Interdisciplinary Journal of Linguistics Volume [10] 2017, Pp.104-113

Divergence Patterns in Kashmiri –English Machine Translation: Revisiting Dorr's Classification

Sajad Hussain Wani *

Abstract

This paper presents an attempt to present an overall overview of the major types of divergence patterns that are encountered in translation between Kashmiri and English language pair. The paper has a twin aim of revisiting classification of divergence types as outlined by Dorr and outlining some of the new divergence patterns that have not been reported in literature till date.

Key words: Divergence Patterns, Machine Translation, Syntactic Divergence, Lexical Divergence, Inflections, Modals, Pronominals.

1.1. Introduction

The age of translation has been a point of discussion for considerable period of time. "Encyclopedia Americana" states that it is as old as the written language. Parts of the Sumerian Epic of Gilgamesh, among the oldest known literary works, have been found in translations into several Asiatic languages of the second millennium BCE. The Epic of Gilgamesh may have been read, in their own languages, by early authors of the Bible and of the Iliad. With the advent of computers, attempts have been made to computerize or otherwise automate the translation of natural-language texts (machine translation) or to use computers as an aid to translation (computer-assisted translation). Etymologically, translation is a "carrying across" or "bringing across".

Translation is sometimes referred as the most difficult task among the academicians. Translation is not a uni-dimensional process but is multidimensional interweaving linguistic, social, cultural and other factors. Translation is not only the transference of meaning between languages but is a negotiation between two cultures, between two mind sets and between two time periods. Translation can never reach to the level of absolute equivalence but is always accompanied by loss of some kind. The terms like absolute equivalence, near equivalence, linguistic loss, cultural loss are very common in the translation literature.

Machine translation, sometimes referred to by the abbreviation MT, is a sub-field of computational linguistics that investigates the use of computer software to translate text or speech from one natural language to another. At its basic level, MT performs simple substitution of words in one natural language for words in another. Using corpus techniques, more complex translations may be attempted, allowing for better handling of differences in linguistic typology, phrase recognition, and translation of idioms, as well as the isolation of anomalies.

Current machine translation software often allows for customization by domain or profession (such as weather reports) — improving output by limiting the scope of allowable substitutions. This technique is particularly effective in domains where formal or formulaic language is used. It follows that machine translation of government and legal documents more readily produces usable output than conversation or less standardized text. Improved output quality can also be achieved by human intervention: for example, some systems are able to translate more accurately if the user has unambiguously identified which words in the text are names. With the assistance of these techniques, MT has proven useful as a tool to assist human translators and, in

^{*} Department of Linguistics, University of Kashmir

a very limited number of cases, can even produce output that can be used as is (e.g., weather reports). In order to translate one language into another, one has to understand the grammar of both languages, including both morphology (the grammar of word forms) and syntax (the grammar of sentence structure). In order to understand syntax, one had to also understand the semantics and the lexicon (or 'vocabulary'), and even to understand something of the pragmatics of language use. The biggest challenge that is faced in the MT at present is the issue at the linguistic level and one such issue is the issue of divergence.

Divergence is a common phenomenon in translation between two natural languages. Typically, translation divergence occurs when structurally similar sentences of the source language do not translate into sentences that are similar in structure in the target language" (Dorr, 1993). As a consequence, dealing with divergence assumes special significance. Dorr (1993) categorizes translation divergences into two broad types. They are: (A) syntactic Divergences, (B) Lexical-semantic Divergences. They are further subcategorized as follows:

- (A) Syntactic Divergence: i. Constituent order divergence, ii. Adjunction divergence, iii. Preposition-stranding divergence, iv. Movement divergence, v. Null subject divergence, vi. Dative divergence and vii. Pleonastic divergence
- **(B)** Lexical-semantic Divergence: i. Thematic divergence, ii. Promotional divergence, iii.Demotional divergence, iv. Structural divergence v. Conflational divergence, vi. Categorial divergence and vii. Lexical divergence.

In Dorr (1994), she has examined the structure of the lexical-semantic divergences. This classification takes into account various sources of differences between a set of translation languages and captures a large sets of translation divergences. The classification is based on the Government and Binding framework (Chomsky 1986, Jackendoff 1990) of linguistic theory which assumes a deep structure to capture the surface structure variations.

The deep structure functions as the universal structure, i.e. applicable across languages. Thus both the classification and the resolution of the translation divergences are largely discussed from the perspective of the universal grammar. The classification captures the major grammatical issues in translation divergence across languages. However, it also misses a number of points that pertain to a particular set of translation languages.

The issue of divergence between a set of languages is associated with a number of factors ranging from linguistic to sociolinguistic and psycho-linguistic aspects of the languages involved. Although Dorr's classification takes into account many of the major linguistic factors associated with translation divergence, there still remains a number of points related to both linguistic and extra-linguistic factors that may exist in different sets of translation languages. Furthermore, the parameters of the classification does not take into account subtle semantic factors to the extent they are relevant for the classification of translation divergences in various languages.

In the existing literature, the issue of translation divergence for Hindi and English MT has not been exhaustively examined. Gupta et al (2003) and Dave et al (2001) discuss some of the translation divergences pertaining to English-Hindi MT and Hindi-English MT. Dave et al (2001) discusses the issue within the UNL-based Interlingua approach and only some of the obvious types of divergences have been discussed. These works do not explore further areas of divergence. Similarly other scholars like R.K. Sinha and Anil Thakur(2004) have also given a very exhaustive account of translation divergence. All these researches have stressed on the need to study divergence patterns in other language pairs so that the accuracy in machine translation should reach to it's target goal.

Divergence patterns need to be studied in great detail as capturing these divergence patterns can be very helpful in developing algorithms which will be very useful in the development of machine translation systems in future.

It is in this back drop that this paper deals with a general overview of divergence patterns for Kashmiri-English machine translation system. The paper shall begin with a description of various translation divergence types which have been described by Dor(1993)and shall look for the Kashmiri equivalents of these divergence types. This shall be followed by various issues which surround this classification and how this can be extended for Kashmiri and English language pair. The paper shall give an overview of some other divergence patterns which have not been covered in the earlier classification of divergence types. This paper intends to present an overall view of major divergence patterns which are found for Kashmiri English language pair. It should be noted that some of these divergence patterns have been brought to discussion for the first time and can be very useful for further research in this direction.

2.1 Methodology

This research paper forms part of a doctoral dissertation on morpho-syntactic divergence patterns in Kashmiri English machine translation. The methodology adopted is based on the principles of descriptive and structural Linguistics. The data has been collected from many informal and formal settings and data consists of about 30 hrs of recorded speech. Some examples have been self created to illustrate a particular divergence pattern.

3.1 Analysis

Divergence in Kashmiri English language occurs due to many structural differences between Kashmiri and English languages. The main differences between Kashmiri and English languages is that while English has lost many of it's inflections, Kashmiri language is highly inflectional in nature. Kashmiri has comparably a greater degree of flexibility when compared to English language. While English is a gender neutral language; Kashmiri has gender for all the animate and inanimate nouns and nouns as well as verbs are marked with suffixes showing the gender of subject and object. Following Dorr's classification, some of the examples of such types of divergence in Kashmiri English language pair can be presented as follows:

1.Thematic Divergence: The verbal object in one language becomes as the subject of the main verb in other language.

1.1 soni	chu	saja:d	tang	kara:n	K		
Soni	is	sajad	irritate	do			
Sajad irritates Soni.							

In the example 1.1, the verbal object in English *Soni* becomes the subject of the main verb in Kashmiri.

2. Promotional Divergence: The modifier is realized as an adverbial phrase in one language but as the main verb in other language.

1.2 ti:vi	chu	chala:n	K		
TV	is	run-prg			
TV is on					

In the example 1.2, English modifier *on* (an adverb) is realized as the main verb chala:n in Kashmiri.

3.Structural Divergence: The verbal object is realized as a noun phrase in one language and as a prepositional /postpositional phrase in other language.

1.3sajad	oes	kha:ndras	manz	тији:d	K	
Sajad	was	marriage-dat	in	present		
Sajad attended the marriage						

In the example 1.3, *the marriage* is the noun phrase but in Kashmiri it becomes post-positional phrase *Khandras manz* (in the marriage).

4. Conflational Divergence: The sense conveyed by a single word in one language requires at least two or more words of the other language.

1.4 tAm	kour	mye	shraptsi	si:t	hamli	K
He-erg	did	me	knife-abl	with	attack	
He stabbed me						

In the example 1.4, English word *stab* has no one-word equivalent in Kashmiri, and therefore many words have to be used to convey the same in Kashmiri and hence the conflational divergence.

5. Categorical Divergence: When translation results in the change in the category of a word, it is said to exhibit categorical divergence. For example, the verb in one language is realized as a noun in another language.

1.5mye	di	Khabar	K		
me	give	News			
Inform me					

In the example 1.5, English word *inform* (verb) is realized as a noun *khabar*(news) in Kashmiri language.

6. Lexical Divergence: The event is lexically realized as the main verb in one language but as a different verb in other language.

1.6 tim	gAy	kamras	manz	davi	Atsith	K	
They-erg	went	room-dat	in	run-prg	enter-pst-pc		
They run into the room							

In the example 1.6,the event is lexically realized as the main verb *run* in English but as *Atsith* a different verb (literally (to enter)) in Kashmiri, and *run* is used as participle.

7. Constituent Order Divergence:When translation results in the change of word order between two languages; constituent order divergence is said to take place. The constituent order divergence in Kashmiri English language pair is a very common phenomenon as words move

freely to a very considerable extent in Kashmiri due to it's highly inflectional nature. Consider the example 1.7:

1.7 mye	chu	pasand	gindun	K		
I-dat	is	like	playing	E		
I like playing						

The alternative translations of example 1.7 exemplifying constituent order divergence can be shown as:

mye	chu	gindun	pasand
gindun	chu	mye	pasand
gindun	chu	pasand	mye
chu	mye	pasand	gindun
chu	pasand	mye	gindun

It must be noted that the most natural order of the Kashmiri sentence is one where the auxiliary is at the second position of the sentence.

A close look at the morpho-syntactic differences between Kashmiri and English can reveal that Dorr's classification in no way is exhaustive for covering divergence in Kashmiri English machine translation. There are other types of divergence patterns which are very common in Kashmiri English language pair. Aspects related to tense, mood, number, gender, causatives, questions, honorifics, modals, indirect speech, mapping of time, case expressions are some of the other phenomena which need to be studied in detail before one can conceptualize about a Kashmiri English Machine translation system.

Some New Areas for Study in Kashmiri English Machine Translation: All the new areas which have been identified for the study of Kashmiri English machine translation require detailed studies as these areas are surrounded by a number of complexities which need elaboration and a detailed analysis. For example verbs in English usually inflect for tense. (Examples 2.1 and 2.2) and agree with the subject as in 3rd person singular as in example no. 2.3.

- 2.1 We play games.
- 2.2 We played games.
- 2.3 He plays games.

However, most verbs do not show any inflection for number and gender in English. Only verb "be" has more person-number distinction than other verbs. Thus, inflections in English language are comparatively reduced to minimum. This is, in fact, true in the light of the history of English language where a transition from old to modern English has proceeded with a great reduction in the number of inflections. Tensed verbs are prototypical verbs. Verbal nouns and adjectives are usually non-tensed. This is true in case of Kashmiri as well as a prototypical verb carries the tense inflection whereas verbal nouns and adjectives carry inflections which are untensed. E.g; asvun(smiling), natsvun(dancing), asun(smile), parun(reading) are all non tensed and are different from verbs which carry appropriate tense inflections for present, past, future in addition to the number and gender of the subject and object. Verbs have tensed forms. Most verbs usually have six inflectional forms in English. E.g.;

Tensed			Non-	Non-tensed		
Past			Base form	-ing form	-en form	
	3p-sng	General	4 -1 -1 -	5 al alia	(alamban	
1.shook	2.shakes	3.shake	4.shake	5.shaking	6.shaken	

Morphologically, the base form is the lexical stem, and the general present tense is identical with it. However, there are some exceptions to the above general inflection paradigm. E.g.; verb "be" has extra person-number forms: was and were instead of a single past tense form, am and are instead of the normal general present tense form. In contrast, the modal auxiliaries can, may, must, etc., have no non-tensed forms and have no contrast in the present tense between 3rd person singular and general forms. This lack of inflectional paradigm has been referred to as defectiveness. In addition to modal auxiliaries, there are a few defective verbs. e.g.; beware. It must be noted that the processes used in the formation of verb stems are fewer and typically less productive than those used in the formation of nouns.

Due to inflectional richness of Kashmiri language, verbs in Kashmiri exhibit different properties. Kashmiri verbs are highly inflectional in nature and inflect for number, tense, and gender.

2.4 su	khe	-yi	amb	K		
he-nom	eat	fut-sng-fem	mango			
He will eat the mango.						

2.5	su	chu	khe	-va:n	amb	K	
	he-nom	is	eat	-pr	mango		
	He eats a mango/ He is eating a mango						

2.6 tAm	khe	- y	mazda:r	amb	K	
he-erg	eat	pst-sng-fem	tasty	mango		
He ate a tasty mango						

The above examples (2.4,2.5,2.6) illustrate how the verb is inflected for the future, present and past tense respectively. In these examples basic verb root *khe "eat"* is inflected with different suffixes (-yi,-va:n,-y) to show the different tenses. It can be easily seen that when the verb is in the past tense; the subject assumes an ergative case. This is one of the differences between nominative-accusative English and ergative-absolutive Kashmiri language. The ergative case in Kashmiri occurs when the verb is in past tense and past perfect tense. The ergative subject takes an altogether different shape. This provides the first rule of divergence between Kashmiri and English where a nominative case form is reduced to ergative case form as in the example 2.6. The examples 2.7 and 2.8 represent the change that takes place in the verb with number. The suffixes yi "sng-fut" and yan "pl-fut" are added to the verbal root khe "eat" for showing the singular and plural number.

2.7 su	khe	-yi	dach	K	
he-nom	eat	fut-sng	grapes		
He will ed	at the grapes				

2.8	tim	khe	-yan	dach	K	
	They-nom	eat	fut-pl	grapes		
	They wi	ll eat the gra	pes			

2.9	tAm	khe	-y	ayis kri:m	K
	he-erg	eat	pst-sng-fem	ice cream	
		He ate	an ice cream		

2.10	tAm	khe	-ov	thu:l	K	
	he-erg	eat	pst-sng-mas	egg		
		He at	te an egg			

In the examples 2.9 and 2.10; the verb *khe "eat"* takes -y(fem) and -ov(mas) which directly correspond to the gender of the object *ice cream (fem)* and egg(thu:l(mas)).

Traditionally, verbs are classified as intransitive, transitive, causative, dative, conjunct, or compound in Kashmiri. Most intransitives mark their subjects in the nominative across all tenses. A few intransitives mark their subjects in the ergative in the past tense as is the case with transitive verbs. Transitive subjects are marked nominative in the non-perfective. Direct objects are marked nominative or dative as dictated by the person hierarchy. In the perfective, transitive subjects are marked ergative. Direct objects are in the nominative with the exception of la:yun 'to beat.' la:yun marks its direct object in the dative. Kashmiri transitive verbs may be basic or derived from intransitives. Basic transitives in Kashmiri are only few which may be simple or ditransitive. All ditransitives are basic in Kashmiri. Simple transitives take a direct object. It may be marked nominative or dative depending on the context as in examples 2.11 and 2.12. In example 2.11, the object *tso:nth(apple)* is marked nominative(unmarked in Kashmiri) whereas in 2.12, the object *Raja* is marked dative.

2.11 tAm	khy	-ov	tso:nth	K
he-erg	eat	pst-sng-mas	apple	
	Не с	ate an apple.		

2.12 tAm	lou	-у	ra:j	-as	K	
he-erg	beat	pst-sng-mas	Raj	dat		
	Н	le beat Raj.				

Ditransitives take two objects: direct, and indirect. Direct objects follow the pattern in the simple transitive. Indirect objects are marked in the dative as in example 2.13.

2.13 sajad	an dyut	shabir	-as	tohfI	K
Sajad e	rg gave.msg	Shabir	dat	present	
Se	ajad gave a present i	to Shabir.			

From the above discussion, it becomes clear that developing a Kashmiri English Machine Translation system shall require an exhaustive study of the morpho-syntactic systems of the given languages. Some of the areas which can be identified for the study in designing a Kashmiri English machine translation system include:

- 1. Translation across tenses.
- 2. Translating agreement patterns.
- 3. Translation across case forms.
- 4. Translating question types.
- 5. Translating pronominal expression.
- 6. Translating subject-less constructions.
- 7. Translating honorifies.
- 8. Translating across word classes.
- 9. Translating across verb classes.
- 10. Translating vocatives.
- 11. Translating modal systems.
- 12. Translating Passive structures.
- **4.1 Conclusion:** English and Kashmir languages have many similarities but at the same time, the differences are also very much noticeable. The two languages have many differences at phonological, morphological and syntactic levels. English has a fixed word order with accompanying loss of inflections. Kashmiri on the other hand has a free word order and is inflectionally a very rich language. This basic morpho-syntactic difference lies at the roots of morpho-syntactic divergence which occurs when one translates between Kashmiri and English languages. The divergence is noticeable when one starts translating simple sentences containing a subject and object. Different types of divergences were observed between Kashmiri and English languages as were observed by Dorr including addition, replacement, substitution, promotion, thematic and other types of divergence. However, the fact remains that divergence

has to be studied in much detail for the said pair of languages as there are many areas of study where a considerable amount of time and energy needs to be spent for the study of the divergence patterns. Many such areas have been outlined in this paper. From a preliminary study of the two languages, it appears that for designing a machine translation system for Kashmiri English language pair; a number of approaches need to be combined (a hybrid approach) as an approach based on examples alone or a rule based approach alone can't yield the required output.

List of Abbreviations:

Nominative case. Dative case . nom dat gen Genitive case. erg = Ergative case. poss = Possessive case. = Masculine mas Feminine. = Singular. fem sng Plural. = Present. plpr = Progressive. Past. prg pst Future. Postposition. fut pp = Kashmiri. K abl ablative pc participle = infinitive

5.1 References

Dorr, B. J. 1993. *Machine Translation: A View from the Lexicon*. MIT Press: Cambridge. Dorr, B.J. 1994. Classification of Machine Translation Divergences and a Proposed Solution. *Computational Linguistics* 20(4), pp.597-633.

Dorr, B. J., Jordan, P. W. and Benoit, J. W. 1998. *A Survey of Current Paradigms in Machine Translation*, Technical Report LAMP-TR-027, UMIACS-TR-98-72, CS-TR-3961. University of Maryland. College Park: USA.

Grierson, G.A.1932. *A Dictionary of the Kashmiri Language*. B.R. Publishing Corporation: New Delhi.

Gupta, D. and Chatterjee, N.2002a. *Study of Similarity and its Measurement for English to Hindi EBMT*. Proceedings of STRANS-2002. IIT Kanpur, pp.132-139.

Koul, Omkar N. and Wali, Kashi. 2009 . *Modern Kashmiri Grammar*. Delhi : Indian Institute of Language Studies

Sinha, R.M.K. 2002. *Translating News Headings from English to Hindi*. Proceedings of 6th IASTED International Conference on Artificial Intelligence and Soft Computing, ASC 2002. Banff: Canada.

Wani, Sajad H. 2010. Divergence in Kashmiri English Machine Translation: A Study in Passivisation Process, In *Interdisciplinary Journal of Linguistics IJL* Vol.3.University of Kashmir. pp.235-250.

......2011. Handling Divergence Patterns of Modals in Kashmiri English Machine Translation, *Interdisciplinary Journal of Linguistics IJL* Vol.4.University of Kashmir., pp 263-284

......2013. Divergence Patterns in Kashmiri –English Machine Translation: A View from Translation of Tenses, *Interdisciplinary Journal of Linguistics IJL* Vol.6.University of Kashmir, pp201-222.

......2014. Verb Word Class and Kashmiri –English Machine Translation Divergence, *Interdisciplinary Journal of Linguistics IJL* Vol.7.University of Kashmir, pp. 198-219.