Clause Boundary Identification for non-restricted type complex Sentences in Punjabi Language

Sanjeev Kumar Sharma
Department of Computer Science and Applications
DAV University, Jalandhar (Punjab), India
Sanju3916@rediffmail.com

ABSTRACT

Clause boundary identification for non-restricted type complex sentences in Punjabi language is one of the basic requirements for processing of complex sentences. For grammar checking of complex sentences, it is necessary to identify the structure of clauses present in the sentence. Once the sentence is identified as complex sentence, the next step is to identify its pattern i.e. position of dependent and independent clauses. After identification of patterns, various clauses present in the sentence are extracted. In this paper, author has explored a technique to identify and extract the clause boundaries of various clauses present in non-restrictive type of complex sentence. This study will be helpful in simplification of complex sentences. Also this study will be helpful in developing other Natural Language Processing (NLP) applications like paraphrasing, Improving Machine translation system and grammar checking of complex sentences.

KEYWORDS

NLP, Complex Sentences, Non-restrictive, Grammar checking.

INTRODUCTION

A clause is a largest unit of the sentence that contains a predicate and an explicit or implied subject. A sentence may have any number of clauses. Clause boundary identification means to split the sentence into clauses by identifying the starting and the ending position of the clause. The task of clause boundary identification is not only detecting a non-recursive phrase of the sentence, rather it is a three step process: identifying start of clause, identifying end of clause and finding complete clause (Sang and Dejean, 2001). Consider the following example:

Punjabi:ਮੀਂਹ ਪੈ ਰਿਹਾ ਸੀ ਤੇ ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ।

Tranlitrated in Roman: (mīṃh pai rihā sī lōk bhij jrahē san)

English Translation: It was raining and people were on spree

In the above sentence, there are two clauses; one is ਮੀਂਹ ਪੈ ਰਿਹਾ ਸੀ (mīṃh pai rihā sī) and second is ਲੋਕ ਭਿੱਜ ਰਹੇ ਸਨ (lōk bhijj rahē san). Both these clauses are joined by conjunction (ਤੇ). In clause identification, problem to be identified is the start and end position of both the clauses. As shown in figure 1, s1 and e1 represent start and end point of the first clause (clause 1) and s2 and e2 represent start and end point of the second clause.

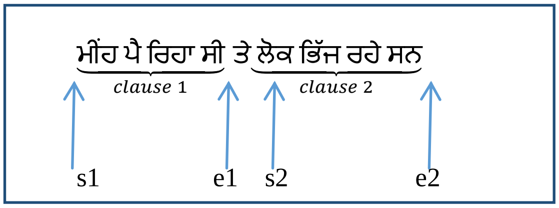


Figure 1 :Marking starting and end points of clauses in a sentence

CLAUSE

All the phrases (postpositional, nominal, adjectival, verb) combine to constitute the clauses. If a sentence is highest unit, then clause is the second highest unit of the sentence. These are composed of phrases. A clause may contain any number of phrases. Verb phrase is the essential component of every clause. Even a single clause constituted by a verb phrase can construct sentence. There is no need of any other element in the sentence. Clauses can be classified on the basis of this verb phrase; the verb phrase is the essential element of every clause. There are two types of clauses in Punjabi language one is independent and other is the dependent clause. The clause having the finite verb phrase is called independent clause and the other having non-finite verb phrase is called dependent clause. In the following section, an overview of these two types of clauses is given in the next section.

INDEPENDENT CLAUSE

Independent clause is essential part of all types of sentences. The structure of independent clause is same as that of simple sentence. A clause is called independent clause if it can exist in-dependently as a complete sentence. Verb phrase is the essential part of the independent clause. The independent clause contains exactly one verb phrase (Puar, 1990) along with other elements of the clause. Other than verb phrase, an independent clause may contain one or more noun phrase, adjective phrase, adverb phrase etc. as other elements (Bray,2008). The verb phrase present in the independent clause is finite verb phrase. In compound sentences, these independent clauses are joined by coordinate conjunctions. In complex sentences, an independent clause and dependent clause are joined using subordinate conjunctions. Independent clauses are used to give more identity in grammar checking system.

DEPENDENT CLAUSE:

Subordinate verb phrase as one of the essential element of type of independent clause. These clauses convey an incomplete thought and hence, cannot constitute a sentence without the help of other clauses. To constitute a sentence it requires at least one independent clause. These clauses participate in the construction of complex sentences. Dependent clause also contains verb phrase as one of its essential element. Like independent clause, this verb phrase can be finite or nonfinite.

NON-RESTRICTIVE DEPENDENT CLAUSE

Nonrestrictive clauses provide some information about the preceding subject and conjunctions starting with 'ਜ' character are used to provide relevant information of the clause. Comma is used to separate different clauses. Consider the following examples:

Sr. No.	Example
1	Punjabi: ਥੋੜ੍ਹੀਆਂ ਜਿਹੀਆਂ ਚੀਜ਼ਾਂ,ਜੋ ਅੱਗੇ ਪਿਛੇ ਪਈਆਂ ਹਨ,ਬੜੀਆਂ ਹੀ ਸੋਹਣੀਆਂ ਤੇ ਮੂੰਹੋਂ ਬੋਲਦੀਆਂ ਸਨ। Transliteration: (thōṛhīāṃ jihīāṃ cīzāṃ, jō aggē pichē paīāṃ han, baṛīāṃ hī sōhṇīāṃ tē mūṃhōṃ bōldīāṃ san.) Translation: A few objects, which have been placed together, look lively.
2	ਮੇਰੇ ਮਾਮਾ ਜੀ, ਜੋ ਕਿ ਜੱਜ ਹਨ, ਅੱਜ ਕੱਲ੍ਹ ਪਟਿਆਲੇ ਰਹਿੰਦੇ ਹਨ । (mērē māmā jī, jō ki jajj han, ajj kallh paṭiālē rahindē han.)

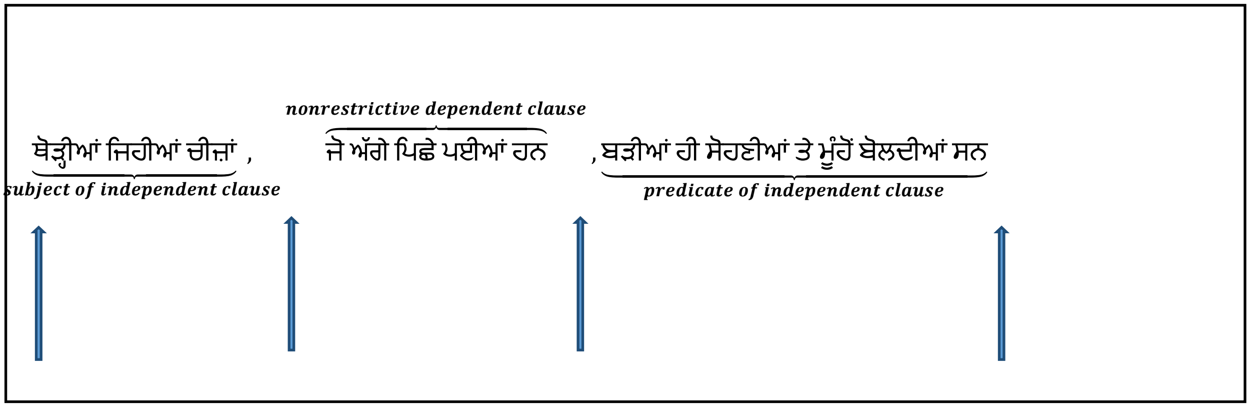
Like restrictive clause, this clause is also embedded between the subject and the predicate of the independent clause. Therefore, like restrictive clause, nonrestrictive clause also splits the sentence in three parts. One part containing start and end of the subject of independent clause, second part containing the start and end of the nonrestrictive clause and the third part containing the start and end of the predicate of independent clause. Consider the sentence 1 from above table:

Punjbi:ਥੋੜ੍ਹੀਆਂ ਜਿਹੀਆਂ ਚੀਜ਼ਾਂ,ਜੋ ਅੱਗੇ ਪਿਛੇ ਪਈਆਂ ਹਨ,ਬੜੀਆਂ ਹੀ ਸੋਹਣੀਆਂ ਤੇ ਮੂੰਹੋਂ ਬੋਲਦੀਆਂ ਸਨ।

Transliteration: (thōṛhīāṃ jihīāṃ cīzāṃ, jō aggē pichē paīāṃ han, baṛīāṃ hī sōhṇīāṃ tē mūṃhōṃ bōldīāṃ san.)

English: A few objects, which have been placed together, look lively.

As shown in the above complex sentence, the nonrestrictive clause "ਜੋ ਅੱਗੇ ਪਿਛੇ ਪਈਆਂ ਹਨ" (jō aggē pichē paīāṃ han) is embedded between the subject "ਥੋੜ੍ਹੀਆਂ ਜਿਹੀਆਂ ਚੀਜ਼ਾਂ" (thōṛhīāṃ jihīāṃ cīzāṃ) and the predicate "ਬੜੀਆਂ ਹੀ ਸੋਹਣੀਆਂ ਤੇ ਮੂੰਹੋਂ ਬੋਲਦੀਆਂ ਸਨ" (baṛīāṃ hī sōhṇīāṃ tē mūṃhōṃ bōldīāṃ san) of the independent clause. In this way, nonrestrictive clause splits the sentence into three parts. The first part contains the subject of independent clause, the second part contains the nonrestrictive clause and the third part contains the predicate of independent clause. The first part of the sentence starts with the first word of the sentence i.e. ਥੋੜ੍ਹੀਆਂ (thōṛhīāṃ) and ends with the word having comma and just previous to J-conjunction i.e. with the word ਚੀਜ਼ਾਂ (cīzāṃ). Similarly, the second part of the sentence starts with the J-conjunction that appears just after the first comma and ends with the word just before the second comma i.e. ਹਨ (han), and the third part starts with a word just after the second comma i.e. ਬੜੀਆਂ (baṛīāṃ) and ends with the last word of the sentence. These three separate parts of the sentence can be represented as:


nonrestrictive dependent clause
ਥੋੜ੍ਹੀਆਂ ਜਿਹੀਆਂ ਚੀਜ਼ਾਂ ,
ਜੋ ਅੱਗੇ ਪਿਛੇ ਪਈਆਂ ਹਨ
,ਬੜੀਆਂ ਹੀ ਸੋਹਣੀਆਂ ਤੇ ਮੂੰਹੋਂ ਬੋਲਦੀਆਂ ਸਨ
subject of independent clause
predicate of independent clause

Clause Boundaries

Figure 2: Marking clause boundaries in complex sentence having non-restrictive type clause

As shown above in figure 2, the start and end point of the nonrestrictive clause can be identified with J-conjunction and comma (,) respectively. J-conjunction separates the subject of independent clause from nonrestrictive clause and comma separates the nonrestrictive clause from the predicate of dependent clause.

METHODOLOGY USED

In this research work, the syntactic cue and morphological information have been used for clause boundary identification. The morphological information used includes suffix information of non-finite verb and even part of speech tag at some places. Syntactic cue includes presence of conjunction or comma. Different morphological and syntactic cues have been used for different type of dependent clauses. For example, suffix information of non-finite verb has been used for marking clause boundaries in complex sentences containing predicate bound type of clauses; subordinate conjunctions are used to mark the clause boundaries in complex sentences containing non-predicate bound type of clauses. In the following section, detailed description about identification of dependent clauses has been provided.

Algorithm used: Clause boundary identification of non-predicate bound non-restrictive type.

Input: Annotated Punjabi sentence

Database used: List of J-conjunction

Output: Punjabi sentence with marked clause boundaries.

```
1. Tokenize the input sentence.
2. Mark the first word of the sentence as beginning of subject of independent clause.
3. Repeat step 4 for all the tokens of the sentence.
4. If the current word is comma and the next word is J - conjunction then go to step 5.
5. Mark the previous word as end of subject of independent clause and go to step 6.
6. Mark the next word i.e. J-conjunction as beginning of dependent clause.
7. If current word is auxiliary verb with comma and it is not the last word of the sentence then mark this word
   as end of dependent clause and next word as beginning of predicate of independent clause.
8. Mark the last word as end of predicate of independent clause.
```

Flow chart representing above mentioned algorithm is shown in figure 1. An example to illustrate the working of flowchart/algorithm is also provided with this flow chart.

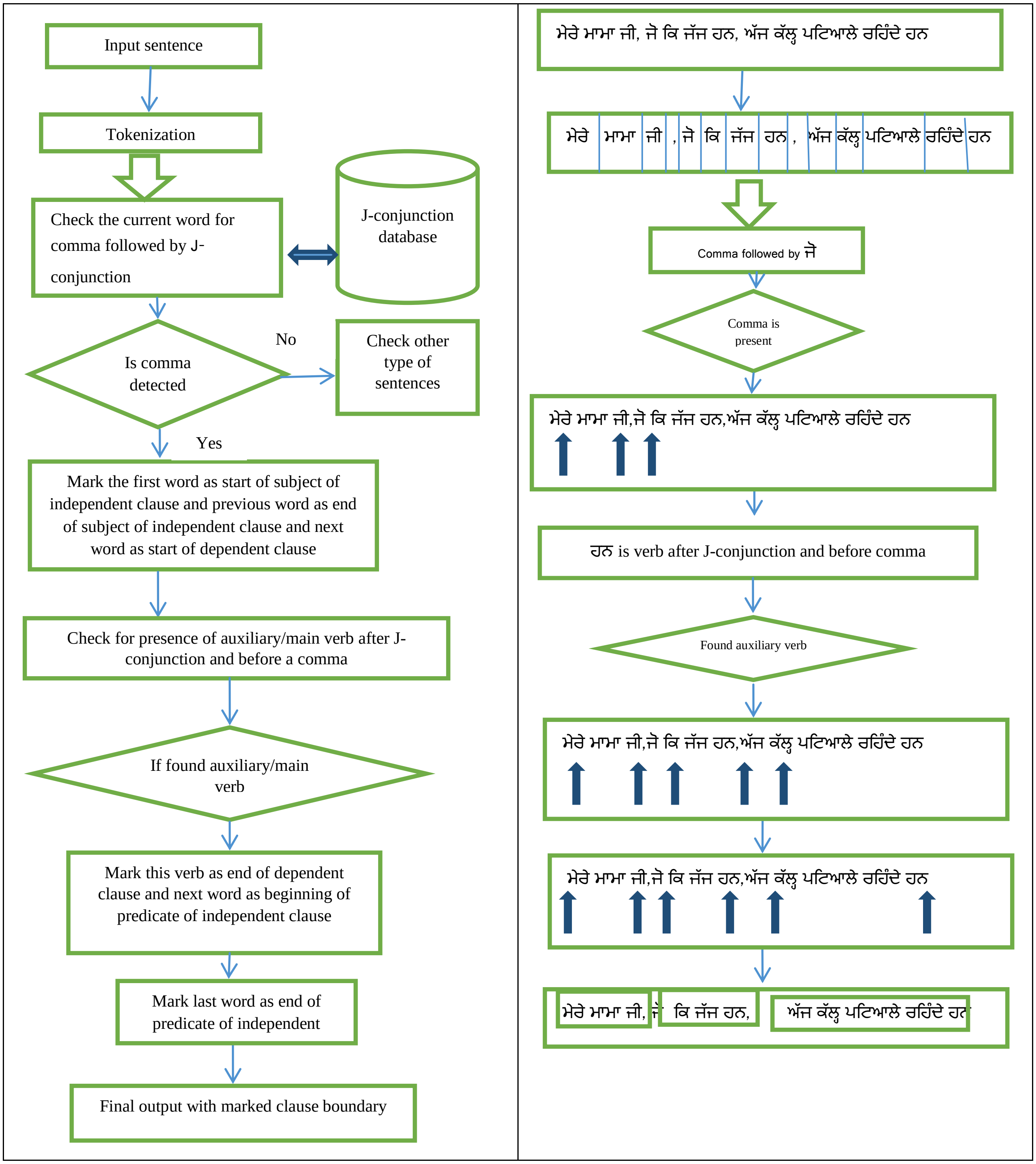

Input sentence
Tokenization
Check the current word for comma followed by J-conjunction
J-conjunction database
Is comma detected
No
Check other type of sentences
Yes
Mark the first word as start of subject of independent clause and previous word as end of subject of independent clause and next word as start of dependent clause
Check for presence of auxiliary/main verb after J-conjunction and before a comma
If found auxiliary/main verb
Mark this verb as end of dependent clause and next word as beginning of predicate of independent clause
Mark last word as end of predicate of independent
Final output with marked clause boundary
ਮੇਰੇ ਮਾਮਾ ਜੀ, ਜੋ ਕਿ ਜੱਜ ਹਨ, ਅੱਜ ਕੱਲ੍ਹ ਪਟਿਆਲੇ ਰਹਿੰਦੇ ਹਨ
ਮੇਰੇ | ਮਾਮਾ | ਜੀ | , | ਜੋ | ਕਿ | ਜੱਜ | ਹਨ | , | ਅੱਜ | ਕੱਲ੍ਹ | ਪਟਿਆਲੇ | ਰਹਿੰਦੇ | ਹਨ
Comma followed by ਜੋ
Comma is present
ਮੇਰੇ ਮਾਮਾ ਜੀ,ਜੋ ਕਿ ਜੱਜ ਹਨ,ਅੱਜ ਕੱਲ੍ਹ ਪਟਿਆਲੇ ਰਹਿੰਦੇ ਹਨ
ਹਨ is verb after J-conjunction and before comma
Found auxiliary verb
ਮੇਰੇ ਮਾਮਾ ਜੀ,ਜੋ ਕਿ ਜੱਜ ਹਨ,ਅੱਜ ਕੱਲ੍ਹ ਪਟਿਆਲੇ ਰਹਿੰਦੇ ਹਨ
ਮੇਰੇ ਮਾਮਾ ਜੀ,ਜੋ ਕਿ ਜੱਜ ਹਨ,ਅੱਜ ਕੱਲ੍ਹ ਪਟਿਆਲੇ ਰਹਿੰਦੇ ਹਨ
ਮੇਰੇ ਮਾਮਾ ਜੀ, ਜੋ
ਕਿ ਜੱਜ ਹਨ,
ਅੱਜ ਕੱਲ੍ਹ ਪਟਿਆਲੇ ਰਹਿੰਦੇ ਹਨ

Fig 1: Flow chart and working example to mark clause boundaries in nonrestrictive type complex sentences

CONCLUSION

This paper concerns the grammar checking of complex sentences in various agreement errors within independent clauses in case of complex sentences and between dependent and independent clauses in case of complex sentences. For grammar checking of complex sentences, it is necessary to identify the structure of Clause Boundary for Non-Restrictive type of Punjabi language complex sentences. In this paper we have explored the different types of sentences present in Punjabi language. The structure of complex sentences can be identified on the basis of number of clauses and types of clauses present in them. We have also proposed an algorithm for identification of simple, compound and complex sentences. This study will be helpful in identifying and separating the complex sentences from Punjabi language. We have also proposed an algorithm for identification of simple, compound and complex sentences. Also this study will be helpful in developing other Natural Language Processing (NLP) applications like converting a complex sentence in simple sentences, grammar checking of complex sentences.

REFERENCES

[1]. Sobha, L. D., & Lakshmi, S. Malayalam. 2013. Clause Boundary Identifier: Annotation and Evaluation. WSSANLP-2013, p. 83.

[2]. Kaur, N., Garg, K., Sharma, Sanjeev. Kumar. 2013. Identification and Separation of Complex Sentences from Punjabi Language. International Journal of Computer Applications, 69(13), pp. 21-24.

[3]. Sharma, Sanjeev Kumar ‘Assigning the Correct Word Class to Punjabi Unknown Words using CRF’ International Journal of Computer Applications (0975 – 8887) Volume 142 – No.2, May2016

[4]. Brill, E. 1992. A simple rule-based part of speech tagger. In Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics. pp. 112-116

[5]. Brill, E. 1993. Automatic grammar induction and parsing free text: A transformation-based approach. In Proceedings of the workshop on Human Language Technology. Association for Computational Linguistics. pp. 237242

[6]. Kasbon, R., Amran, N., Mazlan, E., &Mahamad, S. 2011.Malay language sentence checker. World Appl. Sci. J. (Special Issue on Computer Applications and Knowledge Management),12, pp. 19-25.

[7]. Kubon V., & Platek, M. 1994. A grammar based approach to a grammar checking of free word order languages. In Proceedings of the 15th conference on Computational linguistics-Volume 2. Association for Computational Linguistics. pp. 906-910

[8]. Leffa, V. J. 1998. Clause processing in complex sentences. In Proceedings of the First International Conference on Language Resources and Evaluation Vol. 1, pp. 937-943.

[9]. Narula, R., & Sharma, S. K. 2014. Identification and Separation of Simple, Compound and Complex Sentences in Punjabi Language. International Journal of Computer Applications& Information Technology. Vol. 6, Issue II Aug-September 2014.

[10]. Orasan, C. 2000. A hybrid method for clause splitting in unrestricted English texts. Proceedings of ACIDCA' 2000

[11]. Parveen, D., Sanyal, R., & Ansari, A. 2011. Clause Boundary Identification using Classifier and Clause Markers in Urdu Language. Polibits Research Journal on Computer Science, 43,pp. 61-65.