A Finite-State Morphological Analyzer for Saraiki

Meesum Alam[1], *Alexandra O'Neil*[1], *Daniel Swanson*[1], *Francis Tyers*[1]

[1]Indiana University, USA

`{meealam,aconeil,dangswan,ftyers}@iu.edu`

Abstract

Saraiki (also Sirariki) (skr) is the first language of almost 25 million speakers in Pakistan and nearly one million speakers in India. Our study documents the process of creating an Apertium module for Saraiki and contributes to future efforts to generate computational resources for Saraiki. Apertium is chosen for the development of a Saraiki morphological analyzer since the platform has shown to adequately handle morphological complexity. In discussing the process of creating an analyzer for Saraiki, we detail our implementation by discussing our treatment of Saraiki morphology in regard to gender, number, and case marking for nouns and adjectives, verb categorizations (basic stem forms, direct causatives, and indirect causatives), and cases of ambiguity in nominal gender inflections.

Index Terms: Finite State Technology, Apertium, Saraiki Morphology

1. Introduction

Saraiki is an Indo-Aryan language widely used in Pakistan and India [1]. The language is one of the new Indo-Aryan languages, a grouping that includes Punjabi and Hindko, which emerged in the second millennium of the common era [2] [3]. Saraiki is spoken by around 24 million people in Southern and Southwestern Punjab, Northern Sindh, the Southern district of Dera Ismail Khan and Tank of Khyber Pakhtunkhwa, and the Eastern part of Balochistan, especially the Loralai and Naseer Abad divisions. Western scholarly literature also refers to Saraiki using alternative names, such as Jataki, Multani, Western Punjabi, and Lahnda [4] in various Punjabi areas. The Muhajir settlers in the Saraiki belt are also part of the Saraiki identity; the Saraiki people mainly populate southern and western parts of Pakistani Punjab.

Developing a morphological analyzer for Saraiki is the first step for NLP related research and having morphological lexicons smooth the NLP tasks such as parts of speech tagging, morphological disambiguation and spelling correction. To our knowledge there is not a morphological analyzer for the Saraiki language that gives Part of Speech (POS) analysis and lemmas for the given words.

The paper is structured as follows: In the introduction we discuss related languages and Saraiki dialects. Section 2 discusses related work, section 3 gives an overview of the orthography, and section 4 details the corpus and annotation process. In section 5 we explain the rules we implemented for Saraiki morphology. Section 6 shows the evaluation of the analyzer and section 7 offers some concluding remarks.

1.1. Related Languages

Saraiki is counted among the widely spoken languages in the Pakistani provinces of Punjab and Khyber Pakhtunkhwa (KPK). Saraiki is the sister language of Punjabi and Sindhi. Similarly, Saraiki uses the basic Subject Object Verb (SOV) structure. Saraiki shares adjectival concord patterns with Punjabi. Otherwise, Saraiki has distinct morphological, syntactic, and phonological features when compared to Punjabi, Hindko and Sindhi [4].

1.2. Saraiki Dialects

As the language has been spoken in the northwestern Indo-Aryan region for a large part of history, multiple dialects have emerged over time. [5] distinguishes six varieties: Southern Saraiki is spoken in the south of Punjab, which includes Dera Ghazi Khan, Muzafargarh and Rahim yar khan. Northern Saraiki is spoken in the district Dera Ismail khan and Mianwali, which is close to the Central Saraiki dialect in its various features. Sindhi Saraiki is a mixure of the Saraiki and Sindhi languages and is spoken in the areas of Sindh and bordering part of the Punjab. Jhangi Saraiki is mostly spoken in the area of the Jhang District. It has some distinctive phonological features, such as dental implosives [5]. Shahpuri Saraiki is spoken in the Sargodha District of Punjab and shares some features with the Central dialect, but is also transitional to Punjabi. The Central dialect of Saraiki is spoken in Multan and in its surrounding regions.

Saraiki, as well as its neighbouring languages, are relatively low-resourced in terms of existing linguistics research. However, some work has been done on Western dialect Punjabi and Sindhi which can be used to provide a basic understanding of the linguistics features of Saraiki. Saraiki is often considered a neglected language of the region for political reasons. This work not only adds to the existing work on the Saraiki language but provides a foundation for related languages.

2. Related Work

There is not much work on Saraiki linguistics; however, Bashir and Conners[1] explains the concept of Saraiki grammar in relation with Punjabi and Hindko. Bashir and Conners's work on Saraiki, Punjabi, and Hindko serves as a guide for linguists who are interested in work-

ɓ	ٻ	ɲ	ڃ	ŋ	ڱ
b^{ɦ}	ڀ	ʄ	ڄ	ɠ	ڳ
d^{ɦ}	ڌ	c^{h}	ڇ	k	ڪ
ɗ	ڏ	ʈ^{h}	ٺ	ɳ	ڻ
ɖ	ڊ	ʈ	ٽ		
ɖ^{ɦ}	ڍ	t^{h}	ٿ		

Table 1: *Characters used to write Saraiki which are not used in Persian script, along with their phonetic values.*

ing on marginalized languages in Pakistan. Her work provides a detailed description of the grammar for Saraiki, Punjabi, and Hindko and covers the common and unique grammatical features of each language. While discussing a morphological analyzer for Punjabi, [6] provide insight on their implementation of morphology, the development of a corpus, and the building of a lexicon for the Punjabi language. They provide a detailed description of Punjabi morphology for nouns, adjectives, and verbs.

The work of [7] on Saraiki verbs focuses on causativity and the morphology of transitive and intransitive verbs. Their paper divides the intransitive and transitive verbs into six categories with each category corresponding to a set of inflections that can create additional verbal forms. Additionally, they explain that verbs with number and gender also change their endings according to the direct or oblique case of the object/subject. Additionally, the work of [8] on a Sindhi morphological analyzer is helpful in understanding morphological processes in Saraiki. This paper explains the morphology of Sindhi which overlaps well with Saraiki morphology for many forms. Further, this paper explains the corpus, rules, and implementations of the finite state analyzer for Sindhi.

3. Orthography

The orthographical history of the Saraiki language has been strongly influenced by introduction of Islam in the region, resulting in the historical use of different orthographies based on the script of those in the position of dominance. Before the Islamization of the region, Saraiki was written in the Devanagri script. Following this period, the language started using the Perso-Arabic script. The modern orthography is slightly adapted from the Perso-Arabic script with the addition of characters that are necessary to represent the sounds present in Sindhi, Saraiki, and Khetrani. A complete list of these additional characters can be found in Table 1.

4. Corpus and Annotation Process

We used the Mozilla Common Voice dataset of Saraiki for our morphological analyzer. Common Voice is a publicly available voice and text dataset that is powered by the voices of volunteer contributors around the world. Researchers that want to build voice applications can use this corpus to train machine learning models. The corpus is licensed under CC-0 which means that it can be used freely for research [9]. The corpus contains the proverbs and translations of Quranic verses in the Saraiki language. In the first step, we extracted six thousand sentences into an excel sheet. The average length of the sentences were between 7 to 10 tokens. We then tagged the tokens in the sentences according to their part of speech. These sentences were manually tagged by a native speaker of Saraiki who is the first author of the paper. Table 4 shows a list of stems with part of speech tags that were added to the lexd file of the analyzer.

5. Morphology

Indo-Ayran languages are known for their rich morphology. As Saraiki is spoken in the central region between Sindhi, Western Punjabi, Khetrani and Baloching, many of the features common to these languages are reflected in Saraiki. Nouns follow the patterns of these languages with double-borrowing. Double-borrowing refers to a two-step borrowing process that words undergo before being adopted into a language. For example, nouns like قوم رن, /ɹʌn kɔm/ which entered into the Balochi language from Persian, were reshaped to Balochi morphology, and then were borrowed into Saraiki, adopting Saraiki morphological features. Verbs in Saraiki have some inflectional patterns that are related to neighbouring languages, but also have their own unique forms. The following sections of the paper give a detailed overview of the morphological features in open and closed classes of Saraiki.

5.1. Nouns

Like most of the Indo-Aryan languages, Saraiki nouns have patterns that reflect gender. Feminine nouns in Saraiki usually end with the vowels [i] or [ɪ] or the consonant [t]. Masculine nouns end with the vowel [a] or the consonant [r]. However, this inventory of noun endings is not comprehensive as there are some nouns that end with feminine vowels, despite being classified as a masculine form, such as ,"قصائی" سخی /sʌχi qʌsaɪ/. Based on the corpus findings and the literature related to Saraiki nouns, we have divided nouns into three main categories: Case, Caseless, and Unmarked.

Case nouns are inflected for gender, number and case. These nouns can inflect for any of the following four cases: direct, oblique, vocative or ablative. Case nouns also use syncretism in inflection, as the inflection for the direct case plural and the oblique singular is the same for masculine forms [ے]. The difference between the inflection of masculine oblique plural and vocative singular is nasalization [ں]. Case feminine nouns have the same singular form for the direct, oblique and vocative case, but differentiate in their marking for the ablative case, represented by [و].

Caseless nouns are inflected for gender and number, but do not inflect for case. Caseless nouns that end in nasalized vowels [ں] are pluralized by the addition of [واں], while Caseless nouns ending with consonants use [اں] to produce plural forms.

The third category of nouns is Unmarked nouns. Unmarked nouns are borrowings that originate from Arabic, Persian or Turkish words. These nouns do not have a standardized plural form, but observational input from speech indicates that Saraiki speakers have begun inflecting Unmarked nouns in the same way as Case nouns. There are numerous examples of Unmarked nouns being inflected as Case nouns in the Common Voice corpus, reflecting the necessity of further research on Saraiki nouns and the potential need to further segment nouns into subclasses.

5.2. Verbs

The Saraiki verb system is complex and has features that are unattested in neighbouring languages such as Punjabi, Urdu and Hindko. Intransitive and transitive verbs are formally distinct in Saraiki. This feature is also present in Khetrani and Sindhi. A unique morphological process found in Saraiki is the creation of transitive verbs from intransitive verbs – a process that can be achieved through various patterns of affixation. In this analyzer we worked on the inflections of both intransitive and transitive forms, with first and second causativity of verbs. These verb inflections overlap with each other with different base forms; moreover, some of these inflections combine with auxiliary verbs to express tense/mood and aspect. In addition, not all inflections attach to every verb root. Table 3 shows verb inflections for پڑھ /pʌɽh/ 'read, study'.

5.3. Adjectives, Pronouns and other categories

Adjectives in Saraiki are either inflected or unmarked. Inflected adjectives can take one of four cases: direct, oblique, vocative, and ablative. The inflected adjectives have syncretism in their inflections. Adjective inflections are the same as the inflections of nouns in Saraiki.

The reflexive pronouns take most noun inflections. The other pronouns for first, second, third proximal, and third distal pronouns take the inflections in oblique, agentive, possessive and dative-accusative cases. [1] refers to dative-accusative as a postposition for pronouns. However, throughout the corpus we found it attached to pronouns, so we analyzed this construction as a postposition in compound form with pronouns.

Other categories are the close categories including adverbs, auxiliaries, post-positions, conjunctions and determiners. Auxiliaries in Saraiki function to explain tense and sometimes function in tandem with nouns that are functioning as verbs. The adverbs in Saraiki are unmarked as in Sindhi and Punjabi. Additionally, postpositions are attached to nouns and verbs and sometimes there is a combination of post-positions.

6. Implementation

We implemented the morphological analyzer as a finite-state transducer in the Apertium framework [10], [11] using the lexicon compiler lexd [12].

An analyzer source file in lexd consists of a collection of "lexicons" and "patterns". Lexicons are lists of pairings of analyses and surface forms, optionally with sets of labels to reduce duplication in patterns, with the most frequent type of pairing being between a dictionary headword ("lemma") and the minimal orthographic form onto which affixes are added ("stem"). Patterns, meanwhile, define how lexicons can be concatenated to form full words. An example is given in (1).

(1)
```
PATTERN NOUN
NounRoot[m,-unmarked,-I] NounInflMasc
NounRoot[f,-unmarked,-I] NounInflFem

LEXICON NounInflMasc
<n><m><sg><dir>:ہ
<n><m><sg><dir>:ا

LEXICON NounInflFem
<n><f><sg><dir>:ی
<n><f><pl><dir>:یاں

LEXICON NounRoot
باز:بازی[f]
باغ:باغ[m]
```

Here we define a pattern for nouns, the first line of which states that a noun may be formed by selecting any entry from NounRoot which is labeled as masculine (m) and which is not labeled as unmarked (unmarked) or caseless (I) and concatenating it with any entry from NounInflMasc.

We then define the sets of inflectional suffixes. For example, in the first line of NounInflMasc it states that the analysis tags for noun (<n>), masculine (<m>), singular (<sg>), and direct (<dir>) together correspond to the surface string ہ.

Finally, in NounRoot, we list the lexical forms, stating in the first line that there exists a noun whose lemma is بازی IPA [bazi] and whose stem is بازIPA [baz]. We further specify that this is a feminine noun ([f]).

Lexical entries were collected in consultation with a native speaker by examining the forms in the Common Voice corpus in order of frequency. The distribution of stems by part of speech is given in Table 4. At present, our lexicon is approximately 60% nouns and 15% verbs.

7. Evaluation

To evaluate the analyzer, we randomly sampled distinct forms from the Common Voice corpus and removed ones that were not words. The correct analyses for each of these 545 forms were then manually checked to create a gold standard for our evaluation. We then compared our gold standard forms to the output generated by our analyzer and calculated precision and recall.

The precision score compares the output generated by the analyzer to the gold standard forms. Our score of 93.18% means that roughly 14 out of every 15 analyses returned by our analyzer are correct. That is, if our analyzer returned 3 analyses for each of 5 words, on average we would expect one of those 15 to be wrong.

The recall score indicates the percentage of correct analyses that are produced by our analyzer when considering the full set of analyses that exist in the gold standard. Our score of 96.99% shows that our analyzer produces most of the possible analyses of forms, but is short of producing all of the possible analyses for all forms. That is, if our analyzer returns 29 total analyses for a set of words, we would expect, on average, that only one correct analysis is missing from that set.

Naive coverage is the percentage of forms that have at least one analysis. Our naive coverage of 83% is a bit low for a production analyzer, with a bit more than one word in seven missing an analysis, but the high precision

Gender	Number	Direct	Oblique	Vocative	Ablative
Masculine	Singular	بندہ /bndja/	بندے /bnde/	بندیا /bndɦ/	بندیو /bndjv/
	Plural	بندے /bndjv/	بندیاں /bndjv/	بندیو /bndjã/	بندیو /bnde/
Feminine	Singular	بندی /bndj/	بندی /bndj/	بندی /bndj/	بندیو /bndjv/
	Plural	بندیاں /bndjã/	بندیاں /bndjv/	بندیو /bndjã/	بندیاں /bndjã/

Table 2: *The paradigm of the Saraiki noun* بندہ /banda/ *'person'.*

Stem: پڑھ /pʌɽh/	1st Person		2nd Person		3rd Person	
Masculine Singular	/pɽ^{h}sã/	پڑھساں	/pɽ^{h}sj̃/	پڑھسیں	/pɽ^{h}sj/	پڑھسی
Masculine Plural	/pɽ^{h}sj̃/	پڑھسیں	/pɽ^{h}sv/	پڑھسو	/pɽ^{h}sn/	پڑھسن
Feminine Singular	/pɽ^{h}sã/	پڑھساں	/pɽ^{h}sj̃/	پڑھسیں	/pɽ^{h}sj/	پڑھسی
Feminine Plural	/pɽ^{h}sj̃/	پڑھسیں	/pɽ^{h}sv/	پڑھسو	/pɽ^{h}sn/	پڑھسن

Table 3: *The basic paradigm of Saraiki verb* پڑھ /pʌɽh/ *'read, study'.*

Part of Speech	Stems
Adjective	341
Adverb	215
Auxiliary	22
Conjunction	17
Determiner	43
Noun	2031
Number	8
Postposition	24
Pronoun	46
Proper Noun	17
Verb	527
Total	3291

Table 4: *Number of stems for each part of speech in the analyzer.*

Metric	Score
Naive Coverage	82.94%
Precision	93.18%
Recall	96.99%
F1 Score	95.05%

Table 5: *Evaluation scores*

indicates that the analyses that are provided can be expected to be highly accurate. In this context, coverage and recall provide very similar information, with recall being higher because analyses outnumber surface forms.

Using the precision, and recall scores, we compute a F1 score. The F1 score is the harmonic mean of these scores and indicates the percentage of tokens in the corpus which receive at least one analysis. Thus, the higher F1 score is indicative of the analyzer's ability to correctly produce analyses for high frequency tokens. The results are listed in Table 5.

8. Conclusion

This paper presents a free open-source finite state analyzer for Saraiki, available at `https://github.com/apertium/apertium-skr`. Using the Common Voice dataset for Saraiki morphology as the corpus, our analyzer produces a coverage of 83% with excellent precision and recall. Thus, the primary next step for improving the analyzer is to simply add more roots.

In addition to expanding the analyzer, we are planning to develop a morphological disambiguator. This will help the analyzer determine which of the analyses is most likely for a form. Once the disambiguator is developed, the monolingual dictionary will be used in the creation of an Apertium-based, Saraiki-English bilingual dictionary that can be used for translation.

Saraiki belongs to Lahnda group of Indo-Aryan languages [4] which are mostly similar in terms of morphology and syntax. Due to the similarity of the languages, our Saraiki analyzer offers a foundation for building a comprehensive Apertium module for all seven languages of the Lahnda group. This multi-lingual module will then be able to function as a base NLP resource for all languages and dialects in the Lahnda group, further enhancing the accessibility of language-technology for speakers of Lahnda languages.

This research also contributes to the current development of a spell checker using the Apertium framework. The progress on creating a parts of speech tagger for Saraiki also paves the way for further technological advancement in Saraiki NLP, such as the creation of a Universal Dependency treebank.

9. Acknowledgements

We would like to thank Pervaiz Qadir for developing the Saraiki Common Voice dataset. We would also like to thank Sandra Kuebler for feedback on a draft of this paper and Jonathan Washington for helping with LaTeXissues.

References

[1] E. Bashir and T. J. Conners, “A descriptive grammar of hindko, panjabi, and saraiki,” in *A Descriptive Grammar of Hindko, Panjabi, and Saraiki*, De Gruyter Mouton, 2019.

[2] C. P. Masica, *The Indo-Aryan languages*. Cambridge University Press, 1993.

[3] I. Ullah, M. G. Abbasi, M. M. Abbasi, Y. Arafat, and G. Asghar, “Historical background of the origin and evolution of Saraiki language in central Pakistan: A case study of Saraiki language in Dera Ghazi Khan,” *PalArch’s Journal of Archaeology of Egypt / Egyptology*, vol. 18, no. 10, pp. 621–627, Aug. 2021.

[4] G. A. Grierson, “Linguistic survey of India,” 1927.

[5] C. Shackle, “The Siraiki language of central Pakistan: A reference grammar,” 1976.

[6] M. Humayoun and A. Ranta, “Developing Punjabi morphology, corpus and lexicon,” in *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, Tohoku University, Sendai, Japan: Institute of Digital Enhancement of Cognitive Processing, Waseda University, Nov. 2010, pp. 163–172. [Online]. Available: `https://aclanthology.org/Y10-1020`.

[7] J. J. Lowe and A. H. Birahimani, “Causative alternations in Siraiki,” *Transactions of the Philological Society*, vol. 117, no. 2, pp. 266–293, 2019.

[8] R. Motlani, F. Tyers, and D. M. Sharma, “A finite-state morphological analyser for sindhi,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 2016, pp. 2572–2577.

[9] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common voice: A massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.

[10] M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O’Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers, “Apertium: A free/open-source platform for rule-based machine translation,” *Machine translation*, vol. 25, pp. 127–144, 2011.

[11] T. Khanna, J. N. Washington, F. M. Tyers, S. Bayatlı, D. G. Swanson, T. A. Pirinen, I. Tang, and H. Alòs i Font, “Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages,” *Machine Translation*, vol. 35, no. 4, pp. 475–502, 2021.

[12] D. Swanson and N. Howell, “Lexd: A finite-state lexicon compiler for non-suffixational morphologies,” in *Multilingual Facilitation*, M. Hämäläinen, N. Partanen, and K. Alnajjar, Eds., University of Helsinki Library, 2021, ISBN: 978-951-51-5025-7. [Online]. Available: `https://helda.helsinki.fi/handle/10138/327807`.