Asesorías y Tutorías para la Investigación Científica en la Educación Puig-Salabarría S.C.
José María Pino Suárez 400-2 esq a Lerdo de Tejada. Toluca, Estado de México. 7223898475
RFC: ATI120618V12

Revista Dilemas Contemporáneos: Educación, Política y Valores.
http://www.dilemascontemporaneoseducacionpoliticayvalores.com/
Año: VI Número: Edición Especial Artículo no.:39 Período: Agosto, 2019

TÍTULO: Desarrollo de marcas diacríticas para los nombres y verbos de Punjabi Shahmukhi.

AUTORES:

1. M.Phil. Muhammad Ahmad Hashmi.

2. Ph.D. Muhammad Asim Mahmood.

3. Ph.D. Muhammad Ilyas Mahmood.

RESUMEN: El estudio ha sido diseñado para aplicar marcas diacríticas a 1000 palabras del Punjabi, incluidos 800 nombres y 200 verbos. El corpus de 2 millones de palabras ha sido tomado de diferentes libros, periódicos, revistas, artículos y novelas. Punjabi Shahmukhi carece de recursos digitales en línea para desarrollar diferentes herramientas del Procesamiento de Lenguaje Natural (PLN), lo que ayudaría a reconocer su estado internacional. Punjabi Shahmukhi tiene una escritura "Perso-árabe" y ha sido ignorada por los lingüistas para digitalizar su literatura. El estudio es significativo ya que servirá para el desarrollo de WordNet, ayudando a desarrollar un etiquetador de la parte del habla del Punjabi Shahmukhi, digitalizará su literatura y ayudará a los maestros y no nativos a desarrollarse una armonía intercultural.

PALABRAS CLAVES: digitalización, diacríticos, Procesamiento del Lenguaje Natural (PNL), Punjabi Shahmukhi, WordNet.

TITLE: Development of Diacritical Marks to Punjabi Shahmukhi nouns and verbs.

AUTHORS:

1. M.Phil. Muhammad Ahmad Hashmi.

2. Ph.D. Muhammad Asim Mahmood.

3. Ph.D. Muhammad Ilyas Mahmood.

ABSTRACT: The study has been designed to apply diacritical marks to 1000 Punjabi words including 800 nouns and 200 verbs. The corpus of 2 million words has been taken from the different books, newspapers, magazines, articles and novels. Punjabi Shahmukhi lacks any online digital resource to develop different tools of Natural Language Processing (NLP), which will help to recognize the international status of it. Punjabi Shahmukhi has "Perso-Arabic" script and has been ignored by linguists to digitize its literature. The study is significant as it will serve its part in the development of WordNet and will help to develop a Part of Speech (POS) tagger of Punjabi Shahmukhi, digitize the literature of Punjabi Shahmukhi and be helpful for the teachers and non-natives to develop an intercultural harmony.

KEY WORDS: digitization, diacritics, Natural Language Processing (NLP), Punjabi Shahmukhi, WordNet.

INTRODUCTION.

The purpose of present study is to develop diacritical marks to the nouns and verbs of Punjabi Shamukhi, which will help to disambiguate the different senses of words with same orthography. Punjabi language has its two scripts: Gurmukhi which is in practice in Eastern Punjab of India and Shahmukhi which is spoken in Western Punjab of Pakistan.

Punjabi Shahmukhi is a script of Punjabi language, which is called "Perso-Arabic" used by the people of Western Punjab Pakistan. It is a combination of various local scripts of Urdu language. It

follows right-to-left writing system and shapes of words are assumed according to the context (Saini, Lehal & Kalra, 2008).

The orthography of Shahmukhi follows the Nastalique script which is thought to be a cursive one and highly complex system of writing (Lehal, 2009). Furthermore, it has 35 consonants sounds including 15 aspirated consonants, 15 marks for diacritical, one nasal sound and 10 digits (Lehal et al., 2009).

Malik (2006) identifies that Shamukhi script consists of 37 simple consonants, 4 long vowels and 3 short vowel symbols. Unfortunately, Punjabi Shamukhi lacks any online digital resource which could serve as a base to digitize the literature of Punjabi Shahmukhi. It has been started an effort to develop the latest form of language dictionary in the form of WordNet at Government College University, Faisalabad.

In the processing the syntactic categories of Punjabi Shahmukhi, it has been observed that one word has more than one senses in different situations. These senses are recognized with the change of accent while speaking. On the other hand, diacritical marks help to distinguish among the orthographic text of Punjabi Shahmukhi which are yet to be developed. So, the present study is designed to develop diacritical marks for the selected number of nouns and verbs.

According to Motlani (2016). Lack of diacritics creates semantic ambiguities, where words may have more than one interpretation; for example: ملک can have the meaning of mulk (country) or can have the meaning of milk (milk). Moreover, the word کڑی can be interpreted as kari (proof) or can be taken as kuri (girl). The absence of diacritics also leads to the syntactic ambiguities, e.g., the word دل changes its meaning to an ‘organ system of body’ if a short vowelZer (ِ), is put to it like دِل which is a noun, otherwise it will lead to the ‘process of grinding’ which is a verb.

Writers do not take it standard to miss diacritics because it creates a number of problems including the wrong interpretation of target words. In contrast to these standard directions, Motlani (2016) has reported a qualitative evaluation that is inputed with diacritics and makes it difficult to look up the target word in lexicon. One of the main reasons behind this problem is the natives of perso-Arabic script who do not use diacritics while writing text because their readers could understand the target words in their context, but there is still needed to develop an inter-cultural understanding of language to whom for those Punjabi Shahmukhi serves as a second language.

Malik (2006) describes that Punjabi Shahmukhi is based on Nastalique style of Persian and Arabic script. Its characters are context-sensitive with thirty eight letters, including four long vowels Alif (ا), Vao (و) [v], Choti-ye (ی) [j] and Badi-ye ﴿ے﴾ [j], three short vowels Zer (ِ), Pesh (ُ) and Zabar (َ), diacritical marks like Shad (ّ), Khari-Zabar (ٰ), do-Zabar (ً) [ən] and doZer (ٍ) [In], or symbol hamza [(ء)]. Furthermore, it has ten aspirated consonant sounds (**پھ ، بھ، تھ، ٹھ، جھ، چھ، دھ، ڈھ،کھ، گھ**) which are used frequently in the company of six aspirates (**لھ، رھ، ڑھ**) **[**ھ**، نھ، مھ**].

The present digital lexicon and online dictionaries are not taken as a reliable database for Punjabi Shahmukhi because of many flaws (Malik, 2006), e.g., the written material in Punjabi Shahmukhi does not use long and short vowels and diacritical imprints. It takes the data without diacritics and interprets without its any limitation of reader's desire (Hasan, Iqbal, Azeemi & Javeed, 2015). Moreover, the words cannot be distinguished with their relevant connotation to recognize their meanings in different senses, e.g., the word **ترنا** can be written in two ways with diacritics **تُرنا**(To walk/move) and**تَرنا** (To swim/ float). Similarly, the word **بیل**means 'the vine' and **بَیل**means 'the bull'. In this way diacritical marks are necessary to distinguish among a number of connotations of same words.

Diacritical marks are taken as the backbone of the vowel system in Punjabi Shahmukhi because these are the reasons for correcting understand and pronunciation of a word (Malik, 2005). No doubt, these are sparingly used in common writing including newspaper, books and magazines etc., but it creates serious problems while handling it digitally. These diacritical marks help in Word Sense Disambiguation (WSD) process which serves as a main part in the development of WordNet.

DEVELOPMENT.

The present study is associated with the project of developing WordNet for Punjabi Shahmukhi. WordNet is taken as a most modern digital form of any language in order to digitize the literature of any language. It serves as a base to run the applications of Natural Language Processing (NLP) including artificial intelligence like Automatic Summarization, Computational Lexicography, Machine Translation, Automatic Morphological analyzer, Named Entity Recognition, Optical Character Recognition, Digital documentation, Parsing, POS-Tagging, Sentence Breaking, Text Mining, Sentiment analysis, Text to Speech, Automatic speech recognition, Speech to Text, Speech to Speech, Information Retrieval, Speech Identification and many more (Hashmi, Mahmood & Mahmood, 2019).

Significance of the Study.

The present research is significant as it will be used in the development of WordNet for Punjabi Shahmukhi in near future. It will also provide a reliable digital database which will be able to distinguish among the different senses of same words so, to help the better understanding of Punjabi Shahmukhi to those who are not the native of this language for the sake of cultural harmony.

Research Question.

How the diacritical marks help to distinguish among the different sense of same words in Punjabi Shahmukhi nouns and verbs?

Limitations of the Study.

The current study is limited to the nouns and verbs of Punjabi Shahmukhi and does not deal with the other open syntactic categories including adverbs and adjectives.

Methodology.

Data Collection.

A corpus of 2 million words has been taken from the different books, newspapers, magazines, articles and novels. To tag this corpus, it has been transliterated into Gurmukhi script of Punjabi language because the tagger of Punjabi Shahmukhi has not been developed yet. After tagging, it has been again transliterated into Punjabi Shahmukhi script.

Using an online tool of Laurence Anthony's AntConc, which is freely available on internet, the lists of nouns and verbs have been extracted. Then machine errors have been removed manually by looking at each verb and noun in online and manually available resources of Punjabi Shahmukhi. The final lists of 800 nouns and 200 verbs have been developed.

Data Analysis.

The final lists of nouns and verbs have been analyzed manually form the available lexicon including Punjabi Wikipedia, manual dictionaries and asking the native speakers where necessary. The online resources include Punjabi dictionary https://www.ijunoon.com/punjabi_dic/, Punjabi Wikipedia https://pnb.wiktionary.org/wiki/, which has more than 9000 words in its database and another Punjabi Wikipedia https://pnb.wikipedia.org/wiki/, which has a huge data of 46546 articles; then, diacritical marks have been applied to disambiguate among the different senses of words. The final draft has been checked by a Punjabi Expert from the Department of Punjabi at Government College University, Faisalabad. The amendments have been made where necessary according to the guidelines of Punjabi expert.

Results and interpretations.

The results have showed that the same word is having different meanings in different contexts. The meaning of these words can only be understood by native speakers of Punjabi Shahmukhi. Here, the following table shows three columns including ‘received ontology, corrected ontology and meaning.

The words in received ontology represents the final product after tagging from Gurmukhi and transliterated into Shahmukhi script. The extracted list of 800 nouns has been analyzed using different online and manual resources including dictionaries. These words have been given diacritical marks to set their one particular sense for the sake of Word Sense Disambiguation (WSD) and are put in the column of ‘corrected ontology’ On the biases of correct ontology with proper diacritics the meanings have been deduced.

Table 1.

Received Ontology	Corrected Ontology	Meaning
دکھ	دُکھ	غم
بیر	بیِر	اک فروٹ دا ناں
جھاں	جہان	دُنیا
سکھ	سُکھ	سکون
اٹ	اِٹ	چکڑ نوں پکا کے مستطیل مورت چ کوٹھے کند بنان والا
انڈا	آنڈا	آنڈا اک ایسا جیوندا پانڈا اے جیدے وچ کوئی جمن توں پہلے جنور ودھدا اے
حسن	ُحسن	رُوپ
مک	مُک	گھونسا
ونڈو	وِنڈو	باری
ویر	وِیر	بھرا
مکھ	ُمکھ	مُکھڑا
ابی	آبی	پانی بھرا

انت	آخر	اخر
رات دا اُڈاری	اُلّو	الو
پُتلا	بُت	بت
جُثہ	ِجسم	جسم
آگ دی مخلوق	جِن	جن
جگتاں مارنا مذاق کرن	ُجگت	جگت
انگ پاسہ کسے شے دا وکھرا ہویا یا تڑوڑیا گیا انگ	حِصّہ	حصہ
حُقا اک یا زیادہ نالیاں آلا تماکھو نوشی لئی استعمال کیتا جاݨ آلا اک قدیم ینتر اے۔ جیہدا جنمستھان پنجاب نوں قرار دتا جاندا اے۔ حُقے دے بھانڈیاں چ اک نیچا تے اک چِلم ہُندی اے۔ نیچا تھللے پاسے ہُندا اے جیہدے وِچ پاݨی پایا جاندا اے تے چِلم حُقے دے اُتلے پاسے ہُندی اے جیہدے وِچ کولے رکھے جاندے نیں تے ذائقے لئی ون سونیاں جڑیاں بوٹیاں وی پائیاں جاندیاں نیں	حُقہ	حقا
زمین دا اک ٹوٹا	خِطّہ	خطہ
سند (اردو:سندھ) پاکستان دے چار صوبیاں وچوں اک صوبہ اے ۔ پرانے ویلیاں توں ایہہ سندیاں دا دیس اے ۔	سِند	سند
کرودھ	غُصّہ	غصہ
دئیں دُدھ توں پین لئی بنایاگیاپانی سار	لسّی	لسی
جوان لڑکا	مُنڈا	منڈا
روپ	مُورت	مورت
مٹی نمکیات، نامیاتی مادے، گیساں، پانی سار تے اوہناں جانداراں دا گڑھ اے جو بوٹیاں دا جیون بناندے نیں۔	مِٹی	مٹی
لمبائ ناپن دا اک ناپ	مِیل	میل
کرکٹ دے ویڑھے چ بھیج تے کھڑے 6ڈنڈے	وِکٹ	وکٹ
مرد نیانا	پُتر	پتر
سیکنڈ	پَل	پل
کتا بھگیاڑ دی اک ونڈ اے۔ ایدا جوڑ دد پلانے آلے جانوراں نال اے۔ کتا انساناں دے نال پرانے ویلیاں توں رہ ریا اے۔	کُتا	کتا
کرکٹ بال بلے دا کھیڈ اے جہنوں دنیا چ چوکھا کھیڈیا جاندا اے۔ ایہہ دو ٹیماں دے وشکار ہوندا اے تے ہر ٹیم چ یاراں کھڈاری ہوندے نیں۔	ِکرکٹ	کرکٹ

جاتکڑی, مادہ	کُڑی	کڑی
اک جنگلی جانور	ِگدڑ	گدڑ
کڑیاں دے کھیڈن لئی آپ بنائی گئی گ ئ نکی کڑی ورگی	گُڈی	گڈی
اک پھَل دا ناں	آڑو	آڑو
بلا بلی ٹبر دی اک جنس اے۔ ایہدیاں چار ونڈاں نیں۔ باب بلا، یوریشی بلا، کینیڈی بلا، آئیبیری بلا۔	ِبلا	بلا
امب دے پھل توں پہلاں لگن والی شے	ُبور	بور
انسان تے جانواں چ لہو نوں اگے ٹورن والا انگ	دِل	دل
کہر	دُھند	دھند
دیوا، چراغ یا دیپک توں مراد تیل جلا کے چانڑ کرن دا برتڑ اے	دِیا	دیا
ایمان	دِین	دین
موسم	رُت	رت
کپاہ	ُروئی	روئی
رُوح کسے بندے، جانور یا ہور جیوندی شے دے وجود دی اوہ نہ دکھن والی طاقت اے جیہڑی جسم نوں حیاتی بخشدی اے۔	رُوح	روح

The word **دکھ** has different interpretations in a variety of contexts. The current orthography of **دکھ** means 'lice' but when the Pesh (ُ) is put to it becomes**دُکھ** which means 'sorrow'. Likewise, the word **بیر**means 'animosity but after putting Zer (ِ) it becomes **بِیر** 'a kind of fruit'. In this way diacritics play an important part in deciding the sense of particular word in specific context. **سکھ، اٹ، انڈا، حسن، تصویر، مک، ونڈو، ویر، ٹھاں، مکھ، ابی، اخر، الو، بت، جسم، جن، جگت، حصہ، حقا، خطہ، سند**…etc., have different meanings when applied to diacritical marks as **سُکھ، اِٹ ، آنڈا، ُحسن، تصویر، مُک، وِنڈو، وِیر، تھاں، ُمکھ، آبی، آخر، اُلّوم بُت، ِجسم، جِن، ُجگت، حِصّہ، حُقہ** …etc.

There occurs certain derivational changes when diacritical marks are applied to specific nouns which also represent some actions for example, **دیا** means 'to give something away' which is a verb but when Zer (ِ) is applied, it becomes **دِیا** means 'lamp' which is a noun. Likewise, **اٹ** means 'to

undergo something’ which is a verb and when Zer (ِ) is put to it, it becomes اِٹ and gives the meaning of ‘a brick’ which is a noun.

The same nature of results has been deduced while analyzing the verbs of Punjabi Shahmukhi. The following list of verbs has been developed with specific senses by applying diacritical marks.

Table 2.

Received Ontology	Correct Ontology	Received Ontology	Correct Ontology
رل	رُلنا	گھما	گُمنا
چگدے	چُگنا	آکھ	آکھنا
مڑ	مُڑنا	کھچن	کِھچنا
ونڈو	ونڈنا	سکھن	سِکھنا
اواں	آنا	مٹا	مِٹانا
سکے	سُکنا	اکھڑ	اُکھڑنا
ہسدا	ہسنا	پسر	پِسرنا
رڑکے	رڑکنا	کلنا	گھلنا
پڑھ	پڑھنا	اڈنا	اُڈنا
اڑنا	اُڑنا	ٹٹنا	ٹٹنا
پچھ	پُچھنا	ٹرنا	ٹُرنا
لکھی	لِکھنا	گزری	گُزرنا
ملدی	ملنا	کترن	کُترنا
بلا	بُلانا	سونا	سُونا
سن	سُننا	پھڑ	پَھڑنا
چکیا	چُکنا	پکڑ	پھڑنا
کھر	کُھرنا	سنگڑ	سُنگڑنا
مکا	مُکنا	گُم	گُمنا
ڈگ	ڈِگنا	الٹ	اُلٹنا
تھک	تُھکنا	ابال	اُبالنا
اٹھی	اُٹھنا	ڈبو	ڈبونا

پُورنا	پور	چُھپنا	چھپی
کُوکنا	کوک	اُترنا	اتار
چُبھنا	چبھن	کُٹنا	کٹی
نِچُڑنا	نچڑ\	لُکنا	لکے
اڈانا	اڈا\	جُڑنا	جڑے
گھلنا	گھلّ\	بُجھنا	بجھے
ڈِگنا	ڈیگ\	چُمنا	چم
چِلّانا	چلان\	اُگنا	اگدی
دِسنا	دسیا	چِھڑنا	چھڑن
لُکانا	لکاو\	ٹِکنا	ٹکی
دھرَونا	دھرو\	پِنجنا	پجن
نکھیڑنا	نکھیڑے	جِتنا	جتیا
ورھنا	برس	بھجنا	بھجے
چُوسنا	چوس	چُننا	چنن
بڑھکنا	بڑک	اُبھرنا	ابھر
پِٹنا	پٹن	رُسنا	رسن
جننا	جنن	سَدنا	سدیا
رِڑھنا	رڑو	ٹنگنا	ٹنگی

Results have shown that diacritical marks play an important role in defining proper meaning of the given word; for example,رلنا means ‘to mix’ but when Pesh (ُ) is put it becomes رُلنا which has a meaning of ‘insulted’.

The verb اڑنا has a meaning of ‘getting jam’ but applying the Pesh (ُ) changes its meaning to ‘fly’. Many of the verbs can be changed into nouns when put to diacritical marks as we have observed about nouns above. For example, the verb پُور represent the process of covering something underground but without Pesh (ُ) it becomes پورmeans ‘the fingertips’ which is noun.

CONCLUSIONS.

To conclude this study, it is obvious to say that diacritics play a basic role in defining the syntactic categories and Word Sense Disambiguation (WSD) of any context sensitive language.

Punjabi Shahmukhi has “Perso-Arabic” script and is context sensitive language. It lacks its digital representation to develop various application of Natural Language Processing. It is digitally a young language, which lacks the availability of WordNet, yet.

Word Sense Disambiguation is a critical component in developing its WordNet, and it works by applying diacritical marks to classify possible meaning of same orthography. It has been classified the 800 nouns and 200 verbs with diacritics showing their specific meanings. The same orthographies showing different syntactic categories after applying diacritics have also been classified in their respective tables.

BIBLIOGRAPHIC REFERENCES.

1. Hasan, E., Iqbal, M. M., Azeemi, Q. R., &Javeed, A. (2015). An online Punjabi Shahmukhi Lexical Resource. Sci. Int (Lahore), 27, 2529-2535.
2. Hashmi, M. A., Mahmood, M. A., & Mahmood, M. I. (2019). Analysis of Lexicon-Semantic Relations of Punjabi Shahmukhi Nouns: A Corpus Based Study. International Journal of English Linguistics, 13. Retrieved from http://www.ccsenet.org/journal/index.php/ijel
3. Lehal, G. S. (2009). A Gurmukhi to Shahmukhi Transliteration System. In proceedings of ICON-2009: 7th international conference on Natural Language Processing (pp. 167-173).
4. Malik, A. (2005, April). Towards a Unicode Compatible Punjabi Character Set. In 27th Internationalization and Unicode Conference (p. 9).

5. Malik, M. G. (2006, July). Punjabi machine transliteration. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 1137-1144). Association for Computational Linguistics.
6. Malik, M. G. (2006, July). Punjabi machine transliteration. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 1137-1144). Association for Computational Linguistics.
7. Malik, M. G. Abbas. 2005. Towards Unicode Compatible Punjabi Character Set. In proceedings of 27th Internationalization and Unicode Conference, 6 – 8 April, Berlin, Germany.
8. Motlani, R. (2016). Developing language technology tools and resources for a resource-poor language: Sindhi. In Proceedings of the NAACL Student Research Workshop (pp. 51-58).
9. Saini, T. S., Lehal, G. S., &Kalra, V. S. (2008, August). Shahmukhi to Gurmukhi transliteration system. In 22nd International Conference on on Computational Linguistics: Demonstration Papers (pp. 177-180). Association for Computational Linguistics.

DATA OF THE AUTHORS.

1. **Muhammad Ahmad Hashmi.** SESE English, School Education Department, Okara, Punjab, Pakistan. He received Master of Philosophy Degree from Department of Applied Linguistics, Government College University, Faisalabad, Pakistan. Email: abdullahahmad63@yahoo.com

2. **Muhammad Asim Mahmood.** Dean, faculty of Arts and Social Sciences, Government College University, Faislabad, Punjab, Pakistan. He received his PhD degree in Applied Linguistics from Birmingham University, UK.

3. **Muhammad Ilyas Mahmood:** Head of English Department, University of Okara, Pakistan. He is a PhD Candidate at the Faculty of Education, UniversitiKebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia. Correspondence: ahmad453@yandex.com

RECIBIDO: 3 de julio del 2019. **APROBADO:** 12 de julio del 2019.