PUIG-SALABARRÍA S.C.

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.:40 Período: Agosto, 2019.

TÍTULO: Estudio basado en corpus sobre el perfil de vocabulario del lenguaje Shahmukhi Punjabi.

AUTORES:

1. Lect. Muhammad Farukh Arslan.

2. Ph.D. Muhammad Asim Mehmood.

3. M. Phil. Shaukat Hayat.

RESUMEN: Esta investigación es sobre el desarrollo del Perfil de Vocabulario con la ayuda de compilar un corpus de dos millones de palabras de Shahmukhi Punjabi. Un corpus de Shahmukhi Punjabi se transcribió a Gurmukhi Punjabi para el etiquetado de partes del habla. El corpus fue analizado con la ayuda de Antconc. La lista de frecuencias y la lista de diferentes ítemes de vocabulario según sus categorías gramaticales se estudiaron en el corpus desarrollado. Se ha observado que las palabras del idioma Punjabi tienen muchos casos y formas diferentes como contrarias al idioma inglés y similares al idioma Urdu. Los sustantivos, verbos y adjetivos varían según el número y el género. En el corpus también se encontraron abreviaturas y palabras de préstamo del idioma inglés.

PALABRAS CLAVES: adjetivo, adverbio, corpus, sustantivo, etiquetado POS, Shahmukhi Punjabi.

TITLE: Corpus Based Study on Vocabulary Profile of Shahmukhi Punjabi Language.

AUTHORS:

1. Lect. Muhammad Farukh Arslan.

2. Ph.D. Muhammad Asim Mehmood.

3. M. Phil. Shaukat Hayat.

ABSTRACT: This research is about the development of the Vocabulary Profile (VP) with the help of compiling a corpus of two million words of Shahmukhi Punjabi. A corpus of Shahmukhi Punjabi was transliterated to Gurmukhi Punjabi for parts of speech (POS) tagging. Corpus was analyzed with the help Antconc. Frequency list and the list of different vocabulary items according to their grammatical categories were studied in the developed corpus. It has been observed that the words of Punjabi language have many different cases and forms as contrary to English language and similar to the Urdu language. Nouns, verbs and adjectives vary according to number and gender. Abbreviations and loan words from English language were also found in the corpus.

KEY WORDS: adjective, adverb, corpus, noun, POS tagging, Shahmukhi Punjabi.

INTRODUCTION.

The research is a corpus-based study of a VP of shahmukhi Punjabi language. VP is a resource of lexical items which can be used in the field of language teaching, language learning, material development, syllabus design and evaluation (Capel, 2012). Nation (1990) considered VP as an important source for the purpose of language teaching and language learning. VP is a lexical resource which can play its vital role in language learning tasks.

VP of Shahmukhi Punjabi is studied in this research. Punjabi is an Indo-Aryan language; it is mostly spoken in the Punjab region of South Asia. 88 million people speak this language worldwide (Lewis, 2009), it is almost termed as the 13th most frequently spoken language in the world. Total of 110 million people, (66 million) in Pakistan, (44 million) in India and many millions in America, Canada and Europe have Punjabi as their mother tongue.

There are two scripts available two write Punjabi, Gurmukhi (in India) and Shahmukhi (in Pakistan). Shahmukhi Script is the basic concern of this research. Shahmukhi is a deviant form of the Perso-Arabic scripts. Script of Shahmukhi Punjabi is very much similar to Urdu Script. Urdu has 37 letters and Shahmukhi Punjabi contains 38 letters.

Corpus of two million 2(M) words of Shahmukhi Punjabi is collected for this research. Textbooks, Newspapers, short stories and an online source Apnaorg are kept under consideration for corpus compilation. Newspapers Published during 2017 are included in corpus.

In attempts of developing VP of Shahmukhi Punjabi language and to create lists of each Parts of speech, complete corpus was transliterated into Gurmukhi script to assign POS tags. After assigning POS tags the corpus was transliterated back to Shahmukhi Punjabi and VP is developed.

In this research, the researcher aims to find out the word list, frequencies of vocabulary items and grammatical categories of vocabulary items in the developed corpus. This research also includes POS tagging of a corpus of Shahmukhi Punjabi language. Research Questions are stated below:

1) What is the Frequency Profile of different vocabulary items found in this developed corpus?

2) What is the Frequency and Distribution of lexical items found in each grammatical category?

DEVELOPMENT.

Literature review.

VP is a vocabulary resource for teachers, teacher trainers, paper setters, materials writers and syllabus designers (Capel, 2012). An aggregation of the word frequencies is termed as a vocabulary profile (VP) (Graves, 2005). Large amount of vocabulary items is collected for research purposes, teaching and learning processes. VP includes words, phrases and phrasal verbs. VP can be developed and used for the enhancement of a language. VP helps in improving the status of a language, by developing a VP in the form of corpus we can study frequent lexical items used by a speech community, syntactic structures of that particular language by finding out the collocation patterns.

VP can be used for many other purposes like grammar checking, vocabulary testing, and evaluation of language learning and other fields of exploring a language. It provides an online source, including a vast amount of lexical items available for teaching and learning. In a research by (Yoon, 2012) VP was used for measuring the level of sophistication and proficiency in the target language. Dodigovic (2005) worked on VP for designing a course in English for academic or specific purposes (EAP/ESP).

Punjabi is an Indo-Aryan language; it is mostly spoken in the Punjab region of South Asia. 88 million people speak this language worldwide (Lewis, 2009), so it is almost termed as the 13th most frequently spoken language in the world. Another source states that 110 million people, (66 million) in Pakistan, (44 million) in India and many millions in America, Canada and Europe have Punjabi as their mother tongue. According to the Census (2001, p. 107), Punjabi is spoken by 44.15 % population of Pakistan.

Punjabi language is usually written in two scripts: Gurmukhi (in India) and Shahmukhi (in Pakistan). Shahmukhi script is the primary object of this research. Shahmukhi is a deviant form of the Perso-Arabic scripts. Due to this influence, it includes most of their innate qualities such as right to left writing direction, diacritic marks are used optionally, and the short vowels are not considered as letters of their own but placed above or below a consonant by using appropriate diacritics. There are 16 vowels, 16 diacritical marks and 49 consonants in Punjabi language (Malik, 2006).

POS tagging is a process in which POS tags are assigned to all the words given in the text after the word went through the process of morphological analysis and grammatical interpretation Garside

(1995). He considers morphological analysis and grammatical interpretation as an important part of the POS tagging.

According to Dash (2013), in the field of language technology, computational linguistics and NLP the task of POS tagging is associated with assigning particular POS tags to the lexical items on the basis of their grammatical category. This process is also known as grammatical annotation and word category disambiguation. POS tagging is a process of assigning grammatical categories, in this process words are marked with correspondence to their grammatical categories on the basis of their form and function in the context of syntactic pattern of a language.

Methodology.

A corpus of two million 2 (M) words of Shahmukhi Punjabi was collected. Corpus comprised upon Newspapers, News items, Novels, Books, Poetry, Short stories and Articles. Lokai and khabrain are the two Newspapers included in the corpus. Textbooks published by publisher are taken in computer readable form.

Shahmukhi Punjabi is written in Urdu script by using the software InPage. Corpus should be in notepad format to run in Antconc and analyze in the corpus. A software PakinPagetoUnicodeConverter was used to convert InPage files into Notepad files. Software (notepad ++) was used to change the encoding of the corpus. Encoding of Punjabi corpus was converted into UTF-8 to run that corpus in Antconc because corpus has to be analyzed with the help of AntConc to extract the results.

To extract POS categories, it was necessary to assign POS tags. Shahmukhi Punjabi did not have any POS tagger. Complete corpus was transliterated into Gurmukhi Punjabi to assign POS tags. After assigning POS tags corpus was re-transliterated to Shahmukhi with POS tags assigned to each word. Accuracy of those tags was not very much reliable tag of Unknown was assigned to those words which did not belong to the culture of Gurmukhi Punjabi.

Transliteration effected the orthography of Shahmukhi Punjabi. Letters within the words were seen variant form. Few patterns are seen regarding these variations. All these wrong scripted words were corrected manually.

Few words were seen with the different orthography and script style, as compared to the standard script of Shahmukhi Punjabi language. These words are described below; these words often vary on a pattern. Words containing خ letter were transformed into کھ due to the influence of Gurmukhi during the process of transliteration.

Orthography of Shahmukhi Words خ and کھ

کھفا	خفا
کھاتما	خاتما
مکھلوق	مخلوق
کھیال	خیال
کھلوس	خلوص
کھاتمے	خاتمے
ضکھم	زخم
تریکھ	تریخ
مکھدوم	مخدوم
کھیال	خیال

Few words in the text containing ظ،ز،ذ،ژ were replaced by ض during transliteration. Those words are stated in the table below.

Orthography of Shahmukhi Words ظ ، ز ، ذ and ض

لاضمی	لازمی
ضنانیاں	زنانیاں
درواضے	دروازے
پوضیشن	پوزیشن
ضلم	ظلم
ضیادتی	زیادتی
نضر	نظر
جایضا	جائزہ
مضاہمتی	مزاحمتی
ضکھم	زخم

Words of Shahmukhi Punjabi containing ح were changed into ہduring the process of transliteration from Shahmukhi to Gurmukhi and vice versa.

Orthography of Shahmukhi Word حand ہ

ہملے	حملے
ساہب	صاحب
مہنتی	محنتی
مہمدی	محمدی
ہقیقت	حقیقت
ہمضا	حمزا
ہرارت	حرارت
سورتہال	صورتحال
ہضرت	حضرت
ہیران	حیران

Words containing ک ، ق were also observed as wrong according to the standard orthography of Shahmukhi Punjabi language.

Difference in the orthography ک ،ق

متئلک	متعلق
تاکب	تعقب
اناکاد	انعقاد
شوک	شوق
مکابلے	مقابلے
تکریب	تقریب
قتب	کتب

Words containing ژ،ث ،ص were replaced by س during transliteration. All those words are stated in the table below.

Orthography of Shahmukhi Word ص ، ژ ، ث and س

استسواب	استصواب
غوس	غوث
انساف	انصاف
عناسر	عناصر
سبر	صبر
ملوس	ملوث
ساف	صاف
منسف	منصف
سرف	صرف
فاسلا	فاصلا

Letter ط in the words was transformed into ت automatically while transliterating Shahmuhki Punjabi into Gurmukhi and vice versa.

Orthography of Shahmukhi Word ط and ت

متابق	مطابق
تریقے	طریقے
متلوبا	مطلوبا
قہت	قحط
ترہاں	طرحاں
برتانیہ	برطانیہ
متلب	مطلب
اہتیات	احتیاط
غلتی	غلطی
تاقت	طاقت

Differences are seen in the letters such as ا،ع،ی،ئ،ء were transformed into each other in few words and those words are considered wrong by following the standard orthography of Shahmukhi Punjabi.

Difference in the orthography ا،ع،ی،ئ،ء

جایض	جائز
قایم	قائم
ارفان	عرفان
اسمت	عصمت
ناریباضی	نعریبازی
متنازا	متنازع
ایتماد	اعتماد
اضیم	عظیم
اضم	عزم
اناسر	عناصر

Tags were corrected manually by following the grammatical rules of Shahmukhi Punjabi and assistance of Punjabi speakers and experts. Wrong tags were corrected and replaced with correct tags by using the software NotePad ++. POS tags of 5000 most frequent lexical items were checked manually and frequency list of Vocabulary items was developed. Then, with the help of AntConc, frequency Profile and distribution of each grammatical category is developed.

CONCLUSIONS.

The Shahmukhi Punjabi language is written by following the script of Urdu and Arabic language, but like Urdu language Shahmukhi Punjabi language is written without using diacritical marks. Due to this influence, many words become ambiguous in the process of reading in this script. Such as, the word مل (meet), مل (mill), مل (price) and مل (rub).

Words having same orthography are once used as a verb and same words can be used as a noun or adjective in the other sentence, such as the word بولی (language) and بولی (speak), جان (life) and جان (go) and the word جیون (life) and جیون (to live). Context helps in determining that either that word is used as a noun, adjective or verb in that sentence.

Words of Punjabi language have many different cases and forms as contrary to English and similar to the Urdu language. Nouns, verbs and adjectives vary according to number and gender. The word (come) has many equivalents and replacement in Punjabi language, such as ندیاںٓ ۔ ا ،ندےٓا ،ندیٓا ،ندآ ا

Many abbreviations of English language were seen in the corpus. These words were written in Shahmukhi script, such as ئی سی سیآ ، این جی او ، ڈی پی او ، ایس ایچ او and پی سی بی .

Many loan words from English and Urdu were also seen in the corpus such as; match (میچ), university (یونیورسٹی) and professor پروفیسر. Those words did not have any equivalent in Punjabi language, so those words are used at a higher frequency and now have become part of Shahmukhi Punjabi language.

BIBLIOGRAPHIC REFERENCES.

1. Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. English Profile Journal, 3, e1.
2. Dash, N. S. (2013). Part-of-Speech (POS) Tagging in Bengali Written Text Corpus. International Journal on Linguistics and Language Technology, 1(1), 53-96.

3. Dodigovic, M. (2005). Vocabulary profiling with electronic corpora: A case study in computer assisted needs analysis. Computer Assisted Language Learning, 18(5), 443-455.
4. Garside, R. (1995). Grammatical tagging of the spoken part of the British National Corpus: a progress report. Spoken English on computer: transcription, mark-up and application, 161-167.
5. Graves, D. (2005). Vocabulary Profiles of letters and novels of Jane Austen and her contemporaries. A publication of the Jane Austen Society of North America, 26(1).
6. Lewis, M. Paul (ed.), 2009. Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. Online: http://www.ethnologue.com/.
7. Malik, M. G. Abbas., 2006. Punjabi Machine Transliteration. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp 1137-1144.
8. Nation, P. (1990). Teaching and learning vocabulary. New York: Heinle and Heinle.
9. Yoon, S. Y., Bhat, S., & Zechner, K. (2012, June). Vocabulary profile as a measure of vocabulary sophistication. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (pp. 180-189). Association for Computational Linguistics.

DATA OF THE AUTHORS.

1. Muhammad Farukh Arslan. Lecturer in English, department of Applied Linguistics, Government College University, Faisalabad, Pakistan. Email: ahmad453@yandex.com

2. Muhammad Asim Mehmood. Ph.D., Dean Faculty of Social Sciences, Government College University, Faisalabad, Pakistan. Email: masimrai@gmail.com

3. Shaukat Hayat. M. Phil., Lecturer in English, Government Degree College, Shujaabad, Multan, Pakistan.

RECIBIDO: 4 de julio del 2019. **APROBADO:** 17 de julio del 2019.