Developing English-Urdu Machine Translation Via Hindi

R. Mahesh K. Sinha
Department of Computer Science & Engineering
Indian Institute of Technology, Kanpur
Kanpur 208016 India
rmk@iitk.ac.in

Abstract

The paper presents a strategy for deriving English to Urdu translation using English to Hindi MT system. The English-Hindi lexical database is used to collect all possible Hindi words and phrases. These are further augmented by including their morphological variations and attaching all possible post-positions. This list is used to provide mapping from Hindi to Urdu. There may be change in gender and a word or a word group may be of multiple parts of speech. These are resolved using information available from English-Hindi MT. As Urdu is structurally very close to Hindi using similar post-positions, the output obtained is as acceptable as the Hindi translation.

1 Introduction

Hindi and Urdu are sister languages having common origin (Hock, 1991). They are structurally very close to each other and use similar post-positions, verb morphology as well as complex predicate verb structure. Can we exploit this similarity in developing English-Urdu translation from the English-Hindi machine translation (MT) system?. This paper is an attempt to address this problem. An obvious question that emerges is why not consider Hindi to Urdu machine translation as a separate task rather than limit it to the translation as obtained from English-Hindi MT system? If we take Hindi as the source language, we need to perform its grammatical analysis in terms of POS tagging, chunking, parsing and transformation of the source language. A statistical machine translation system (Brown et al., 1990) requires a large representative parallel corpus for Hindi-Urdu. Such parallel corpora are not available.

Yet another approach is to use an interlingua approach (Trujillo, 1999; Goodman,1989; Hutchins and Somers, 1992). However, the success of this approach depends upon the knowledge representation schema used. System's ability to capture adequate knowledge from the source language so as to be able to generate the target language text that is truly a translation of the source language, decides both the quality and accuracy of translation. A compromise is to use a pseudo-interlingua (Sinha, 2004) for a class of structurally similar languages. Here a text generator for each of the target language needs to be developed.

In this paper, we present a more straightforward strategy for deriving translation to Urdu from Hindi English-Hindi MT system without the need for developing a detailed grammatical analysis of the source language and without developing a full fledged text generator from the interlingua representation. We have taken an English-Hindi MT system (Sinha, 2004) for our case study. We primarily use lexical mapping of Hindi words to Urdu and transform the output sentence for gender agreement in case of dissimilarity in gender of the lexicons.

Megerdoomian and Parvaz (2008) use a similar idea to obtain Tajiki MT using an English-Persian system. One of the issues discussed there is the fact that compounds that are written separately in the Perso-Arabic script are attached to each other in the Cyrillic script of Tajiki. Such compositions are not encountered in case of Hindi-Urdu lexical

mapping. However, the substitutions may be for a group of nouns.

It should be noted here that there is no correlation in the scripts in which Hindi and Urdu are written. Hindi is written in Devanagari which is written left to right. Urdu script is based on Perso-Arabic alphabet with six additional characters primarily to map sounds of English and Hindi. 'Urdu' is short form of "zaban-i-urdu" meaning "language of the camp". Urdu as a language evolved during the Moghul rule in India when the soldiers in camps interacted with the local population and started mixing local language words in Arabic, Farsi and Turkish languages. Thus the difference in Hindi and Urdu languages is primarily the nature and origin of the words used. Urdu as a language has a number of regional variations. Urdu language purists tend to use more of Arabic-Farsi words. Hindi on the other hand is a Sanskrit based language. Often, Hindustani (Gilchrist, 1796) is referred to as a language which is in between Hindi and Urdu in terms of the degree of usage of words of different origin. In this work, our focus has been on Hindustani language written in Urdu script.

2 System Design

As pointed out earlier, in this study we have taken English to Hindi MT system developed by us. It uses a pseudo interlingua based rule-based approach. The input English sentences are transformed to an intermediate form called PLIL (pseudo lingua for Indian languages). A rule base is used for this transformation. The PLIL representation is then transformed to the target language using a text generator. Figure 1 shows a block schematic of the process at a broad level.

The English-Hindi MT system produces a Hindi translation. All grammatical analysis of English is performed by this translation engine. The parse structure of the input English sentence is transformed to the corresponding Hindi structure. The output Hindi sentence is generated from this structure.

Given below are a few examples of the English to Hindi translation process:

```
(1)
English Input:
He wants cold water.
PLIL:
<aff {sub_np ( he noun masculine singular third
[human] [वह:m 8] [] [] ) } {obj1_np ( cold ad-
jective positive [NIL] [ठंढा] [] [] ) ( water
noun neuter singular third [edible] [पानी :m 17]
[] [] ) } k1 {main_vp_active ( want verb_2 nor-
mal normal masculine singular third [चाह] 11 []
[]  ) } > . sviram
Hindi Output:
वह ठंढा पानी चाहता है
(2)
English Input:
He should get success in life .
PLIL:
<aff {sub_np ( he noun masculine singular third
[human] [वह:m 8] [] [] )} {pp ( life noun neu-
ter singular third [concept] [जीवन:m 8] [] [] )
(in prep [ in ] ) } {obj1_np ( success noun
neuter singular third [concept]
[सफलता:f 1] [] [] ) } k1 {main_vp_active (
get_5 verb_1 normal should masculine singular
third [पा] 1 []  []  ) } > . sviram
Hindi Output:
उसे जीवन में सफलता पानी चाहिए
```

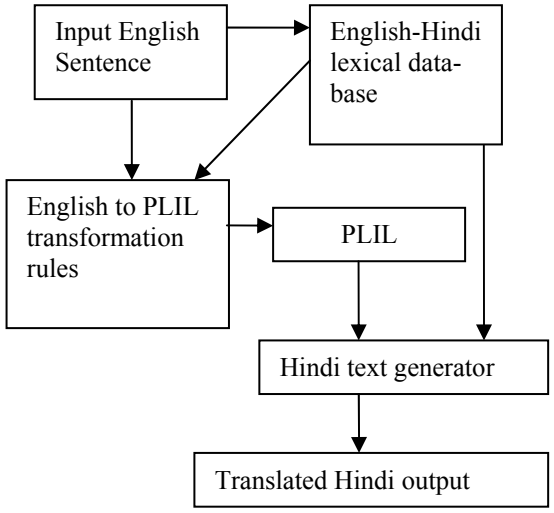

Figure 1. Block Schematic of English to Hindi MT

It is observed that even though Hindi and Urdu sentence structure remain the same, the individual words may differ in gender, number and may have multiple parts of speech (with different meanings). Moreover, a word may have multiple POS. Therefore, in order to derive the correct form of the Urdu sentence from Hindi with number and gender agreement, the English-Hindi MT system must also produce the POS, gender and number information for each of the Hindi word or word groups produced by the translation system. This informa-

tion is readily available in all rule-based MT systems. Figure 2 shows a block schematic of the translation process at a broad level that is relevant to Urdu sentence generation.

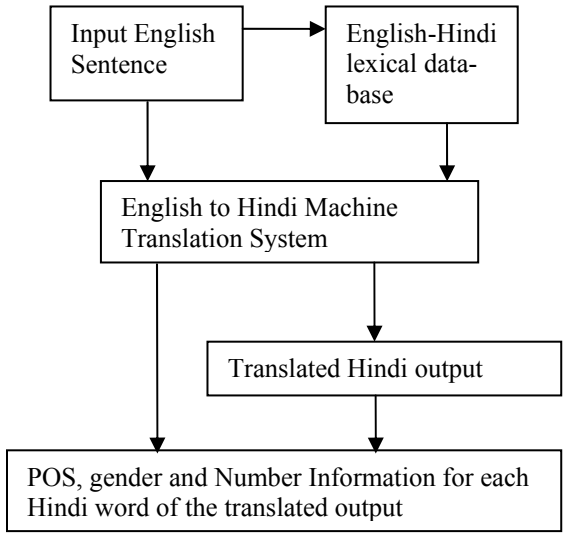

Figure 2. Block Schematic of English to Hindi MT

2.1 Hindi to Urdu mapping

The English to Hindi lexical database used in the above MT system is used to derive Hindi to Urdu mapping. The English-Hindi lexical database has syntactic, semantic and morphological information for each of the lexicon items with Hindi meaning. It also has information about constraints on the selection of a meaning. As Hindi and Urdu are structurally similar and use similar postpositions in the same order, the Hindi-Urdu mapping table needs to store only Urdu meanings along with only that information that affect the Urdu text composition. The mappings involving predicate verbs require special care in the choice of words. A predicate verb has a noun or an adjective/adverb followed by a light verb. Each of these constituent words in this case will have corresponding entries in the mapping table. In case a light verb of Hindi maps on to the same light verb in Urdu, no new entry needs to be made. However, if a Hindi predicate verb maps on to a non-predicate Urdu verb or a predicate verb with a different light verb (or vice-versa), then all combinations of morphological derivations have to be entered. For example, for the English verb 'achieve', the Hindi meaning from the lexical database is a predicate verb प्राप्त करना (پراپت کرنا) where प्राप्त(پراپت) is a noun and करना (کرنا) is a light verb. The Urdu mapping for the noun is हासिल (حاصل) and it already exists. The light verb करना(کرنا) remains as it is. Therefore, no fresh entry in the mapping table is required. Now, consider the English verb 'get' with Hindi meaning पाना(پانا). The corresponding Urdu mapping is हासिल करना(حاصل کرنا) which is a predicate verb. Now all forms of the Hindi verb will have to be entered into the mapping table. Some of these entries are shown in table 1. Similarly, for the English verb 'stop' or 'stall', the corresponding Hindi meaning is a Hindi predicate verb स्थगित करना (کرنا استھگت) and its corresponding Urdu mapping is a non-predicate verb रोकना(روکنا). Here also, all forms of the predicate verb will have to be entered.

A Hindi word may differ in number and gender and these are important for text composition. Besides this, a Hindi word may have multiple parts of speech (POS). In general, affixing some of the postpositions may affect the target text composition. In Urdu, this change is mostly due to change in gender, number or inflection of the associated word. For all words that get inflected, both inflected and non-inflected words are entered as options to the user for human post-editing. Similarly, for all adverbs, the postposition से(سے) is added as an option. One of these options are picked up by human post-editing. At present, our system is not capable of performing automatic selection.

The different morphological forms are automatically generated using paradigm numbers assigned to all the nouns, verbs and adjectives in the lexical database. Figure 3 depicts the process of creation of the Hindi-Urdu mapping table. As a large number of words of Hindi and Urdu (as spoken in India) are common, these entries are deleted from the mapping table. The entire process of creation of mapping table is automated and only corresponding Urdu entry is required to be done manually.

A set of sample entries of the mapping table for Hindi-Urdu is shown in table 1. This mapping table is directly used in construction of English to Urdu translation via English to Hindi translation.

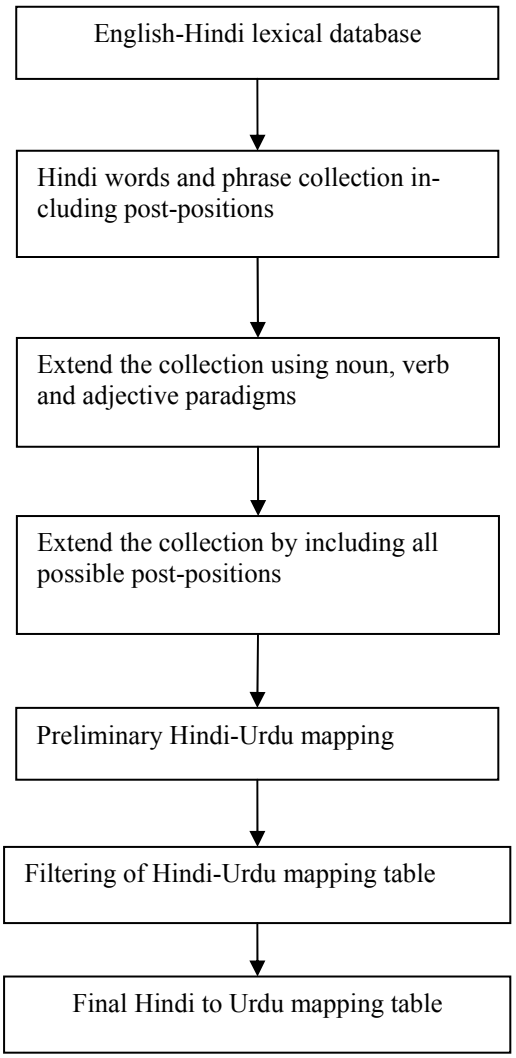

Figure 3. Creation of Hindi-Urdu mapping

2.2 The Translation Process

Figure 4 depicts the flow of the overall translation process. The parts of speech information for all the words of the translated Hindi sentences as obtained through the English-Hindi MT system are already available and are used by the POS resolution module. A stemming is performed for trimming the plural and feminine suffixes before the search is performed.

All the words and word groups are searched in the Hindi-Urdu mapping table. The search is made on the basis of maximal length string match. Matching for all the multi-word verbs (predicate and compound verbs) are performed by allowing the permissible particles such as नहीं (نہیں), भी (بھی), ही(ہی) or तो(تو), between them. The words that are not found in the mapping table are treated as unknowns and are transliterated into Urdu.

Hindi word				Mapped Urdu word				
	POS	G	N		POS	G	N	
आया	Verb	M	S	आया	Verb	M	S	آیا
आया	Noun	F	S	नौकरानी	Noun	F	S	نوکرانی
कारण	Noun	M	S	वजह से | { }	Noun	F	S	وجہ سے | {}
सफलता	Noun	F	S	कामयाबी	Noun	F	S	کامیابی
सफल	Adj			कामयाब	Adj			کامیاب
बलिदान	Noun	M	S	कुर्बानी	Noun	F	S	قربانی
जीवन	Noun	M	S	जिंदगी	Noun	F	S	زندگی
सुंदर	Adj			खूबसूरत	Adj			خوبصورت
क्रान्ति	Noun	F	S	इनक्लाब	Noun	M	S	انقلاب
स्वतन्त्रता	Noun	F	U	आज़ादी	Noun	F	U	آزادی
संकल्प-शक्ति	Noun	M	S	पक्का इरादा/पक्के इरादे	Noun	M	S	پکّا ارادا /پکّے ارادے
समय	Noun	M	U	वक्त	Noun	M	U	وقت
जन्मा	Verb	M	S	पैदा हुआ	Verb	M	S	پیدا ہوا
तृतीय	Adj			तीसरी / तीसरा				تیسری / تیسرا
सुन्दर	Adj			खूबसूरत				خوبصورت
विख्यात	Adj			मशहूर	Adj			مشہور
स्वतन्त्रता सेनानियों	Noun	M	P	आज़ादी के सिपाहियों	Noun	M	P	آزادی کے سپاہیوں
बुद्धिमान	Adj			ज़हीन	Adj			ذہین
पानी	Noun	M	U	पानी	Noun	M	U	پانی
पानी	Verb	F		हासिल करनी	Verb	F		حاصل کرنی
पाना	Verb	M		हासिल करना	Verb	M		حاصل کرنا

Table 1. Sample Hindi-Urdu mapping table

Whenever the mapping table indicates a change in gender, the gender change module changes the gender of the associated postpositions. Since the gender of the verb depends upon the gender of the subject or the object in both Hindi and in Urdu, it may also require changing gender of the verb. The gender change module performs these operations.

In figure 5 shows an illustration of different aspects of searching, POS resolution, substitution and gender change. In example (1), the gender of the words कारण(کارڑ) and वजह(وجہ) are different and

so the Urdu text generator changes the gender of के(کے) to की(کی). The word से(سے) /<null> is introduced as the post-edit option. In example (2), the word पानी(پیان) has two POS (verb/noun). Now the actual POS is checked with the POS of the output Hindi word as obtained from English-Hindi MT system and it is resolved to be a noun. However in example (3), the same word पानी(پیان) is resolved to be a verb using the information generated by the English-Hindi MT system and accordingly a substitution is performed. Similarly, in example (4), the word आया(آیا) is resolved to be a verb whereas in example (5), the same word आया(آیا) is resolved to be a noun.

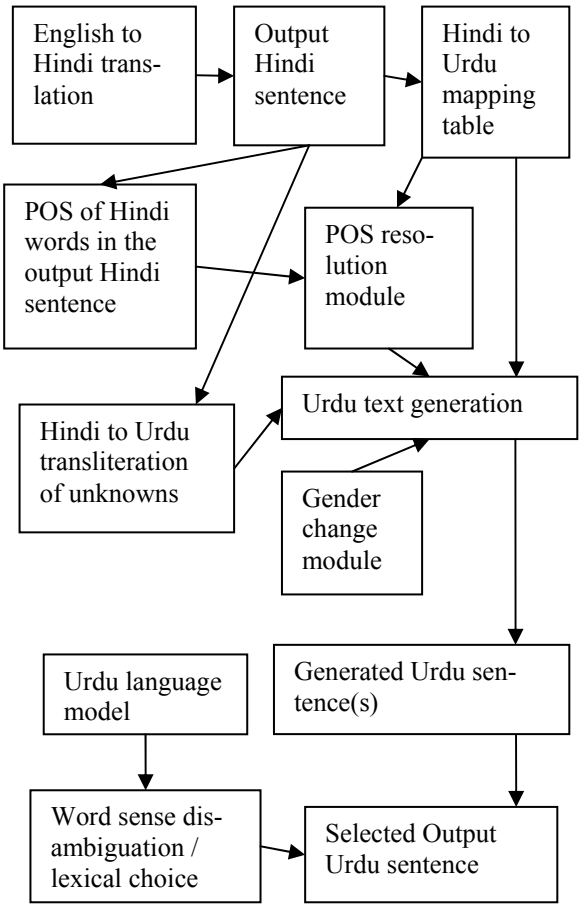

Figure 4. Translation System Flow diagram

(1) He did not go to school due to fever.
वह (وہ) ज्वर (جور) के (کے) कारण (کارڑ)
पाठशाला (پاٹھشالا) नहीं (ںیںہ) गया (ایگ) .
वह (وہ) बुखार (بخار) की (کی) वजह (وجہ) से (سے)
मदरसे (مدرسے) नहीं (ںیںہ) गया (ایگ) .

(2) He wants cold water.
उसे (اسے) ठण्डा (ٹھنڈا) पानी (یپان) चाहिये (ےچاہ)
उसे (اسے) ठण्डा (ٹھنڈا) पानी (یپان) चाहिये (ےچاہ)

(3) He should get success in life.
उसे (اسے) जीवन (ونیچ) में (مےں) सफ़लता (سفلتا)
पानी (یپان) चाहिये (ےچاہ)
उसे (اسے) ज़िन्दगी (یزندگ) में (مےں) क़ामयाबी (یابیکام)
हासिल (حاصل) करनी (یکرن) चाहिये (ےچاہ) .

(4) He has come to my house.
वह (وہ) मेरे (رےیم) घर (گھر) आया (آیا) है (ہے)
वह (وہ) मेरे (رےیم) घर (گھر) आया (آیا) है (ہے)

(5) She is maid of my house.
वह (وہ) मेरे (رےیم) घर (گھر) की (کی) आया (یا آ) है (ہے)
वह (وہ) मेरे (رےیم) घर (گھر) की (کی) नौकरानी (یںوکرانی)
है (ہے) .

Figure 5. Illustration of Urdu sentence generation from Hindi.

It is quite possible that the above process yields more than one translation as a word in the mapping table may have multiple meanings with the same parts of speech. We need to perform word sense disambiguation and a lexical choice in these cases. An Urdu language model is used for this purpose. A trigram model is constructed and used to compare different alternatives. However, due to inadequate size of the available Urdu corpus, the language model is found to be inadequate for the task. In this implementation we simply use the word frequency information for the lexical choice.

3 Experimentation and Results

As outlined in the preceding section, we first need to create the Hindi entries for the Hindi-Urdu mapping table. The number of such entries is about 400000. However, as the post-positions remain the same in Urdu, this number is about 100000. Entry of the corresponding Urdu meanings is a very time consuming task. This process is semi-automated. Urdu paradigm file is used to generate different

morphological forms from the root and entered based on correspondence with the tags associated with Hindi word. These are then manually validated. Sometimes alternative post-positions are entered in case of alternatives and the appropriate post-position is picked up by human post-editor. The POS resolution module and gender change module get gradually refined with experimentation. The system is currently undergoing testing and refinement. The quality of Urdu output translation is almost the same as that obtained for English to Hindi translation.

English Input:

```
In those days such a revolution had spread all over the country.
People were full of determination to win freedom.
It was that time when Bhagat Singh was born on September 28, 1907.
He was the third son of Sardar Kishan Singh and Vidyavati.
Bhagat was a lovely child.
His smile was charming.
People used to say that he would become very famous.
It was a family of freedom fighters and one or the other always would be in
jail.
Vidyavati herself had to look after the affairs of the family.
They were intelligent and brave and this made her forget her misery.
Bhagat Singh was her favourite.
```

Hindi Output :

उन दिनों में ऎसी क्रान्ति सम्पूर्ण देश में फैल चुकी थी.
लोग स्वतन्त्रता प्राप्त करने के लिए संकल्प-शक्ति से भरे थे .
यह उस समय था जब भगत सिंह सितम्बर 28 1907 को जन्मा था.
वह सरदार किशन सिंह और विद्यावती का तृतीय पुत्र था .
भगत एक सुन्दर बच्चा था .
उसकी मुस्कान मुग्ध कर रही थी .
लोग कहते थे जो वह बहुत विख्यात बनेगा .
यह स्वतन्त्रता सेनानियों का परिवार था और एक या दूसरा हमेशा जेल में होगा .
स्वयं विद्यावती को परिवार के कार्यों की देखभाल करना पड़ा .
वे बुद्धिमान और बहादुर थे और इसने उसे उसकी विपत्ति भुलाने के लिये मजबूर किया .
भगत सिंह उसकी मनपसंद था.

Urdu Output:

ان دنوں میں ایسی انقلاب پورے ملک میں پھےل چکی تھی
لوگ آزادی حاصل کرنے کے لئے پکّا ارادا سے بھرے تھے
یہ اس وقت تھا جب بھگت سنگھ ستمبر 28, 1907 کو پیدا ہوا
وہ سردار کشن سنگھ اور ودیاوتی کا تیسری / تیسرا بیٹا تھا
بھگت ایک خوبصورت بچّا تھا
اسکی مسکراہٹ دلکش کر رہی تھی
لوگ کہتے تھے جو وہ بہت مشہور بنیگا
یہ آزادی کے سپاہیوں کا کنبہ تھا اور ایک یا دوسرا ہمیشہ جیل میں ہوگا
خود ودیاوتی کو کنبہ کے کام کی دیکھ بھال کرنا پڑا
وے ذہین اور بہادر تھے اور اسنے اسے اسکی مصیبت بھلانے کے لئے مجبور کیا
بھگت سنگھ اسکی پسندیدہ تھا

Figure 6. Sample English-Hindi-Urdu Output

It should be noted that while working out the Hindi-Urdu mapping table, we have taken the commonly used Indian Urdu words (Ahmed, 1998) which may not be the same as spoken in Pakistan. Most of the errors in Urdu are due to gender mismatch and in predicate verb forms. There are transliteration errors due to phonetic differences in the way the names are written in Hindi and Urdu. This is also due to one-to-many mappings of some of the Hindi consonants to Urdu consonants.

Figure 6 shows a sample English input and the corresponding outputs in Hindi and Urdu. The average BLEU scores for Hindi and Urdu translations have been found to be 0.3412 and 0.3544 respectively.

This methodology, tries to preserve fidelity and intelligibility that is obtained for translation from English to Hindi. However, fluency is very much compromised primarily due to lexical choice. The fluency is expected to be better when it is directly translated from English. The lexical choice inappropriateness/errors in English to Hindi and the lexical choice in Hindi to Urdu mapping get multiplied. Never the less this is an interesting experiment that provides quick working translation.

4 Conclusions and Discussion

The paper presents a simple strategy for deriving Urdu translation using an English-Hindi machine translation. No part of speech tagging, chunking or parsing of Hindi has been used as would be required for any source language. Instead, the grammatical analysis of English provides all the necessary information needed for Hindi to Urdu mapping. It should be noted that this kind a system cannot be used for direct translation from Hindi to Urdu due to various ambiguous mappings that have to be resolved, In general, the issues concerning resolving translation divergence between Hindi and Urdu will have to be addressed.

Acknowledgments

Author is thankful to Saleem Siddiqui and Praveen Srivastava for experimentation and testing, and Vijendra Shukla and his group at CDAC Noida for preparing the mapping data.

The work is supported by a grant from Technology Development for Indian Languages (TDIL) program of Govt. of India.

References

A. Trujillo, 1999. Translation Engines: Techniques for Machine Translation, Springer, London.

Bholanath Tiwari, 2004. *हिन्दी भाषा (Hindi Bhāshā)*, Kitāb Mahal, Allahabad.

H. Somers, 1999. Review Article : Example-based Machine Translation, *Machine Translation*, 14(2) : 113-157.

Hans H. Hock, 1991. *Principles of Historical Linguistics*, Walter de Gruyter, Berlin–New York.

John Gilchrist, 1796. *Grammar of the Hindoostanee Language, or Part Third of Volume First, of a System of Hindoostanee Philology*, Chronicle Press,Calcutta.

K. Goodman, (ed.) 1989. Special issue on Knowledge-Based Machine Translation, I and II, *Machine Translation*, 4(1/2).

Karine Megerdoomian and Dan Parvaz, 2008. Low-density language bootstrapping: The case of Tajiki Persian. In Proceedings of *LREC 2008* (Language Resources and Evaluation Conference). Marrakech, Morocco, May 2008.

P.F. Brown,, J. Cocke, S. Della Pierta, V. Della Pierta , F. Jelinek , J.D. Lafferty, R.L. Mercer and P.S. Rossin. 1990. A Statistical Approach to Machine Translation, *Computational Linguistics,* 16 :79-85.

R.M.K. Sinha, 2004. An Engineering Perspective of Machine Translation: AnglaBharti-II and AnuBharti-II Architectures, *Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004)*,Tata Mc Graw Hill, New Delhi, pp. 10-17.

Rizwan Ahmed, 1998. *Hindi-Urdu Kosh,* Ramnarainlal Arunkumar Allahabad.

W.J. Hutchins and H.L. Somers, 1992. *An Introduction to Machine Translation*, Academic Press, London.