PDA)</td>
<td>29</td>
<td>23</td>
<td>4</td>
<td>93.10345</td>
<td>79.31034</td>
<td>85.65517</td>
</tr>
<tr>
<td>Due to Postposition NU (PNU)</td>
<td>132</td>
<td>128</td>
<td>3</td>
<td>99.24242</td>
<td>96.9697</td>
<td>98.0929</td>
</tr>
<tr>
<td>Overall Average</td>
<td>2729</td>
<td>2707</td>
<td>15</td>
<td>99.7435</td>
<td>99.19384</td>
<td>99.46791</td>
</tr>
</tbody>
</table>

### Table 3.2: Experimental evaluation of grammar checking of compound sentences (Real test data)

<table>
<thead>
<tr>
<th>Error type</th>
<th>Number of Errors in input sentences [A]</th>
<th>No. of errors corrected by the system [B]</th>
<th>No. of errors wrongly corrected by the system [C]</th>
<th>Recall $\frac{B+C}{A} \times 100$</th>
<th>Precision $\frac{B}{A} \times 100$</th>
<th>F score $\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modifier and Noun Agreement (MNA)</td>
<td>53</td>
<td>49</td>
<td>2</td>
<td>96.22642</td>
<td>92.45283</td>
<td>94.30189</td>
</tr>
<tr>
<td>Subject Verb Agreement (SVA)</td>
<td>46</td>
<td>43</td>
<td>2</td>
<td>97.82609</td>
<td>93.47826</td>
<td>95.60277</td>
</tr>
<tr>
<td>Error type</td>
<td>Number of Errors in input sentences [A]</td>
<td>No. of errors in input sentences [B]</td>
<td>No. of errors wrongly corrected by the system [C]</td>
<td>Recall (\frac{\mu+C}{\mu} \times 100)</td>
<td>Precision (\frac{\mu}{\mu+C} \times 100)</td>
<td>F score (\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2)</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td>----------------------------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------------------------</td>
<td>-----------------------------</td>
<td>----------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Modifier and Noun Agreement (MNA)</td>
<td>133</td>
<td>125</td>
<td>5</td>
<td>97.74436</td>
<td>93.98496</td>
<td>95.8278</td>
</tr>
<tr>
<td>Subject Verb Agreement (SVA)</td>
<td>113</td>
<td>102</td>
<td>6</td>
<td>95.57522</td>
<td>90.26549</td>
<td>92.8445</td>
</tr>
<tr>
<td>Noun and Adjective Agreement (NAA)</td>
<td>124</td>
<td>118</td>
<td>4</td>
<td>98.3871</td>
<td>95.16129</td>
<td>96.74731</td>
</tr>
<tr>
<td>Order of Modifier of a Noun Phrase (OMNP)</td>
<td>65</td>
<td>58</td>
<td>3</td>
<td>93.84615</td>
<td>89.23077</td>
<td>91.48028</td>
</tr>
<tr>
<td>Order of Words in a Verb Phrase (OWVP)</td>
<td>121</td>
<td>116</td>
<td>3</td>
<td>98.34711</td>
<td>95.86777</td>
<td>97.09161</td>
</tr>
<tr>
<td>Due to Postposition DA (PDA)</td>
<td>80</td>
<td>69</td>
<td>6</td>
<td>93.75</td>
<td>86.25</td>
<td>89.84375</td>
</tr>
</tbody>
</table>

Table 3.3: Experimental evaluation of grammar checking of compound sentences (Hindi to Punjabi Machine translation data)
Due to Postposition NU (PNU)

<table>
<thead>
<tr>
<th></th>
<th>30</th>
<th>24</th>
<th>5</th>
<th>96.66667</th>
<th>80</th>
<th>87.54717</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Average</td>
<td>666</td>
<td>612</td>
<td>32</td>
<td>96.6967</td>
<td>91.89189</td>
<td>94.23309</td>
</tr>
</tbody>
</table>

**Figure 3.1**: Experimental Evaluation of Grammar checking of compound sentences (Dummy Test Data)

**Figure 3.2**: Experimental Evaluation of Grammar checking of compound sentences (Using Real Data)
As shown in tables 3.1, 3.2, 3.3 and figure 3.1, 3.2 and 3.3 the developed system shows a recall as 99.74, precision as 99.14 and F-score as 99.46 for Dummy test data, recall as 95.53, precision as 88.84 and F-score as 92.07 for Real test data and recall as 96.69, precision as 91.89 and F-score as 94.23 for H2P machine translation test data. An overall precision as 93.30, recall as 93.30 and F-measure as 95.25 is reported by the system. Low precision (80%) is also reported in some cases (Error Due to postposition NU in case of H2P machine translation test data) due to presence of unknown or misspelled words in the input sentences. Further the developed system will show low performance when there are number of spelling or typing mistakes in the Punjabi words. Also the system will not work in case of compound – complex sentences because the complex sentences can not be easily split into dependent and independent clauses.

8. Comparison with the existing system

This system can be compared only partially with the existing system because the existing grammar checker for Punjabi language has been developed to check the grammar mainly for simple sentences whereas researcher’s proposed system mainly deals with compound sentences. The existing Punjabi grammar checker system [8] reports precision of 76.79%, recall of 87.08%, and F-measure of 81.61%. Researcher’s system is tested for compound sentences for evaluating the results. For compound sentences, this system shows an overall precision as 93.30, recall as 97.32 and F-measure as 95.25. Also the developed system cannot be compared with the systems developed for other languages, due to the coverage of different types of errors by different grammar checking systems. As per reviewed literature the only grammar checking system that deal with compound sentences is the system developed by Lin et al. (2011) [58] for Myanmar language and this system shows a precision of 83.75% for compound sentences.
9. Conclusion and future scope
In this research author has developed a syntactic analyzer system for compound sentences of Punjabi language that shows an overall precision as 93.30, recall as 97.32 and F-measure as 95.25. Compound sentence simplification approach used in this work can be further used for generating paraphrasing of Punjabi language. Further this compound sentence simplification work can be extended to complex sentence simplification work. Similar grammar checker technique can be implemented to other languages that are similar to Punjabi language like Hindi. Further this work can be extended for grammar checking of complex sentences.

References
Sentence Reduction for Syntactic Analysis of Compound Sentences in Punjabi Language


Integrating Artificial Intelligence and Assistive Technology, pp. 7–10.


