This chapter gives a general introduction of the field of Natural Language Processing (NLP), Computational Linguistics and Machine Translation and how these fields are interrelated. This chapter then presents a discussion of different machine translation types. This is followed by a discussion of various machine translation systems developed in India. One of the very important areas of machine translation systems is the area of divergence studies. The present chapter then discusses this important field of research and gives an outline of studies which have been carried in India on Hindi, English and other Indian languages. This is followed by some of the representative divergence examples in Kashmiri-English language pair. The overall aim of this chapter is to provide a theoretical background for the study which is reported in the following chapters. By providing a theoretical background, this chapter links the present study with the broad field of NLP, Computational Linguistics and Machine Translation.

1.1 Introduction

Natural Language processing (NLP) is an interdisciplinary field combining insights from computer science and linguistics and is concerned with the interactions between computers and human (natural) languages. Natural language generation systems convert information from computer databases into readable human language. Natural language understanding systems convert samples of human language into more formal representations such as parse trees or first-order logic structures that are easier for computer programs to manipulate. Many problems within NLP apply to both generation and understanding, for example, a computer must be able to model
Chapter 1

Introduction

morphology (the structure of words) in order to understand an English sentence, and a model of morphology is also needed for producing a grammatically correct English sentence.

NLP has significant overlap with the field of computational linguistics, and is often considered a sub-field of artificial intelligence. The term natural language is used to distinguish human languages (such as English, Kashmiri or French) from formal or computer languages (such as C++, Java or Lisp). Although NLP may encompass both text and speech, work on speech processing has evolved into a separate field.

Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective. This modeling is not limited to any particular field of linguistics. Traditionally, computational linguistics was usually performed by computer scientists who had specialized in the application of computers to the processing of a natural language. In general, computational linguistics draws upon the involvement of linguists, computer scientists, experts in artificial intelligence, mathematicians, logicians, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists, among others.

Computational linguistics as a field predates artificial intelligence, a field under which it is often grouped. Computational linguistics originated with efforts in the United States in the 1950s to use computers to automatically translate texts from foreign languages, particularly Russian scientific journals into English. There is some dispute about who first had the idea of translating automatically between human languages, but the actual development of Machine Translation (MT) can be traced to conversations and correspondence between Andrew D. Booth, a British crystallographer, and Warren Weaver of the Rockefeller Foundation in 1947, and more specifically to a memorandum written by Weaver in 1949 to the Rockefeller Foundation which included the following two sentences:

“I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text.”
This memorandum sparked a significant amount of interest and research, and by the early 1950s, there was a large number of research groups working in Europe and the USA, representing a significant financial investment. But, despite some success, and the fact that many research questions were raised that remain important to this day, there was a widespread disappointment on the part of funding authorities at the return on investment that this represented, and doubts about the possibility of automating translation in general, or at least in the current state of knowledge. The theoretical doubts were voiced most clearly by the philosopher Bar-Hillel in a 1959 report, where he argued that fully automatic high quality machine translation (FAHQMT) was impossible, not just at present, but in principle. The doubts of funding authorities were voiced in the report which the US National Academy of Sciences commissioned in 1964 when it set up the Automatic Language Processing Advisory Committee (ALPAC) to report on the state of play with respect to MT as regards quality, cost, and prospects, as against the existing cost of, and need for translation. Its report, the so-called ALPAC Report, was damning, concluding that there was no shortage of human translators, and that there was no immediate prospect of MT producing useful translation of general scientific texts. This report led to the virtual end of Government funding in the USA. It led to a general loss of morale in the field, as early hopes were perceived to be groundless.

It was not until the late 1970s that MT research underwent something of a renaissance. There were several signs of this renaissance. The Commission of the European Communities (CEC) purchased the English-French version of the SYSTRAN system, a greatly improved descendent of the earliest systems developed at Georgetown University (in Washington, DC), a Russian-English system whose development had continued throughout the lean years after ALPAC, and which had been used by both the USAF and NASA. The CEC also commissioned the development of a French-English version, and Italian-English version. At about the same time, there was a rapid expansion of MT activity in Japan, and the CEC also began to set up what was to become the EUROTRA project. This was perhaps the largest, and certainly among the most ambitious research and development projects in Natural Language Processing.
The aim was to produce a ‘pre-industrial’ MT system of advanced design for the Languages of European Commission. In the late 1970s, the Pan American Health Organization (PAHO) began development of a Spanish-English MT system, the United States Air Force funded work on the METAL system at the Linguistics Research Center, at the University of Texas in Austin, and the results of work at the TAUM group led to the installation of the METEO system. For the most part, the history of the 1980s in MT is the history of these initiatives, and the exploitation of results in neighboring disciplines. As of 1990, METEO was regularly translating around 45,000 words of weather bulletins every day, from English into French for transmission to press, radio, and television. In the 1980s, the diesel engine manufacturers Perkins Engines was saving around £ 4,000 on each diesel engine manual translated. Moreover, overall translation time per manual was more than halved from around 26 weeks to 9-12 weeks — this time saving can be very significant commercially, because a product like an engine can not easily be marketed without user manuals.

Today different MT systems are functioning all over the world. Worldlingo provides machine translation using both statistical based and rule based systems. Other translation software include SDL Trados, Wordfast, Deja Vu, Swordfish and Alchemy CATALYST, most of them running under Microsoft Windows. A number of translation software programs are available free of charge, e.g. Foreign Desk and the multiplatform Okapi Framework and OmegaT+.

While no system provides the holy grail of fully-automatic high-quality machine translation of unrestricted text, many fully-automated systems produce reasonable output. The quality of MT is substantially improved if the domain is restricted and controlled. Despite their inherent limitations, MT programs are used around the world. Probably, the largest institutional user is the European Commission. Toggle text uses a transfer-based system (known as Kataku) to translate between English and Indonesian.

Google has claimed that promising results were obtained using a proprietary statistical machine translation engine. The statistical translation engine used in the Google language tools for Arabic-English and Chinese-English has an overall score of 0.4281
over the runner-up IBM's BLEU-4 score of 0.3954 (Summer 2006) in tests conducted by the National Institute for Standards and Technology.

With the recent focus on terrorism, the military sources in the United States have been investing significant amounts of money in natural language engineering. In-Q-Tel (a venture capital fund, largely funded by the US Intelligence Community, to stimulate new technologies through private sector entrepreneurs) brought up companies like Language Weaver. At present, the military community is interested in translation and processing of languages like Arabic, Pashto, and Dari. US Air Force has awarded a $1 million contract to develop a language translation technology.

The notable rise of social networking on the web in recent years has created yet another niche for the application of machine translation software in utilities such as Facebook, or instant messaging clients such as Skype, Google Talk, MSN Messenger etc, allowing users speaking different languages to communicate with each other. Machine translation applications have also been released for most mobile devices, including mobile telephones, pocket PCs etc. Due to their portability, such instruments have come to be designated as mobile translation tools enabling mobile business networking between partners speaking different languages, or facilitating both foreign language learning and unaccompanied traveling to foreign countries without the need of the intermediation of a human translator.

With proper terminology at work, with preparation of the source text for machine translation (pre-editing), and with re-working of the machine translation by a professional human translator (post-editing), commercial machine-translation tools can produce useful results, especially if the machine-translation system is integrated with a translation memory or globalization management system. Interactive translations with pop-up windows are becoming more popular. The main impact of MT in the immediate future will be in large corporate environments where substantial amounts of translation are performed.

1.1.1 Linguistic Issues in Machine Translation

Machine translation is basically a linguistic operation involving the linguistic structure of two given languages. 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. Difficulties involved in MT include lexical ambiguities, word sense disambiguation, translation of idioms and Phrases, structural ambiguities, syntactic parsing, and anaphoric references and discourse Structures. Discourse and discourse phenomena such as definiteness and topic/focus structure are particularly troublesome for translation. The same phenomena may be realized in different languages (and within the same language) using different grammatical resources. Conversely, values from one grammatical system (e.g.; word order, articles, and inflections) may realize different discourse phenomena in different languages.

These systems are further complicated by their interaction with syntactic phenomena such as the transitivity patterns of the verbs, and voice of sentences. Computation of equivalents for the values of such attributes as tense and modality is subjected to the same sorts of complexity as discourse features. Modality information is carried by a variety of Linguistic devices (verb inflections, modal verb systems, as part of the meaning of lexical verbs, etc.). Thus, Linguistic complexities form one of the biggest challenges to development of any machine translation system and hence a detailed linguistic study forms the basis of any machine translation project.

1.1.2 Approaches to Machine Translation

Two approaches have dominated MT research over most of the past twenty years. The first is the so-called interlingual approach, where translation proceeds in two stages, by analyzing input sentences into some abstract and ideally language independent meaning representation, from which translations in several different languages can potentially be produced. The second is the so-called transfer approach, where translation proceeds in three stages, analyzing input sentences into a representation which still retains characteristics of the original, source language text. This is then input to a special component (called a transfer component) which produces a representation which has characteristics of the target (output) language, and from which a target sentence can be produced. MT engines can be classified by their
architecture — the overall processing organization, or the abstract arrangement of its various processing modules.

Traditionally, MT has been based on direct or transformer architecture engines, and this is still the architecture found in many of the more well-established commercial MT systems. Newer indirect or linguistic knowledge (LK) architecture which, having dominated MT research for several years, are starting to become available in a commercial form.

The main idea behind transformer engines is that input (source language) sentences can be transformed into output (target language) sentences by carrying out the simplest possible parse, replacing source words with their target language equivalents as specified in a bilingual dictionary, and then roughly re-arranging their order to suit the rules of the target language. Linguistic knowledge (LK) architecture has dominated research in MT design during the past decade and is starting to appear in a number of commercial systems.

The idea behind LK engines is that the high quality MT requires linguistic knowledge of the source and the target languages as well as the differences between them. The term ‘linguistic knowledge’ refers to extensive formal grammars which permit abstract/relatively deep analysis. These also use an additional comparative grammar which is used to relate every source sentence representation to some corresponding target language representation — a representation which will form the basis for generating a target language translation. The LK engine will have grammars for each language it deals with. Researchers are concentrating on developing larger and better dictionaries, better grammars and analyzers, better rules for translation and so on. Others are trying to build large scale parallel corpora so that statistical Machine Learning techniques can be employed to automatically learn such rules. Work is also ongoing on developing a database of examples of translated pieces so that texts can be translated by analogy with already translated examples.

A Sublanguage approach relaxes the criterion of general-purpose translation by restricting the domain of application, or the complexity of the language. One example of a restricted-domain approach or sublanguage approach is TAUM/METEO, which
has been translating weather bulletins from English to French, completely automatically, for a number of years.

Statistical machine translation tries to generate translations using statistical methods based on bilingual text corpora, such as the Canadian Hansard corpus, the English-French record of the Canadian parliament and EUROPARL, the record of the European Parliament. Where such corpora are available, impressive results can be achieved translating texts of a similar kind, but such corpora are still very rare. The first statistical machine translation software was CANDIDE from IBM. Google used SYSTRAN for several years, but switched to a statistical translation method in October 2007. Recently, they improved their translation capabilities by inputting approximately 200 billion words from United Nations materials to train their system.

Example-based machine translation (EBMT) approach was proposed by Makoto Nagao in 1984. It is often characterized by its use of a bilingual corpus as its main knowledge base, at run-time. It is essentially a translation by analogy and can be viewed as an implementation of case-based reasoning approach of machine learning.

Hybrid machine translation (HMT) leverages the strengths of statistical and rule-based translation methodologies. Several MT companies (Asia Online and SYSTRAN) are claiming to have a hybrid approach using both rules and statistics. The approaches differ in a number of ways. One approach is that of rules post-processed by statistics. In this approach, translations are performed using a rules based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine. Another approach is statistics guided by rules in which rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating.

There are now many software programs for translating natural language, several of them online, such as Asia Online which provides a custom MT engine building capability that they claim gives near-human quality compared to the "gist" based quality of free online engines. Asia Online also provides tools to edit and create custom machine translation engines with their Language Studio suite of products.
1.2 Machine Translation in India: An overview

Human translation in India has a rich and ancient tradition. Works of philosophy, arts, mythology, religion, science and folklore have been translated among the ancient and modern Indian languages. Numerous classic works of art, ancient, medieval and modern, have also been translated between European and Indian languages since the 18th century. The earliest efforts on MT in India date from the mid 80s and early 90s. The prominent among these efforts are the research and development projects at Indian Institute of Technology (IIT) Kanpur, Computer and Information Sciences Department, University of Hyderabad, National Center for Software Technology (NCST), Mumbai (now, Center for Development of Advanced Computing (CDAC), Mumbai and Center for Development of Advanced Computing (CDAC), Pune. The Department of Information Technology, Ministry of Communications and Information Technology, Government of India, through its Technology Development in Indian Languages (TDIL) Project since 1990-91 and Department of Official Languages, Ministry of Home Affairs, Government of India, since 1989-90 have played instrumental roles by funding these projects.

The following domains have been identified for the development of domain specific translation systems: government administrative procedures and formats, parliamentary questions and answers, pharmaceutical information, legal terminology and important judgments, and so on. Since the mid and late 90’s, a few more projects have been initiated—at Indian Institute of Technology (IIT) Bombay, International Institute of Information Technology (IIIT) Hyderabad, Anna University – KB Chandrasekhar Research Center (AU-KBC) Chennai and at the Computer Science and Engineering Department, Jadavpur University, Kolkata. There are also a couple of efforts from the private sector like Super Infosoft Private Limited, and more recently, the IBM India Research Laboratory. A survey of the machine translation systems that have been developed in India for translation from English to Indian languages and among Indian languages reveals that the MT soft-wares are used in field testing or are available as web translation service. These systems are also used for teaching machine translation to the students and researchers.
In the current era, human translation finds application mainly in the administration, media and education, and to a lesser extent, in business, arts and science and technology. There is a big market for translation between English and the various Indian languages. MT is an important technology for localization, and is particularly relevant in a linguistically diverse country like India. Some of the prominent MT projects in India are briefly discussed as follows:

1.2.1 **Anglabharati (and Anubharati)**

The Anglabharti project was launched by R. M. K. Sinha at the Indian Institute of Technology, Kanpur in 1991 for machine aided translation from English to Indian languages. The system’s approach and lexicon is general-purpose with provision for domain customization. The system has been applied in several domains such as public health campaign, routine office-correspondence, technical-manual etc. The first prototype was built for English to Tamil in 1991 and later a more comprehensive system was built for English to Hindi translation.

Anglabharti is a pattern directed rule based system with context free grammar like structure for English (source language). It generates a ‘pseudotarget’ (Pseudo-Interlingua) applicable to a group of Indian languages (target languages) such as Indo-Aryan family (Hindi, Bengali, Assamese, Punjabi, Marathi, Oriya, Gujarati etc.), Dravidian family (Tamil, Telugu, Kannada & Malayalam) and others. A set of rules obtained through corpus analysis is used to identify plausible constituents with respect to which movement rules for the ‘pseudo-target’ is constructed. Within each group the languages exhibit a high degree of structural homogeneity. The similarity has been exploited to a great extent in the system. A language specific text-generator converts the 'pseudo-target' code into target language text. Paninian framework based on Sanskrit grammar using Karak (similar to the case) relationship provides a uniform way of designing the Indian language text generators. They also use an example-base to identify noun and verb phrasals and resolve their semantics. An attempt is made to resolve most of the ambiguities using ontology, syntactic & semantic tags and some pragmatic rules.

The unresolved ambiguities are left for human post-editing. Some of the major design considerations in design of Anglabharti have been aimed at providing a uniform
mechanism by which translation from English to majority of Indian languages with attachment of appropriate text generator modules becomes feasible. The translation system has also been interfaced with text-to-speech module and OCR input.

The English to Hindi version named Angla Hindi, based on the AnglaBharti machine aided translation system has been web-enabled. The technical know-how of this technology has been transferred on a non-exclusive basis to Center for Development of Advanced Computing (CDAC) for commercialization. The Anglabharti technology has also been transferred to eight different organizations under AnglaBharti Mission for development of Machine Aided Translation (MAT) systems for English to different Indian languages.

Sinha developed the Anubharti methodology in the year 1995 that follows the example based machine translation strategy. This methodology has been used for Hindi to English translation. The Anubharti approach works more efficiently for similar languages such as among Indian languages. Both of these system architectures, Anglabharti and Anubharti, have undergone a considerable change from their initial conceptualization. In 2004, phase-II of the system development has been launched which addresses many of the shortcomings of the earlier architectures. These are named AnglaBharti-II and AnuBharti-II. Both these systems are hybridized with varying degree of hybridization of different paradigms. AnglaBharti-II uses a generalized example-base (GEB) for hybridization besides a raw example base (REB).

1.2.2 Anusaaraka

The project originated at IIT Kanpur in 1995. The system provides both the robustness in case of failure and no loss of information while translating the text. The output of the system follows the grammar of the source language. The approach for the translation in this system is divided in two parts: 1) The Anusaaraka system which is based on language knowledge. 2) The domain specific knowledge based on world knowledge, statistical knowledge etc. The focus in Anusaaraka is not mainly on machine translation, but on Language Access between Indian languages. Using principles of Paninian Grammar (PG), and exploiting the close similarity of Indian languages, an Anusaaraka essentially maps local word groups between the source and
target languages. In case of languages having greater differences, the system introduces extra notation to preserve the information of the source language.

Thus, the user needs some training to understand the output of the system. The project has developed language accessors from Punjabi, Bengali, Telugu, Kannada and Marathi into Hindi. The approach and lexicon is general, but the system has mainly been applied for children’s stories. The project was later shifted mainly to the Centre for Applied Linguistics and Translation Studies (CALTS), Department of Humanities and Social Sciences, University of Hyderabad. It was funded by TDIL. Of late, the Language Technology Research Centre (LTRC) at IIIT Hyderabad is attempting an English-Hindi Anusaaraka/MT system.

1.2.3 Mantra

The Mantra project is based on the TAG formalism from University of Pennsylvania developed in 1999. It uses Lexicalized Tree Adjoining Grammar (LTAG) for representing the English and the Hindi Language. It uses the TAG for parsing as well as Generation purposes. In addition to translating the content, the system can also preserve the formatting of input word documents across the translation. The Mantra approach is general, but the lexicon/grammar has been limited to the sub-language of the domain.

A sub-language English-Hindi MT system has been developed for the domain of gazette notifications pertaining to government appointments. Recently, work has been initiated on other language pairs such as Hindi-English and Hindi-Bengali, as well as on extending to the domain of parliament proceeding summaries. Now this system is also used in the finance, agriculture, health care, information technology, education and the general purpose activities of the government domains. The system named Mantra Rajyasabha is developed for the Rajyasabha purposes. Currently the work for the language pairs English-Bengali, English-Telugu, English-Gujarati, Hindi-English, Hindi- Marathi, Hindi-Bengali is also in progress.

1.2.4 MaTra

MaTra is a Human-Assisted translation project for English to Indian languages, currently Hindi developed in 2004. It is essentially based on a transfer approach using
a frame-like structured representation. The focus is on the innovative use of man-
machine synergy—the user can visually inspect the analysis of the system, and
provide disambiguation information using an intuitive GUI, allowing the system to
produce a single correct translation.

The system uses rule-bases and heuristics to resolve ambiguities to the extent possible.
e.g., a rule-base is used to map English prepositions into Hindi postpositions. The
system can work in a fully automatic mode and produce rough translations for end
users, but is primarily meant for translators, editors and content providers. Currently,
it works for simple sentences, and work is on to extend the coverage to complex
sentences. The MaTra lexicon and approach is general-purpose, but the system has
been applied mainly in the domains of news, annual reports and technical phrases, and
has been funded by TDIL.

1.2.5 Google Translate

In 2007, Franz-Josef Och applied the statistical MT approach for Google Translate
from English to other Languages and vice-versa. Hindi and Urdu are the only Indian
Languages present among the 57 Languages for which Google Translate provides
translation. Accuracy of the system is good enough to understand the sentence after
translation.

1.2.6 Sampark

In 2009, a MT system among Indian Languages was proposed by the Consortium of
Institutions (IIT Hyderabad, University of Hyderabad, C-DAC Noida, Anna
University, KBC Chennai, IIT Kharagpur, IISc Bangalore, IIIT Allahabad, Tamil
University, and Jadavpur University). Currently released systems are {Punjabi, Urdu,
Tamil, Marathi} to Hindi and Telugu-Hindi and Telugu to Tamil MT system.

1.2.7 Web Based Hindi to Punjabi Machine Translator

Goyal and Lehal proposed a machine translator from Hindi to Punjabi in 2010. The
methodology used for the translation was direct translation and later on improving the
language learning modules for the enhancement of the quality of the system. The
accuracy of the translation is approximately 95%. The web tool developed has many
application areas like sending e-mails from Hindi to Punjabi or vice versa.
1.3 Machine Translation Divergence

Divergence is a common phenomenon in the 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:

1.3.7 Syntactic Divergence

This kind of divergence can be further sub-classified in the following categories:

a. Constituent order divergence,

b. Adjunction divergence,

c. Preposition-stranding divergence,

d. Movement divergence,

e. Null subject divergence,

f. Dative divergence and

g. Pleonastic divergence

1.3.8 Lexical-semantic Divergence

This type of divergence has the following subtypes:

a. Thematic divergence,

b. Promotional divergence,

c. Demotional divergence,

d. Structural divergence,

e. Conflational divergence,

f. Categorial divergence and

g. Lexical divergence.
Dorr (1994) examines the structure of the lexical-semantic divergences and proposed a LCS-based approach for their resolution. 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 remain 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. Among Indian languages, most of the works on divergence have been carried out on English-Hindi Language pair. Dave et al (2001) and Gupta and Chatterjee (2003), Sinha and Thakur (2004, 2005, 2008) are some of the representative works available on this topic. 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. Gupta and Chatterjee (2003) attempted to study MT divergence within the example based MT system approach. These works do not explore further areas of divergence. Other scholars like Sinha and Thakur (2004, 2005, and 2008) have given a very exhaustive account of translation divergence and have
formulated certain new divergence types for Hindi-English language pair. All these researches have stressed on the need to study divergence patterns in other language pairs so that the accuracy in machine translation should increase to a considerable extent.

Many other issues have been dealt with in the research on the divergence between Hindi-English and Sanskrit-English language pairs. The features unique to South Asian languages like echo formation, honorifics, replicative words and mixed languages are being studied by many scholars at present. What has emerged out of the above studies is that more work needs to be done on the divergence patterns in Indian subcontinent. Many issues which have not been addressed are immediately needed to be studied and rules for divergence patterns need to be formulated on a more exhaustive basis. Besides a large scale corpus needs to be developed in Indian languages so that machine translation systems incorporating both rule based and statistical systems will be developed.

1.4 An Introduction to Kashmiri-English Machine Translation Divergence

This study is the first attempt to study translation divergence for Kashmiri-English language pair and there is no available literature which discusses the issues related to translation between Kashmiri-English language pair. Divergence for Kashmiri-English language pair occurs due to many structural differences between Kashmiri and English languages. Translation Divergence between English and Kashmiri occurs due to their different morphological nature. English has lost most of its inflections and is an analytic language. Kashmiri language is highly inflectional in nature and has a rich case system. Kashmiri has comparatively a greater degree of flexibility in terms of word order and positioning of different elements in a sentence when compared to English language. While English is a gender neutral language, Kashmiri assigns 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 important examples of divergence in Kashmiri English language pair can be presented as follows:
1.4.7 Thematic Divergence

The verbal object in one language becomes as the subject of the main verb in other language.

<table>
<thead>
<tr>
<th>1.1</th>
<th>soni:</th>
<th>chi</th>
<th>hani:f-as</th>
<th>pasand</th>
<th>kara:n</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soni</td>
<td>is</td>
<td>Haneef.dat</td>
<td>like</td>
<td>do</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

‘Haneef pleases Sony.’

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

1.4.8 Promotional Divergence

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

<table>
<thead>
<tr>
<th>1.2</th>
<th>ti:vi:</th>
<th>chu</th>
<th>cala:n</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td>is</td>
<td>run.prg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

‘TV is on.’

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

1.4.3 Structural Divergence

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

<table>
<thead>
<tr>
<th>1.3</th>
<th>saja:d</th>
<th>chu</th>
<th>khandr-as</th>
<th>manz</th>
<th>muju:d</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sajad</td>
<td>is</td>
<td>marriage.dat</td>
<td>in</td>
<td>present</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

‘Sajad is attending the marriage.’

In the example 1.3, the marriage is the noun phrase in English but in Kashmiri it becomes a post-positional phrase khandras manz (in the marriage).
1.4.4 Conflational Divergence

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

<table>
<thead>
<tr>
<th>1.4</th>
<th>tAm</th>
<th>ko r</th>
<th>m’e</th>
<th>shrapc-I</th>
<th>sI:t</th>
<th>hamIL</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>He.erg</td>
<td>did</td>
<td>me</td>
<td>knife.abl</td>
<td>with</td>
<td>attack</td>
<td></td>
</tr>
</tbody>
</table>

‘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. Conflational divergence in Kashmiri-English language pair is commonly observed in the translation of verbs when a simple verb in English is replaced with a conjunct and compound verb in Kashmiri as Kashmiri is rich in having a number of compound and conjunct verbs which will be discussed in the forthcoming chapters.

1.4.5 Categorical Divergence

When translation results in a 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.

<table>
<thead>
<tr>
<th>1.5</th>
<th>m’e</th>
<th>dI</th>
<th>khabar</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>me</td>
<td>give</td>
<td>news</td>
<td></td>
</tr>
</tbody>
</table>

‘Inform me.’

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

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

<table>
<thead>
<tr>
<th>1.6</th>
<th>tim</th>
<th>gAy</th>
<th>kamr-as</th>
<th>manz</th>
<th>dav-I</th>
<th>Ats-ith</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>they.erg</td>
<td>went</td>
<td>room.dat</td>
<td>in</td>
<td>run.cp</td>
<td>enter.pst.ptc</td>
<td></td>
</tr>
</tbody>
</table>

‘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.

1.4.7. 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 its highly inflectional nature.

Consider the example 1.7:

<table>
<thead>
<tr>
<th>1.7</th>
<th>m’e</th>
<th>chu</th>
<th>pasand</th>
<th>natsun</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I.dat</td>
<td>is</td>
<td>like</td>
<td>dance.inf</td>
<td></td>
</tr>
</tbody>
</table>

‘I like dancing.’

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

<table>
<thead>
<tr>
<th>m’e</th>
<th>chu</th>
<th>natsun</th>
<th>pasand</th>
</tr>
</thead>
<tbody>
<tr>
<td>natsun</td>
<td>chu</td>
<td>m’e</td>
<td>pasand</td>
</tr>
<tr>
<td>natsun</td>
<td>chu</td>
<td>pasand</td>
<td>m’e</td>
</tr>
<tr>
<td>chu</td>
<td>m’e</td>
<td>pasand</td>
<td>natsun</td>
</tr>
<tr>
<td>chu</td>
<td>pasand</td>
<td>m’e</td>
<td>natsun</td>
</tr>
</tbody>
</table>
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 (KEMT). 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 tenses, 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 (KEMT) system. It becomes clear that developing a KEMT system shall require an exhaustive study of the morpho-syntactic divergence patterns present in the said language pair. It is this study of morpho-syntactic divergence patterns which forms the subject matter of the forthcoming chapters.