



(UN) IDENTIFIED LEXEMES: A FIRST GLANCE AT PUNJABI MORPHOLOGICAL ANALYZER

Harjit Singh

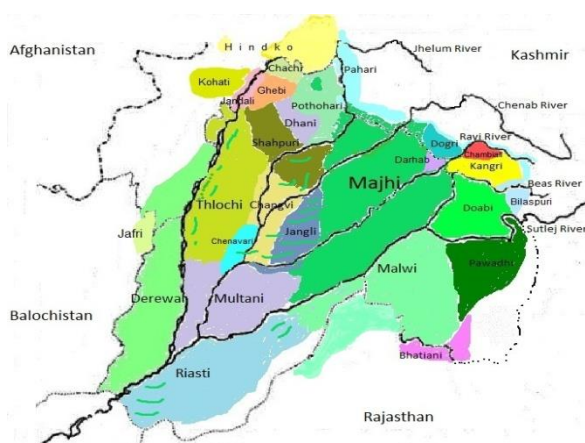
Dept. of Linguistics and CSTLs
Indira Gandhi National Tribal University
Amarkantak, India

Abstract: The paper presents a small survey on Punjabi morphological analyzer and generator, which does not give any morphological response to lexemes of Malawi dialect of Punjabi. This analyzer works well with the standard variety only. However, it fails to provide an answer when the input is colloquial. The random choice of approximately, (42 words/lexemes, e.g. ਕਾਤੇ/ 'why', ਉੱਈ/ 'thus', ਭੁੰਨਹੋਣਾ/ 'Unbearable', ਖੁੰਨੀ/ 'eating', ਅੜੀ/ 'stubborn', ਭਨਾਲੀ/ 'breaking', ਮੰਨਲਾ/ 'suppose', ਧਿਨੀ/ 'annoying' etc.) is not trained either for the analyzer nor the generator. They are unidentified and unrecognized. In this regard, the schemes of SGMA (similar guidelines for morphological analyzer) or SMGMA (similar but modified guidelines for morphological analyzer) could adopt for booming the performance of Punjabi morphological analyzer and widening its scope in the future.

Keywords: Un(identified) lexemes, Punjabi Morphological Analyzer, SGMA

I. INTRODUCTION

Punjabi is a modern Indo-Aryan language with a beautiful story of five rivers (the Jhelum, the Ravi, the Chenab, the Vyas and the Satluj). It covers Eastern Punjab (India), Western Punjab (Pakistan), East Africa, Canada, UK and USA, etc. Gurumukhi is the script, and we usually write it from left to right. It has a (subject-object-verb) word order. However, there are four significant dialects (e.g. Majhi, Malawi, Doabi and Puadhi) of Punjabi, and there is less clarity regarding a standard variety of Punjabi. Majhi considers a standard type, and its area is Amritsar, Tarn Taran Sahib, Pathankot and Gurdaspur in the Eastern Punjab (India) [1]. See map 1.



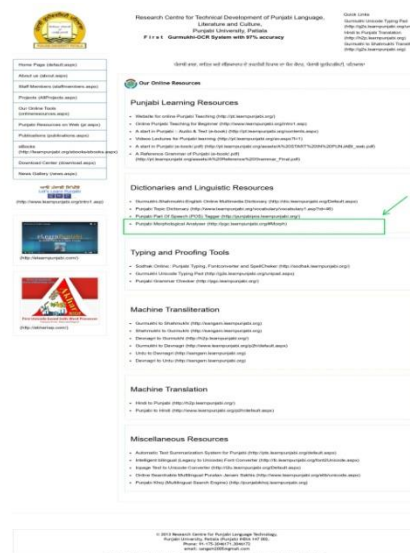
Map 1: Punjabi dialect via Wikipedia

https://upload.wikimedia.org/wikipedia/commons/6/6b/Dialects_of_Punjabi.jpg

Above Map 1 mainly shows four major Punjabi dialects (Majhi, Malawi, Doabi and Puadhi). The variations related to phonetics and grammars are available among all these dialects.

After going through the basic details of Punjabi dialects, we are more interested in studying the specific area of Punjabi

computational linguistics rather than theoretical linguistics. In this regard, we discuss the 'Research Centre for Technical Development of Punjabi Language', which took place in February 2004 at Punjabi University, Patiala (Punjab). It has rich stock of the following online Punjabi resources.



Picture 1: Online Punjabi Resource via Research Centre for Technical Development of Punjabi Language, <http://www.learnpunjabi.org/default.aspx>

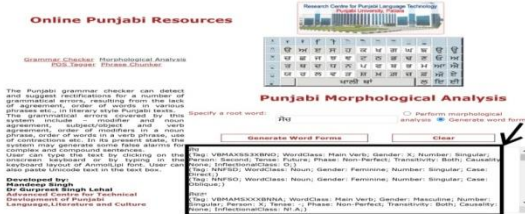
Above Picture 1 provides a list of online Punjabi resources that the 'Research Centre has developed for Technical Development of Punjabi Language' (Punjabi University, Patiala). The green rectangle () and arrow represent that we pick up only 'Punjabi Morphological Analyzer' instead of remaining other online resources.

For Punjabi, 'Morphological Analyzer' works to analyse the submitted word classes root and grammatical information. It provides gender, number, and person and case information related to noun word classes and gives tense, aspect and modal information related to verb word classes only in a standard variety of Punjabi. While the 'Morphological Generator' does

the opposite tasks, it receives root and grammatical information and then produces the word forms accordingly. Consider few instances.



Picture 2: Perform morphological analysis via Research Centre for Technical Development of Punjabi Language, <http://pgc.learnpunjabi.org/#Morph>



Picture 3: Generate word forms via Research Centre for Technical Development of Punjabi Language, <http://pgc.learnpunjabi.org/#Morph>

The yellow colour () identifies the Punjabi lexemes such as 'ਮੈਚ' /'Thinking' in picture 1. The following description shows the morphological analysis of this lexeme.

Root = ਮੈਚ

(Tag: VBMASXXXBNNA; Word Class: Main Verb; Gender: X; Number: Singular; Person: Second; Tense: Future; Phase: Non-Perfect; Transitivity: Both; Causality: None; Inflectional Class: O ;)

(Tag: NNFSO; Word Class: Noun; Gender: Feminine; Number: Singular; Case: Direct ;)

(Tag: NNFSO; Word Class: Noun; Gender: Feminine; Number: Singular; Case: Oblique ;)

The black colour () recognizes the morphological generator that applies to the same lexeme 'ਮੈਚ' /'Thinking' and produces different word forms. Look at the following results.

ਮੈਚ

(Tag: VBMASXXXBNNA; Word Class: Main Verb; Gender: X; Number: Singular; Person: Second; Tense: Future; Phase: Non-Perfect; Transitivity: Both; Causality: None; Inflectional Class: O ;)

(Tag: NNFSO; Word Class: Noun; Gender: Feminine; Number: Singular; Case: Direct ;)

(Tag: NNFSO; Word Class: Noun; Gender: Feminine; Number: Singular; Case: Oblique ;)

ਮੈਚਾ

(Tag: VBMASXXXBNNA; Word Class: Main Verb; Gender: Masculine; Number: Singular; Person: X; Tense: -; Phase: Non-Perfect; Transitivity: Both; Causality: None; Inflectional Class: N! A ;)

We see that ਮੈਚ/Thinking with this tag: VBMASXXXBNNA is a primary root verb. On the other hand, ਮੈਚਾ/Thinking comes

with another tag like VBMASXXXBNNA. But, many colloquial lexemes in Punjabi are unresolved morphologically. We capture those unanalyzed or unknown cases and suggest a procedure to incorporate them with ਮੈਚ/Thinking like lexemes.

Briefly, we have a total of six parts in this paper. In the first part, we discuss few virtual instances of lexemes in a standard variety of Punjabi via Punjabi Morphological Analyzer. At the same time, we see previous literature related to the morphological analyzer in other languages in the second part. Next, we elaborate on the aims and discuss the scope of the paper in the third part. In the fourth part, we mainly present 'unknown' or 'unrecognized' cases in the colloquial Punjabi. The fifth part discusses the procedure to resolve 'unknown' or 'unrecognized' lexemes. In the sixth part, we discuss the conclusions and future of the works.

II. RELATED WORKS

In 2008, a paradigm approach was applied to Hindi words to study their morphological structure [2]. Another language Kokborok (which belongs to Tripura and nearby areas of Bangladesh), has shown 80% correct results when the morphological analyzer trained around 56732 Kokborok words [3]. For Malayalam morphological analyzer, the Memory-Based Language Processing (MBLP) algorithm was adopted. The TiMBL (Tilburg Memory-based Learner) was adopted in this analyzer after annotation for training purposes [4]. Later, the Multi-Task Deep Morphological analyzer (MT-DMA) also morphologically trained Hindi and Urdu [5].

On the other hand, some scholars like Sivan have tried to compare the existing (corpus-based, rule-based, stemmer based, paradigm based, finite state automata and transducer-based) morphological analyzers and notice the reliability for Malayalam [6]. But it points out that the paradigm approach of morphological analyzer increases 84.50%, 81.50% and 80.50% accuracy for nouns, verbs and adjectives, respectively [7].

III. OBJECTIVES AND SCOPE

- To study only Punjabi Morphological Analyzer such as (online Punjabi resources)
 - To find out 'unknown lexemes' related to colloquial Punjabi
 - To suggest the procedure for resolving the unknown lexemes
- To conclude the work and plans in future

IV. UNKNOWN/UNIDENTIFIED LEXEMES

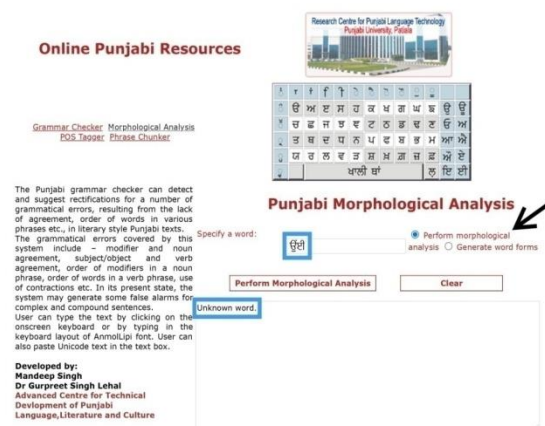
There are two parts (morphological analysis and generate word forms) in the Punjabi morphological analyzer. Earlier, we notice that it works only for the standard Punjabi. After input of the lexeme, it separates the root and then provides the grammatical information, as seen in Picture 2. However, we argue that it does not identify any lexeme that belongs to a colloquial variety (e.g. Malawi) of Punjabi. Moreover, it also does not provide any answer if the input is so standard (like ਪਟਨਾ/Patna, ਚੱਟਨਾ/Stop, ਡੱਟਨਾ/Stable, ਰੱਟਨਾ/Memorization etc.) lexemes from the standard Punjabi. The following Table 1 shows the list of lexemes in Malawi that are possibly unknown/unidentified lexemes.

Table 1: List of unknown lexemes in Malawi (Punjabi)

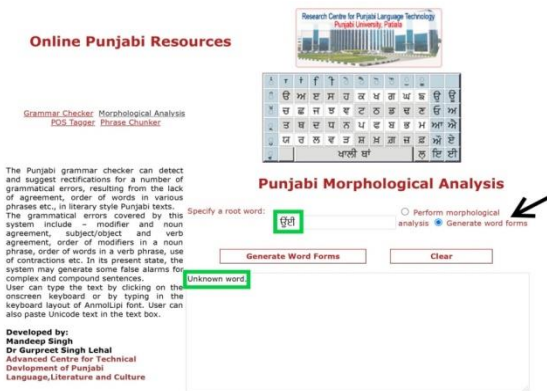
Sr. No	Lexemes of Malwi	Root	Word class	English
1	ਕਾਤੋਂ	ਕਾਤੋਂ	Adverb	Why
2	ਉਂਦੀ	ਉਂਦੀ	Adverb	Thus
3	ਭੁੰਨਹੋਣਾ	ਭੁੰਨ+ਹੋਣਾ	Adjective	Unfavourable
4	ਬੁੰਨੀ	ਬੁੰਨ+ਈ	Noun/Adjective	Eating
5	ਅੜੀ	ਅੜ+ਈ	Adjective	Stubborn
6	ਭਨਾਲੀ	ਭਨਾਲੀ	Verb	Breaking
7	ਦਈ	ਦਈ	Adjective	Giving
8	ਅੜਬ	ਅੜ+ਬ	Adjective	Stubborn
9	ਮੰਨਲਾ	ਮੰਨਲਾ	Verb	Suppose
10	ਜਾਈਦਾ	ਜਾਈਦਾ	Noun	Going
11	ਰੁੜ	ਰੁੜ	Adjective	Falling
12	ਗਏਆ	ਗਏਆ	Verb	Went
13	ਛੱਡੀ	ਛੱਡ+ਈ	Adjective	Giving
14	ਧੱਕਮ	ਧੱਕਮ	Verb/Noun	Push
15	ਪੋਲਿਆ	ਪੋਲਿਆ	Noun	Polya
16	ਤਾਵਾ	ਤਾਵਾ	Noun	?
17	ਖਿਜੀ	ਖਿਜੀ	Adjective	Annoying
18	ਮਗਜ	ਮਗਜ	Noun	Brain
19	ਚਮਲਾਈ	ਚਮਲਾਈ	Adjective	Excited
20	ਰਿੰਗੀ	ਰਿੰਗੀ	Adjective	Crying
21	ਨੀ	ਨੀ	?	?
22	ਬੁੱਥੀ	ਬੁੱਥੀ	Noun	Face
23	ਚੱਜਦੀ	ਚੱਜ+ਦੀ	Adverb/Adjective	Right
24	ਬੁੱਗੇ	ਬੁੱਗੇ	Noun	Child
25	ਰੜਿਆ	ਰੜ+ਇਆ	Adjective	Angry
26	ਕੁਰੇ	ਕੁਰੇ	Noun	Title
27	ਸਿਹਾਂ	ਸਿਹਾਂ	Noun	Title
28	ਗਹਾਂ	ਗਹਾਂ	Adverb	Ahead
29	ਨਨਦਾਂ	ਨਨਦਾਂ	Noun	Sister-in-law

30	ਸਖਾਏ	ਸਖਾਏ	Verb	Teach
31	ਛਾਬਾਛੀ	ਛਾਬਾਛ+ਈ	Adjective	Well done
32	ਮੈਤੇ	ਮੈਤੇ	Pronoun	First person
33	ਕੋਲੇ	ਕੋਲੇ	Post-position	Along
34	ਐਕਨ	ਐਕਨ	Adverb	Likewise
35	ਨਿਆਨਾ	ਨਿਆਨਾ	Noun	Small child
36	ਘੁਰੀ	ਘੁਰ+ਈ	Adjective	Scary
37	ਟਿਕਨਾ	ਟਿਕ+ਨਾ	Noun	Sitting
38	ਬੁੱਜ	ਬੁੱਜ	Adjective	Foolish
39	ਨੀਚੇ	ਨੀਚ+ਓ	Adverb/Adjective	Downstairs
40	ਓਏ	ਓਏ	Noun	Hello
41	ਬਚੀ	ਬਚੀ	Noun	Kid
42	ਨਾਨੂੰ	ਨਾ+ਨੂੰ	Negation	Not

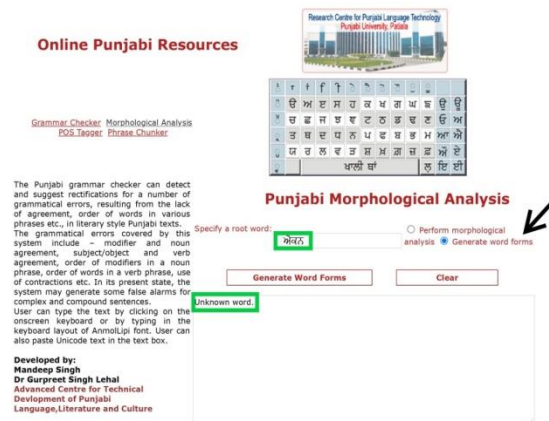
The total 42 instances in Table 1 are highly unknown and unrecognized cases for Punjabi Morphological Analyzer. Not a single instance has yet been resolved [8]. The following cases are like as:



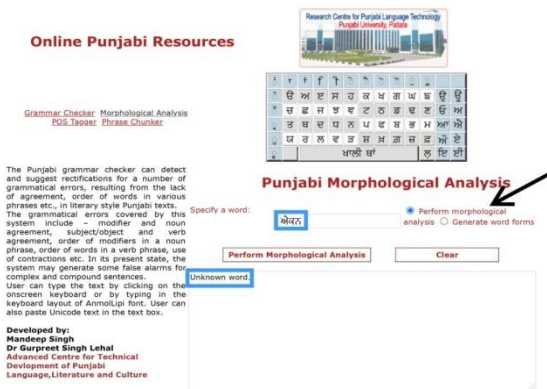
Picture 4: Unknown case of ਉਂਦੀ/Thus (morphological analysis) via Research Centre for Technical Development of Punjabi Language, <http://pgc.leampunjabi.org/#Morph>



Picture 5: Unknown case of ਊੜੀ/Thus (Generate word forms) via Research Centre for Technical Development of Punjabi Language, <http://pgc.leampunjabi.org/#Morph>



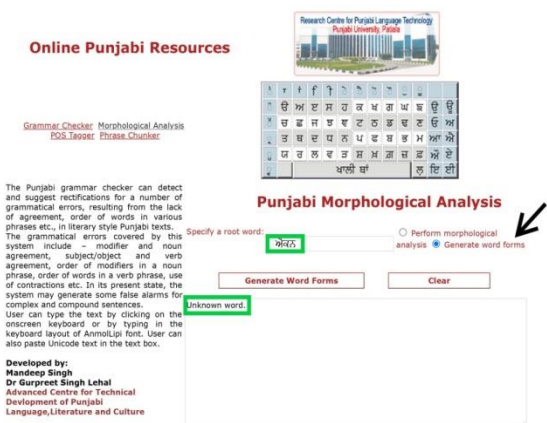
Picture 7: Unknown case of ਐਕਨ/Likewise (Generate word forms) via Research Centre for Technical Development of Punjabi Language, <http://pgc.leampunjabi.org/#Morph>



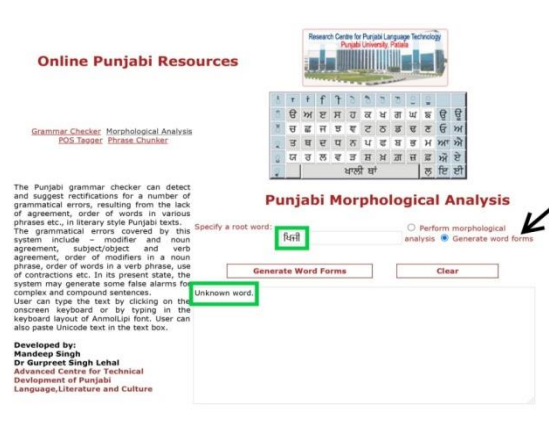
Picture 6: Unknown case of ਐਕਨ/Likewise (morphological analysis) via Research Centre for Technical Development of Punjabi Language, <http://pgc.leampunjabi.org/#Morph>



Picture 8: Unknown case of ਫਿਨੀ/Annoying (morphological analysis) via Research Centre for Technical Development of Punjabi Language, <http://pgc.leampunjabi.org/#Morph>



Picture 6: Unknown case of ਐਕਨ/Likewise (morphological analysis) via Research Centre for Technical Development of Punjabi Language, <http://pgc.leampunjabi.org/#Morph>



Picture 9: Unknown case of ਫਿਨੀ/Annoying (Generate word forms) via Research Centre for Technical Development of Punjabi Language, <http://pgc.leampunjabi.org/#Morph>

From (Picture 4 to Picture 9) related to ਊੜੀ/Thus, ਐਕਨ/Likewise, and ਫਿਨੀ/Annoying lexemes are absolutely ‘unknown’/ ‘unidentified’ cases of Malawi dialect. After performing the morphological analysis, the light blue rectangle () shows ‘no results’ available for such lexemes. Similarly, the green rectangle () also indicates that there is no chance to generate any word form based on these lexemes.

V. PROPOSAL FOR UNKNOWN LEXEMES

In practice, unknown lexemes are nothing but a collection of the Punjabi dialect (e.g. Malawi). Phonetics is a significant factor of the total 42 lexemes, and they are orthographically finding less attractive. Even it is a challenging task to define them linguistically also. So far, we are not sure whether any work (in on-line mode) is available related to the dialect or not. But regarding the unknown lexemes, we approach the following figure 1.

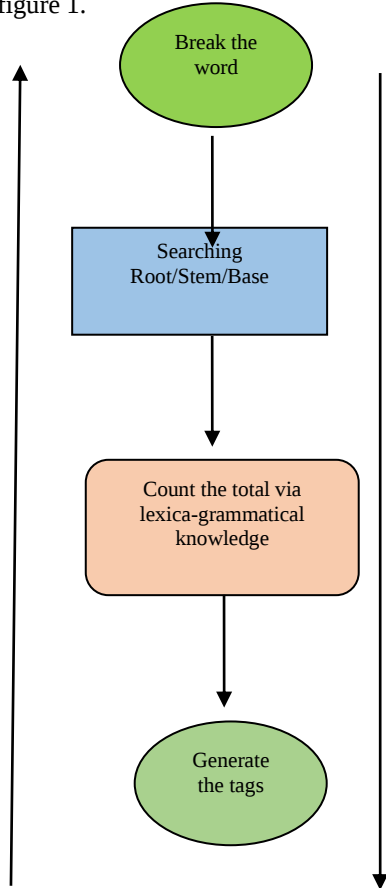
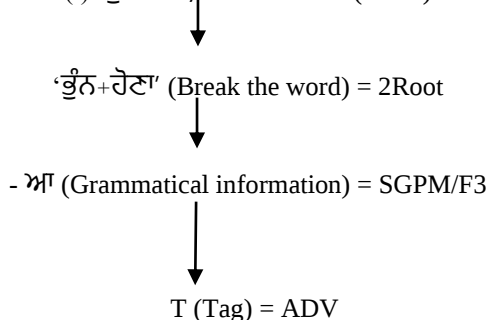


Fig 1. Steps of filtering the unknown lexemes

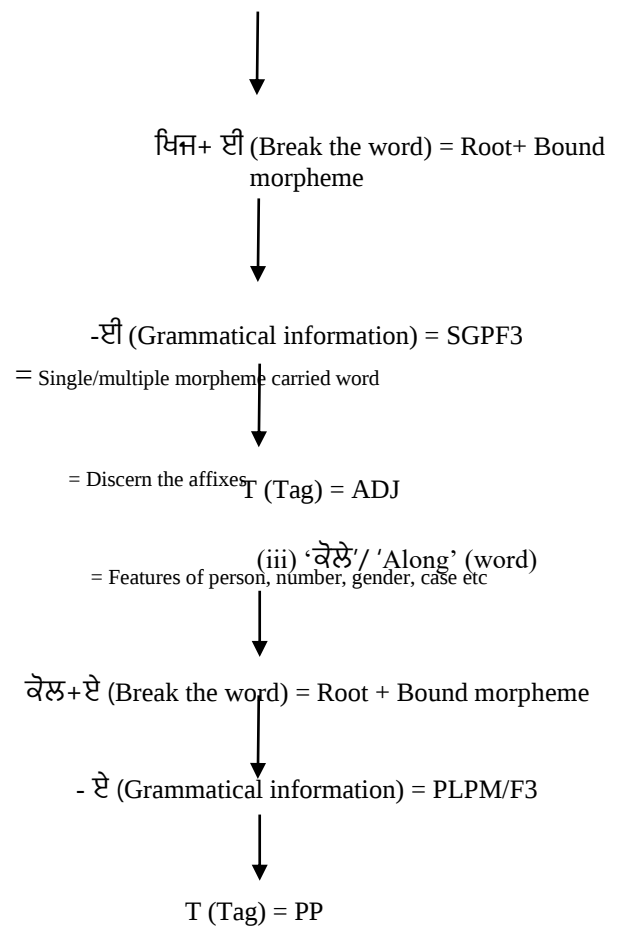
Word/lexeme such as ‘ਭੁੰਨਹੋਣਾ’ / ‘Unfavorable’ breaks down into ‘ਭੁੰਨ+ਹੋਣਾ’ two root parts that mean both are independent forms. The ending – ਆ of ਹੋਣਾ represents singular, masculine/feminine, third-person pronoun etc. information. When they come all together, they function as a compound. Thus, we could quickly analyze the remaining lexemes also. For more details, please go to Table 1 and see the root column carefully.

After the analysis, our next task is to provide a tag to each word class accordingly. The whole procedure defines as,

(i) ‘ਭੁੰਨਹੋਣਾ’ / ‘Unbearable’ (Word)



(ii) ‘ਖਿਜੀ’ / ‘Annoying’ (word)



ਕੋਲ+ਏ (Break the word) = Root + Bound morpheme

- ਏ (Grammatical information) = PLPM/F3

T (Tag) = PP

Following the description, we can resolve almost all the unknown cases at this initial level. Then, SGMA (similar guidelines for morphological analyzer) might be adopted for tagging purposes. See below some instances.

‘ਭੁੰਨਹੋਣਾ’ / ‘Unbearable’ [Compound word]

Tag: ADVXSS3P; Word Class: Adverb, Gender: Male, Number: Singular, Person: Third, Tense: Past, Phase: Perfect etc.

‘ਖਿਜੀ’ / ‘Annoying’ [Complex word]

Tag: ADJXM/FS3P; Word Class: Adjective, Gender: Male/Female, Number: Singular, Person: Third, Tense: Past, Phase: Perfect etc.

‘ਕੋਲੇ’ / ‘Along’ [Complex word]

Tag: PPXM/FS3P; Word class: Post-position, Gender: Male/Female, Number: Singular, Person: Third, Tense: Present, Phase: Perfect etc.

We have successfully but manually annotated the above instances. We used the tag-set of the standard Punjabi. In this way, we suggest that the existing tag-set of ‘Punjabi Morphological Analyzer’ could apply to all 42 unknown lexemes in future.

VI. CONCLUSION

Like other languages, Punjabi is also rich with many online computational resources. ‘Punjabi morphological analyzer’ is an essential resource freely available on the web (<http://pgc.learnpunjabi.org/#Morph>) to analyze the words morphologically. But it is not trained to diagnose the dialectal

vocabulary so far. Malawi is a good example here, and its total of 42 words is unknown. In this regard, we suggest an (SGMA) scheme to minimize the rate of unknown results in the future.

VII. REFERENCES

- [1] Bhatia, K.T. (2013). Punjabi: Cognitive-Descriptive Grammar. Routledge: Taylor and Francis Group. New York, USA.
- [2] Debbarma, K. et al. (2012). Morphological Analyzer for Kokborok. In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP). COLING, pages 41-52, Mumbai, India.
- [3] Goyal, V. and Lehal, S.G. (2008). Hindi Morphological Analyzer and Generator. In Proceedings of the 1st International Conference on Emerging Trends in Engineering and Technology, (ICETET), pages 1156-1159, IEEE, Nagpur, India., <https://doi.org/10.1109/ICETET.2008.11>.
- [4] Jha, S. et al. (2018). Multi Task Deep Morphological Analyzer: Context Aware Neural Joint Morphological Tagging and Lemma Prediction. arXiv preprint arXiv:1811.08619.
- [5] Nisha, M. et al. (2015). Malayalam morphological analysis using MBLP approach. In Proceedings of the International Conference on Soft-Computing and Network Security (ICSNS). IEEE, Coimbatore, India. <https://doi.org/10.1109/ICSNS.2015.7292387>.
- [6] Research Centre for Technical Development of Punjabi Language. (2004). Our online tools: Punjabi Morphological Analyzer. Punjabi University, Patiala. <http://pgc.learnpunjabi.org/#Morph>.
- [7] Shah, N.D. and Bhadka, H. (2019). Paradigm Based Morphological Analyzer for Gujarati Language. In Proceedings of the Intelligent Communication, Control and Devices (ICICCD). Advances in Intelligent Systems and Computing, vol. 989. Springer, Singapore, pages 469-481. https://doi.org/10.1007/978-981-13-8618-3_50.
- [8] Sivan, R. (2019). Study on Morphological Analyzer and Generator for Malayalam. J. IJESI, 8 3:73-77.