Scientific Journal of Impact Factor (SJIF): 4.72
e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Emerging Trends and Innovations in Electronics and Communication Engineering - ETIECE-2017

Volume 5, Special Issue 01, Jan.-2018 (UGC Approved)

Comparison Of English-Kashmiri Language In Context To Machine Translation

[1]Mir Aadil, [2]M. Asger

[1]Department of Computer Sciences Baba Ghulam Shah Badshah University
[2]Department of Computer Sciences Baba Ghulam Shah Badshah University

Abstract - *English is a language that belongs to Indo-European family and has a Latin script writing system while Kashmiri (koshur) on the other hand is from Dardic subgroup of Indo-Aryan family of languages and has a dominant Persian alphabet writing system. However, these contrasts matter very scantily while the language pair is subject to machine translation. The paper gives a brief comparison of the two languages based on their orthography, scripts, consonants, vowels, punctuation marks, grammar and a detailed comparison of language structure (Syntactic Vs Analytical), ambiguity, gender disagreement, Named Entity Recognition (NER), collocations, etc. The latter characteristics determine almost 90 percent whether a machine transition is fairly intelligible or incomprehensible, so there comprehensive study is the first step while attempting to machine translate any language pair.*

Keywords- *NLP; Machine Translation; English-Kashmir Language pair; Divergence.*

I. INTRODUCTION

Natural Language Processing involves modeling a natural language in different ways to make processing of these languages possible. Machine Translation (MT) in-fact was the first application of Natural Language Processing. At the beginning stages of MT as a study, mainly after Georgetown Experiment in 1954, highly optimistic views were common in research community that the problem of MT was to be solved fully in next five years [19]. However, as the study advanced, it was found that there were multiple issues and problems related with MT that needed to be solved some even before attempting translation. Some issues like Word Sense Disambiguation, Named Entity Recognition, Lexical Ambiguities, etc. are common for all language pairs. Most of the issues are targeted language-pair specific and needed to be addressed separately. Languages differ in syntax, pragmatics, word order, grammar and morphology. Each language-pair presents a unique set of differences with each other that are known as Divergences [9][10]. English and Kashmiri Languages are no exception. We consider English as the Source Language and Kashmiri as the Target Language for the study of divergences. These languages differ in script, grammar, word order, syntax and semantics. This paper attempts to provide a study of these differences. English and Kashmir languages are analyzed in the paper on the basis of general language differences like style, culture and background. Furthermore lexicon, syntax and pragmatic differences are also studied in this paper.

II. DIVERGENCES IN ENGLISH AND KASHMIRI LANGUAGES

3.1. General Differences

	English	Kashmiri
Family	Anglo-Frisian –Anglic	Indo-Aryan—Dardic
Native Speakers	400 million	7 million
Region	World-Wide	Jammu & Kashmir
Writing System	Latin	Perso-Arabic
Writing Direction	Left to Right	Right to Left
Status	International	Scheduled Language in India

Table 1. Basic Differences in English and Kashmiri.

No	Name		Transliteration	IPA	Isolated Glyph
1	الف	alif	ā, ʾ, –	/aː, ʔ, ∅/	ا
2	بے	be	b	/b/	ب
3	پے	pe	p	/p/	پ
4	تے	te	t	/t̪/	ت
5	ٹے	ṭe	ṭ	/ʈ/	ٹ
6	ثے	se	s	/s/	ث
7	جیم	jīm	j	/d͡ʒ/	ج
8	چے	če	č	/t͡ʃ/	چ
9	حے	he	h	/h, ɦ/	ح
10	خے	khe	kh	/kʰ/	خ
11	دال	dāl	d	/d̪/	د
12	ڈال	ḍāl	ḍ	/ɖ/	ڈ
13	ذال	zāl	z	/z/	ذ
14	رے	re	r	/r/	ر
15	ڑے	ṛe	ṛ	/ɽ/	ڑ
16	زے	ze	z	/z/	ز
17	ژے	ce	c	/t͡s/	ژ
18	سین	sīn	s	/s/	س
19	شین	šīn	š	/ʃ/	ش
20	صواد	swād	s	/s/	ص
21	ضواد	zwād	z	/z/	ض
22	طوئے	toʾe	t	/t̪/	ط
23	ظوئے	zoʾe	z	/z/	ظ
24	عین	ʿain	ā, ō, ē, ʿ, –	/aː, oː, eː,	ع

25	غین	gain	g	/g/	غ
26	فے	fe	f	/f, pʰ/	ف
27	بڑی قاف	baṛī kāf	q	/q/	ق
28	کیف	kef	k	/k/	ک
29	گاف	gāf	g	/g/	گ
30	لام	lām	l	/l/	ل
31	میم	mīm	m	/m/	م
32	نون	nūn	ñ	/ñ/	ن
33	واؤ	wāˀo	w, ū, ō, ɔ̄	/w, uː, oː, ɔː/	و
34	ہے	he	h	/h, ɦ/ or /ʰ, ɦ/	ہ
35	ہمزہ	hamzah	ˀ, –	/ʔ/, /Ø/	ء
37	چھوٹی یے	choṭī ye	y, ī, ā	/j, iː, aː/	ی
38	بڑی یے	baṛī ye	ɛ̄, e	/ɛː, eː/	ے
31	میم	mīm	m	/m/	م

***Table 2. Perso-Arabic Alphabet for Kashmiri*[2][12][20][21].**

No	Letter	Modern English Name	Modern English Pronunciation	ARPAbet
1	A	a	eɪ	EY
2	B	bee	bi	B
3	C	cee	si	S
4	D	dee	di	D
5	E	e	ə	EH
6	F	ef	ɛf	F
7	G	gee	dʒi	G
8	H	aitch	eɪtʃ	HH
9	I	i	aɪ	IH
10	J	jay	dʒeɪ	JH
11	K	kay	keɪ	K
12	L	el	ɛl	EL
13	M	em	ɛm	EM

14	**N**	**en**	**ɛn**	**EN**
15	**O**	**o**	**oʊ**	**OW**
16	**P**	**pee**	**pi**:	**P**
17	**Q**	**cue**	**kju**:	**Q**
18	**R**	**ar**	**ɑ**:**r**	**R**
19	**S**	**ess**	**ɛs**	**S**
20	**T**	**tee**	**ti**:	**T**
21	**U**	**u**	**ju**:	**UH**
22	**V**	**vee**	**vi**:	**V**
23	**W**	**double-u**	**dʌbəl.ju**:	**W**
24	**X**	**ex**	**ɛks**	**AX**
25	**Y**	**wy**	**waɪ**	**Y**
26	**Z**	**zed**	**zɛd**	**ZH**

***Table 2. Latin Alphabet for English*[12][14][25].**

3.1. Sentence Structure and Word Order

While translating a main effect remains of how well the sentence structure of the output target sentence matches the actual desired output. This usually depends on the word order and sentence type of source sentence. English has sentences usually of following types[17][24]:

1. Simple Sentence containing on independent clause.
   - I am happy.
   - You and me can have dinner together.
   - You met me after long time

2. Compound Sentence contains two independent clauses joined.
   - I am happy, so you and me can have dinner together.
   - You met me after long time, so I am happy

3. Complex Sentence (One or more Dependent Clauses)
   - I am happy, even though I don't make much money.
   - My friends were away while I was planning a surprise.

4. Compound Complex Sentence (3 or more clauses with at-least one dependent clause.
   - I'm happy, even though I don't make much money, but my friends are always complaining since I can't afford a big party.

English language has a Subject-Verb-Object Word Order while Kashmiri follows Subject-Object-Verb Word Order with V2 moment. Table 3 and 4 demonstrate English Word Order and Kashmiri Word Order respectively. Table 5 shows translation of a typical complex sentence and variations in word positions of the sentences[1][13][15][16].

Sentence	**The girl is eating apples.**			
Gloss	**Girl**	**is**	**eating**	**apples**
Parts	**Subject**	**Auxiliary**	**Verb**	**Object**
Translation	کۄر چھے ژونٹہ کھیوان			

***Table 3. Word Order of English Language: SVO* [24]**

Sentence	کور چھے ٹونٹھ کہیوان			
Transliteration	kuur	chhi	tsũũţh	khyevaan
Gloss	girl	is	apples	eating
Parts	Subject	Auxiliary	Object	Verb
Translation	The girl is eating apples.			

Table 4. Word Order Kashmiri Language: SOV with V2 word order **[3][22][23].**

Main clause + Subordinate Clause	I brought the girl who is eating apples.								
Gloss	=>	I	brought	that girl	=>	who	Is	eating	apples
Parts	Main clause =>	Sub	Verb	Object	Relative clause =>	Subject	Auxiliary	Verb	Object
Translation	میے ان سوہ کور یوس ٹونٹھ کہیوان چھے								
Transliteration	=>	mye	eny	swa kuur	=>	ywas	tsũũţh	khyevaan	chhi
Literal English	=>	I	brought	that girl	=>	who	Apple	eating	is

Table 4. Word Order for English and Kashmiri Language at Translation **[29][30].**

3.2. Source Language Ambiguity

For translation is the decoding of sense not mere substitution of words, ambiguities that arise in understanding Source Language input sentence can result in wrong translations. These ambiguities need to be studied and resolved before applying the sentence to translation. The main ambiguities that arise at source level are [4][5][6]:

3.2.1 Lexical Ambiguity

English language contains many words with multiple meanings depending on the context and way in which these are used. e.g; word "run" has 179 different meanings, "take" has 127, "break" has 123, etc. The mapping to actual translation mapping is achieved to some extend only by Contextual Rules in Example based MT, exhaustive Examples in Example based MT and by Frequencies and Language Models in Statistical MT [7][8].

3.2.2 Syntactic and Pragmatic Ambiguity

Ambiguities due to the structure and relation of words in Source Language sentence result in multiple meanings. This ambiguity translates the same English sentence into multiple Kashmiri sentences. e.g;

"We need a book or a copy and laptop for preparing paper" may mean that "a book is required to make some paper or "a copy and a laptop is required" or "a book and laptop is required"[9][27].

3.2.3 Pronoun/Anaphora Resolution

Anaphora or Pronoun may not be well resolved in input sentences and hence may result in wrong translation. Unfortunately this isn't resolved in Rule based or in Corpus Driven MT. The only promising solution seems to be an Intelligent Knowledge Based System that takes context into account. e.g;

"The Police threatened the terrorists while some of *them* lied down". Problem remains with understanding that who lied down "Police or terrorists"[10]

3.3. Cross Lingual Divergences

The divergences arise because of grammatical differences and diverse sense extraction of a sentence in different languages. These divergences vary with each language pair and are independent of methodology used for MT [11]

3.3.1 Categorical Divergences

It is the most common type of divergence between languages and arises due to difference in the representation of parts of speech between languages. Noun in Source Language may be represented as verb in Target Language and vice versa.

Similarly Noun can be represented as adjective and vice versa [19][26]. English and Kashmir languages also exhibit similar type of divergence in most of the sentences. e.g;
Verb → Noun
English: My mom loves me.
Kashmiri: میانہِ ماجہِ چھے میاڑٕ ماے۔
Transliteration: Maineh Majeh che main Maiye.

3.3.2 Conflational & Inflational Divergence.

Conflational divergences occur when multiple words in English are translated into single word in Kashmiri and Inflational divergences occur when a single word in English is translated as n-word phrase in Kashmiri [26].
English: He slipped away.
Kashmiri: ۔سُہ ژول۔
Transliteration:(su) Tsul.

3.3.3 Head Swapping Divergence

Head Swapping results when the role of main verb in English is diminished in Kashmiri and the role of modifier verb is promoted. It is not so common in the language pair however its effect can't be ruled out completely [28].
English: The Jehlum is flowing.
Kashmiri: ۔وۄتھ چھِیہ پکان۔
Transliteration: Wyeth che pakaan.

3.3.4 Thematic Divergence

Thematic differences represent the differences in the argument structure of verb while translating English Sentence to Kashmiri [29].
English: Why are you late.
Kashmiri: ۔توہہِ کیازِ گوۆ ژیر۔
Transliteration: Toye kyazi gov tsheer.

3.3.5 Honorific Divergence

This is the cause of plural inflectional elements that generate verb and the genitive noun to mark respect or honor in some languages. English has no honorific phenomenon while Kashmiri language exhibits quite enough honorifics [29][30].
English: He is my friend.
Kashmiri: ۔سُہ چھُ مےٚ دوس۔
Transliteration: Suh chu meh doos. (He is translated as "*Suh*", singular).
English: He is mv teacher.
Kashmiri: ۔تِم چھِ میانؠ وۄستاد۔
Transliteration: Tem cheh meh ustaad (He is translated as *"Tem",* plural).

3.4 Target Language Variations

Kashmiri Language when treated as Target Language presents some characteristics that affect the translation output [7][18]. These variations are commonly found in all languages and are given as:

3.4.1 Lexical Gap

For some of phrases in English an exact mapping in Kashmiri language is not available. The lack of exact translation mapping is known as Lexical Gap[28].
English: Good luck.
Kashmiri: ۔خوداے سُند فضل آسِنےٚ۔
Transliteration: Khudayih sund fazal aes ney. (Meaning: *May God bless you*).

3.4.2 Multiword Expressions

Idioms are translated as separate words when an exact mapping is not available. Believing that every English idiom and multiword expression shall have a mapping in training corpus is not realistic. So missing phrases and idioms when subjected to translation shall result in bad translations. e.g. It is *raining cats and dogs* shall result in a translation meaning cats and dogs are falling from clouds if there is no exact translation pair for "*raining cats and dogs"* meaning raining heavily. This type of divergence is found in all language pairs as each language pair has its own set of idioms and multiword expressions [19].

V. CONCLUSION

There are multiple divergence patterns between English and Kashmiri Language that can be analyzed for solutions. Some of these can be resolved up to some extend for all languages using conventional methods like developing exhaustive

training sets. However for resource scarce languages like Kashmiri, the application of these methods doesn't prove to be much fruitful. A tradeoff is required between the error rate and efforts to remove the divergences, as solving even single one of the divergences or removing its bad effect on translation fully is a big problem in itself. Still, the study is very necessary as minimizing the effect of some these divergences is already provided by conventional methods and can determine whether a translation is good enough for communication, assimilation or dissemination.

REFERENCES

[1] Bushra Jawaid and Daniel Zeman, "Word-order issues in english-to-urdu statistical machine translation", Number 95, pages 87–106, 2011, Praha, Czechia.
[2] Bhat, R, " A Descriptive Study of Kashmiri", Delhi: Amar Prakashan, 1987. Print.
[3] Bhatt, R.M., "Verb Movement and the Syntax of Kashmiri", Kluwar Academic Press: Dordrecht, 1999. Print.
[4] Dave, S., Parikh, J. and Bhattacharya, P., "Translation Technical Report", LAMP 88, 2002.
[5] Dorr, B., "Classification of Machine Translation Divergences and a Proposed Solution Computational Linguistics". 20 (4), 1994, 597–633.
[6] Dorr. Bonnie, J., "Machine Translation Divergences: A Formal Description and Proposed Solution", Computational Linguistics, 20:4, 1994, pp. 597--633.
[7] Dorr, Bonnie, J. and Nizar Habash,,. "Interlingua Approximation: A Generation-Heavy Approach",. In Proceedings of Workshop on Interlingua Reliability, Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002, Tiburon, CA, 2002, pp. 1—6.
[8] Dorr, Bonnie J., Clare R. Voss, Eric Peterson, and Michael Kiker, "Concept Based Lexical Selection", Proceedings of the AAAI-94 fall symposium on Knowledge Representation for Natural Language Processing in Implemented Systems, New Orleans, LA, 1994, pp. 21—30.
[9] Dorr, Bonnie J., Lisa Pearl, Rebecca Hwa, and Nizar Habash, "DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word-Level Alignment," Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002,Tiburon, CA, 2002, pp. 31—43.
[10] Dorr, Bonnie, J. and Nizar Habash, "Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation", In Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002,Tiburon, CA, pp. 84—93.
[11] Goyal, P., and Sinha., R.M.K.. "A Study towards English to Sanskrit Machine Translation system". SISSCL, 2008.
[12] Grierson, G.A., "A Dictionary of the Kashmiri language", New Delhi: B.R. Publishing Corporation, 1932. Print
[13] Hook, P.E. "Is Kashmiri an SVO Language?" Indian Linguistics 37 (1976): 137-142. Print
[14] Huddleston, R. "Introduction to the Grammar of English", Cambridge: Cambridge University Press, 1984. Print.
[15] Jawaid, B., Zeman, D., Bojar, O., "Statistical Machine Translation between Languages with Significant Word Order Difference". PBML, 2010.
[16] Kachru, B.B., "A Reference Grammar of Kashmiri", Urbana: University of Illinois, 1969a. Print.
[17] Kak, A.A., "Acceptability of Kashmiri-English Mixed Sentences: A Sociolinguistic Study", Diss. University of Delhi, 1995. Print.
[18] Kameyama, Megumi and Ochitani, Stanley Peters, "Resolving Translation Mismatches With Information Flow" Annual Meeting of the Assocation of Computational Linguistics, 1991.
[19] Koehn, P. "Statistical Machine Translation":, Cambridge University Press, 2010.
[20] Koul, O. N., "Linguistic Studies in Kashmiri", Bahri publications: New Delhi, 1977. Print.
[21] Koul, O. N., "Kashmiri Language, Linguistics and Culture: An Annotated Bibliography", Mysore: Central Institute of Indian Languages, 2000. Print.
[22] Koul, O. N., "Kashmiri in the Indo-Aryan Languages", Ed. G. Cardona. and D. Jain. London: Routledge, 2003. 895-952. Print.
[23] Koul, O. N. and K. Wali, "Modern Kashmiri Grammar", Delhi: Indian Institute of Language Studies, 2006. Print.
[24] Levin, B. "English Verb Classes and Alterations: A Preliminary Investigation", The MIT Press 1997.
[25] Lewis, Paul, M., Simons, G.F., Fennig., C.D. "Ethnologue: Language of the World". Seventeenth edition. Dallas, Texas: SILI, 2013.
[26] Sinha, RMK and Thakur, A. "Translation Divergence in English-Hindi MT EAMT", 10th Annual Conference, Budapest, Hungary, 2005.
[27] Shauq, S. "A Contrastive Study of Some Syntactic Patterns of English and Kashmiri with Special Reference to Complementation and Relativization." Diss. University of Kashmir, 1983. Print.
[28] Wali, K. and O. N. Koul. "Kashmiri: A Cognitive-Descriptive Grammar", London and New York: Routledge, 1997. Print
[29] Wani, S. H. "Formal and Functional Aspects of English-Kashmiri Code Mixing." Diss. University of Kashmir, 2004. Print.
[30] Wani, S. H. "CM Grammar and Relativized Constraints: A Study in Kashmiri-English Mixing." Ed. N.A. Dhar. Interdisciplinary Journal of Linguistics IJL.Vol.1. Srinagar, India: University of Kashmir, 2008.119-132. Print.

