# Identification of Divergence for English to Hindi EBMT

**Deepa Gupta and Niladri Chatterjee**

Department of Mathematics
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi - 110016
India
{gdeepa, niladri}@iitd.maths.ernet.in

## Abstract

Divergence is a key aspect of translation between two languages. Divergence occurs when structurally similar sentences of the source language do not translate into sentences that are similar in structures in the target language. Divergence assumes special significance in the domain of Example-Based Machine Translation (EBMT). An EBMT system generates translation of a given sentence by retrieving similar past translation examples from its example base and then adapting them suitably to meet the current translation requirements. Divergence imposes a great challenge to the success of EBMT. The present work provides a technique for identification of divergence without going into the semantic details of the underlying sentences. This identification helps in partitioning the example database into divergence / non-divergence categories, which in turn should facilitate efficient retrieval and adaptation in an EBMT system.

## 1 Introduction

In an Example-Based Machine Translation (EBMT) (Brown, 1996) system translation for a given input sentence is generated by resorting to some past similar translation examples, and then modifying (*adapting*) them to suit the current requirements. The intuitive idea here is that structurally similar sentences in the source language (SL) should translate into sentences that are structurally similar in the target language (TL). However, this assumption is violated due to *divergence* which "arises when the natural translation of one language into another results in a very different form than that of the original" (Dorr, 1993). Although identification of divergence is important for various aspects of MT, such as word-level alignment (Dorr et. al., 2002), in this work we look at divergences from EBMT point of view in order that an EBMT system can identify divergence-prone sentences quickly, and can employ suitable retrieval and adaptation schemes for generation of the correct output.

Need to deal with divergence for adaptation in EBMT arises primarily because of two counts:

a) A sentence that upon translation gives rise to divergence may be difficult to adapt using past examples. This happens because the translation of such a source language sentence is structurally different from the translations of structurally similar source language sentences.
b) A retrieved translation example that involves divergence may not be helpful in generating the translation of a given input. This is because the retrieved translation may have a pattern that is not typical for the source language sentences of similar structures.

For illustration, let us consider the following English sentences (A) "She is in a shock", (B) "She is in trouble" and (C) "She is in panic". All the three sentences have similar syntactic structures. However, their Hindi translations have structural variations as discussed below.

The translation of sentence (A) is "*wah* (she) *sadme* (shock) *mein* (in) *hai* (is)". This the typical Hindi structure for these type of sentences. Hence although a preposition "*mein*" comes after the noun this is not to be considered as a divergence. In a similar way, the sentence (B) is translated into Hindi as "*wah* (she) *pareshanii* (trouble) *mein* (in) *hai* (is)". On the contrary, the Hindi translation for sentence (C) is "w*ah* (she) *dar* (panic) *rhaii* (..ing )

*hai* (is)", a structural deviation from the usual pattern, as the sense of the prepositional phrase (PP) "in panic" is realized by the verb "*dar rhaii hai*" ("is panicking"). Hence this is a divergence.

Now suppose the sentence (A) above is given as the input to an English to Hindi EBMT system. Two scenarios may be considered:

(i) The retrieved example is "She is in trouble". If this one is used for generating the translation, one may get a correct Hindi translation in a straightforward way.

(ii) If sentence (C) above is retrieved for an adaptation the generated translation may be "*wah* (she) *sadmaa* (shock) *rahii* (…ing) *hai* (is)", which is an improper Hindi sentence both syntactically and semantically.

Let us now look at the similarity of these two sentences with the input. Both sentences have the same structure except for the prepositional phrase (PP). For the input it is "in a shock", while for the two retrieved sentences these are "in trouble" and "in panic", respectively. Therefore, similarity between the sentences may effectively be measured by the semantic distance between "shock" and "trouble" in case (i), and the semantic distance between "shock" and "panic" in case (ii). According to Princeton University Wordnet 1.7 [http:// www. cogsci.princeton.edu/cgi-bin/webwn], "shock" and "trouble" have only one common ancestor (psychological feature); while "shock" and "panic" have two common ancestors (feeling and psychological feature). Thus the sentence (A) is semantically more similar to sentence (C) than sentence (B). Yet, the latter produces the appropriate translation for the EBMT system. This is because of the presence of divergence in the first case. Therefore, identification of divergence examples seems very important for an EBMT system.

Various approaches have been pursued in dealing with translation divergence. These may be classified into three categories (Habash and Dorr, 2002):

1. *Transfer approach.* Here transfer rules are used for transforming an SL sentence into TL by performing lexical and structural manipulations. The rules may be formed in several ways: by manual encoding (Han et. al., 2000), by analysis of parsed aligned bilingual corpora (Watanabe, 2000) etc.
2. *Interlingua approach.* Here, identification and resolution of divergence are based on two mappings (the GLR, the CSR) and a set of LCS parameters. In general, translation divergence occurs when there is an exception either to the GLR or to the CSR (or to both) in one language but not in the other. This premise allows one to formally define a classification of all possible lexical-semantic divergences that could arise during translation. UNITRAN (Dorr, 1993) uses this approach.
3. *Generation-Heavy Machine Translation (GHMT) approach* (Habash, 2002). This scheme works in two steps. In the first step rich target language resources, such as word-lexical semantics, categorical variations and sub-categorization frames, are used to generate multiple structural variations from a target-glossed syntactic dependency representation of SL sentences. This is the "symbolic-overgeneration" step. This step is constrained by a statistical TL model that accounts for possible translation divergences. Finally, a statistical extractor is used for extracting a preferred sentence from the word lattice of possibilities. Evidently, this scheme bypasses explicit identification of divergence, and generates translations (which may include divergence sentences) otherwise.

In our approach, we are neither using heavy TL resources as in GHMT, nor are we using language-independent LCS structures of the Interlingua approach. Rather we are comparing *functional tags* (FT) and the *syntactic phrasal annotated chunk* (SPAC) structures of SL and TL sentences for identification of divergence. Note that in this work SL and TL are English and Hindi respectively. Since Hindi is structurally similar to many other Indian languages (e.g. Bengali), a similar approach may be used for handling translation divergences from English to these languages as well.

Section 2. discusses some observations on English to Hindi translation divergences. Section 3 presents an analytical view of divergence identification, and provides algorithms for identification of two types of divergences in English to Hindi translations. This section also provides a critical evaluation of the algorithms.

## 2 Divergences in English to Hindi Translation: An Overview

Lexical-semantic divergences may be of seven types: *structural, conflational, categorial,*

*promotional, demotional, thematic* and *lexical* (Dorr, 1993). Of these, the "lexical" divergence is a mixture of more than one divergence type. In a more recent work (Dorr et. al., 2002), the divergence categories have been redefined in the following way. Under the new scheme six different types of divergences have been considered: *light verb construction, manner conflation, head swapping, thematic, categorial,* and *structural*. The difference in the two categorizations may be summarized as follows:

1. A *light verb construction* involves a single verb in one language being translated using a combination of a semantically "light" verb and another meaning unit (a noun, generally) to convey the appropriate meaning. In English to Hindi (and perhaps in many other Indian languages) context such happenings are very common. Hence this is not considered as a divergence for English to Hindi translation. One of our earlier works (Gupta and Chatterjee, 2003a) discussed this point in detail.
2. *Head swapping* essentially combines both *promotional* and *demotional* divergences under one heading.
3. *Lexical* divergence, which is a mixture of more than one divergence, has not been considered.
4. All other divergence categories remain as they are under the new scheme.

A critical analysis of these divergence categories put them in two broad classes:

- *Role-Preserving Divergence:* Here the roles of the functional tags of a sentence do not change upon translation. But morphological transformations of the constituent words/phrases are required to generate the correct translation. Structural and conflational divergences fall under this category.
- *Role-Changing Divergence:* Here, upon translation, the roles of some of the constituent words may change. Categorial, thematic, promotional and demotional divergences belong to this class.

We illustrate the proposed approach with the help of two divergences. In particular, we deal with structural and categorial divergences with respect to English to Hindi translation. These divergence types belong to role-preserving and role-changing classes respectively. In the following subsections these two divergences are discussed briefly with respect to English to Hindi translations. Understanding these divergences require some knowledge of Hindi grammar. One may refer to (Kellog, 1965), (Kachru, 1980) for details.

### 2.1 Structural divergence

This divergence occurs when the verbal object is realized as noun phrase (NP) in the source language, and as a prepositional phrase (PP) in target language. In the context of English to Hindi translation this divergence occurs in the same way. Consider for example the following translation "Andre will marry Steffi "~ "*andre* (Andre) *steffi* (Steffi) *se* (with) *vivaah* (marriage) *karegaa* (will do)". Here "Steffi", the object, is a NP that is mapped to the PP "*steffi se*" (literal translation "with Steffi") in Hindi. Since in Hindi a preposition comes after noun/pronoun, the roles of the functional tags remain unchanged upon translation. Thus structural divergence belongs to the role-preserving class.

### 2.2 Categorial divergence

Definition of this divergence is slightly different in the context of English to Hindi translation from its usual definition as given in (Dorr, 1993). As explained in one of our earlier works (Gupta and Chatterjee, 2003a) this divergence occurs when a subjective complement (SC) upon translation is realized as a verb in Hindi. This happens irrespective of the type of the SC in the corresponding English sentence. In particular, there we have shown examples where the SC of the English sentences may be of one of the following types: adjective, noun, adverb or prepositional phrase. Examples for these four sub-types of categorial divergence are given below.

.(A) *Adjective ~ Verb*: Consider, for example, the following English sentence and its Hindi translation: The patient was dead → *rogii* (patient) *mar gayaa* (died) *thaa* (did). The adjective of the English sentence "dead" is realized in Hindi by the verb "*mar jaanaa*" ("to die"). The past form of the verb along with the auxiliary verb "*thaa*" gives its past indefinite form.

(B) *Noun ~ Verb*: The following example illustrates this sub-type: Tom is an occasional listener of Jazz → *tom* (Tom) *kabhii kabhii* (occasinally) *jazz* (jazz) *suntaa* (listen) *hai* (is). Here the focus is on the word "listener"

which is a noun and has been used as the SC in the above English sentence. Upon translation it gives the main verb "*sunnaa*" ("to listen") of the Hindi sentence.

(C) *Adverb ~ Verb*: Consider the English sentence "The meeting is over". Its Hindi translation is: *sabhaa* (meeting) *samaapt ho gayii* (finished) *hai* (has).The main verb of the Hindi sentence is "*samaapt ho jaanaa*" i.e "to finish". Its declension for past indefinite is "*samaapt ho gayii*". However, this is not the main verb of the English sentence. Rather, its sense comes from the subjective predicative (adverb) "over" of the English sentence.

(D) *Prepositional Phrase* (PP) *~ Verb:* In the English sentence "She is in tears", "in tears" is a PP. In the corresponding Hindi translation its sense is realized by using the verb "*ronaa*" ("to cry"). Thus the Hindi translation for the above sentence is: "*wah ro rahii hai*", whose literal meaning is "She is crying".

Note that under this category, we are not considering other parts of speech variations, such as subjectival adjective to noun or PP. In this regard our observations are as follows:

(a) No example has so far been found where a subjectival adjective has been realized in Hindi as a prepositional phrase.

(b) There are instances where subjectival adjective becomes noun in the target langauge sentence. However, in such a case the realized noun becomes the subject of the target language sentence. The subject of the source language sentence becomes an object in the target language sentence. See (Gupta and Chatterjee, 2003b) for further details.

Section 3 provides algorithms for identification of divergences for an EBMT sytem. Algorithms have been developed for all the divergence types, but due to limitation of space we restrict our discussion to structural and categorial divergences only.

## 3 Identification of Divergence

Given an input English sentence and corresponding Hindi sentence, the proposed technique aims at identifying occurrence of divergence, if any, in the translation. When a divergence occurs, the algorithm also points out the place of occurrence. The technique uses *functional tags* (FT) and *syntactic phrase annotated chunk* (SPAC) of both the source language sentence and its translation. For the present discussion, the FTs that have been used are: subject (S), object (O), verb (V), subjective complement (SC) and modifier (M). The SPAC categories considered are: noun (N), adjective (Adj), verb (V), auxiliary verb (AuxV), preposition (P), adverb (ADV) and determiner (DT). The N, Adj, V, ADV and P are called the "lexical heads" of the phrases. For each category a suffix "P" is used to denote a phrase.

For illustration, we consider the sentence "Ram moved hurriedly to the car". The SPAC of this sentence is the following:

$[_{NP}$ [ Ram / N]] $[_{VP}$ [moved / V]] $[_{ADVP}$ [ hurriedly /ADV]] $[_{PP}$ [ to / P] $[_{NP}$ [ the /DT] [ car / N]]]

Hindi translation of the above sentence is "*ram* (ram) *jaldii se* (hurriedly) *gaadi* (car) *kii taraf* (to) *badaa* (moved)" having the following SPAC:

$[_{NP}$ [ ram /N]] $[_{ADVP}$ [ *jaldi se* / ADV ] ] $[_{PP}$ $[_{NP}$ [*gaadi*./ N]] [ *kii taraf* / P ]]$[_{VP}$ [ *badaa* / V ]]

This representation will be followed for all subsequent examples given in this paper.

### 3.1 Analytical view of the proposed algorithm

Here we present the theoretical background of the proposed technique.

Let L be a language. A *language* over a set A of words (of some lexicon) is defined as a collection of valid sentences according to the underlying grammar. Let F be the set of the funtional tags of L and F′ be the set of finite multi-subsets of F. A multi-set is a set where repetition of elements is allowed. If there is no ambiguity, hereafter we shall use the term "functional-tag set" instead of "functional tag multi-set". For example, {S, V, O}, {S, V, O, O}, {S, V, M} are valid functional-tag sets and are therefore members of F'.

Let ρ be a mapping from a language to the set of its functional-tag sets i.e. ρ: L $\rightarrow$ F′, where ρ ($x$) = the set of functional tags of $x$, for all $x \in$L. Therefore, ρ ($x$) may be considered as an ordered set $\{f_1, f_2, f_3, \ldots f_n\}$, say, of FTs. If $x$ is of the form $x_1x_2x_3\ldots x_n$ , where each $x_i$ is a word or a group of words, then $f_i$ is the functional tag of $x_i$. For example, if L is English and $x \in$L is the sentence "Ram moved hurriedly to the car", then ρ ($x$) = {S,

V, M, M}∈ F′. Similarly, for Hindi, if $x$ is " *ram jaldi se gadii kii taraf badhaa*", then ρ ($x$) = {S, M, M, V}∈ F′.

We define another map $\sigma_x$ in a similar way that gives for a sentence $x$ its SPACs from its functional-tag set. Let P be the set of all possible phrase structures of a language L. Let P′ be the set of finite multi subsets of P. Then $\sigma_x$: ρ(x) → P′, such that $\sigma_x(f_i)$ = the phrase set of $x_i$ for i ∈ {1,2,…..n}, where $f_i$ is the functional tag of $x_i$.

For illustration, consider the sentence "Ram moved hurriedly to the car". It has four functional tags $f_1$, $f_2$, $f_3$, $f_4$, where $f_1$ is S (Ram), $f_2$ is V (moved), $f_3$ is M (hurriedly) and $f_4$ is M (to the car). The SPACs corresponding to the functional tags are follows: $\sigma_x(f_1)$ = [NP [Ram /N]], $\sigma_x(f_2)$ = [VP [ moved / V]], $\sigma_x(f_3)$ = [ ADVP [ hurriedly /ADV]] and $\sigma_x(f_4)$ = [ PP [ to/P] [NP [ the /DT] [ car /N]]]. Similar mappings can be defined for any language L.

Given a parallel database of two languages $L_1$ and $L_2$, say, the divergence examples may be identified as follows. Since $L_1$ and $L_2$ are translations of each other, there is a bijection map ξ: $L_1 \rightarrow L_2$ such that for each $x \in L_1$, ξ($x$) is the translation of $x$ in $L_2$. Let $\rho_1$ and $\rho_2$ be functional tags mapping on $L_1$ and $L_2$ respectively.

1) If $\rho_1(x) \neq \rho_2(\xi(x))$ then it is a *role-changing divergence* in the translation of $x \in L_1$ to $L_2$.
2) In case $\rho_1(x) = \rho_2(\xi(x))$, we consider their funtctional-tag sets arranged in the same order. Let $\rho_1(x) = \{f_1, f_2, f_3, \ldots\ldots f_n\}$ and $\rho_2(\xi(x)) = \{g_1, g_2, g_3, \ldots\ldots g_n\}$, such that $f_i = g_i \; \forall \; i \in \{1,2,\ldots..n\}$. By "same order" we mean that if '$f_j$ ' corresponds to some word/words in a sentence $x$ which is kept in the $j^{th}$ position in the set $\rho_1(x)$, then the functional tag of the corresponding translation word/words in ξ($x$) should also be put in the $j^{th}$ position in the set $\rho_2(\xi(x))$. This process will avoid the confusion occurring due to multiplicity of a particular element in $\rho_1(x)$ or $\rho_2(\xi(x))$. If $\rho_1(x) = \rho_2(\xi(x))$, but $\sigma_x(f_i) \neq \sigma_{\xi(x)}(g_i)$ for some i ∈ {1, 2, .., n}, it is a *role-preserving divergence*.

The above technique has been used to develop algorithms for identifying different divergences. In particular, structural and categorial divergences have been used for illustration of the technique. The algorithms presented below return zero if the corresponding divergence are not found. It returns a number $k$ if divergence of sub-type k is detected. Note that structural divergence has only one sub-type, whereas categorial has four sub-types with respect to English to Hindi translation. In this respect, one may refer to sections 2.1 and 2.2 for definitions of different sub-types. In our discussion $L_1$ and $L_2$ are English and Hindi respectively, and divergence is being checked for sentences $x \in L_1$ and $\xi(x) \in L_2$.

### 3.2 Identification of structural divergence

Structural divergence deals with the objects of both $x$ and ξ($x$). Therefore, if no object is present in either of these sentences structural divergence cannot occur. If $x$ has declension of "be" verb at the main verb position and there is no auxiliary verb, then structural divergence cannot occur, as there is no object in the source language sentence. Further, if both the sentences have objects, and their phrase structures are same then also no structural divergence can occur. Otherwise, if the object of $x$ is a noun phrase and the object of ξ(x) is a prepositional phrase then structural divergence is identified. Figure 1 gives the corresponding algorithm.

The above algorithm is explained with respect to the example given in section 2.1. Here $x$ is "Andre will marry Steffi", and ξ($x$) is "*andre steffi se vivaah karegaa*". The SPACs of these two sentences and their correspondences are given in Figure 2. Here dotted arrows represent correspondence, and bold lines indicate no correspondence. Further, the objects of $x$ and ξ($x$) are not null; in $x$ the object is "Steffi", whereas in ξ($x$) the object is "*steffi se*". But their SPACs are [NP [N]] and [PP [NP[N][P]] respectively, which are not equal. Therefore, a structural divergence is identified.

### 3.3 Identification of categorial divergence

Categorial divergence cannot occur (in English to Hindi) if the tense (i.e. present/ past/ future), and form (i.e. indefinite, continuous etc.) of the sentences are same. Categorial divergence deals with the subjective complement. In all the examples that we have encountered so far, we observed that a categorical divergence is associated with the following: the main verb is a declension of "be", and the auxiliary verb is absent. Similarly, for ξ($x$) (i.e the Hindi sentence), if the root word of the

**Step1.** If main verb root word of *x* is "be" **AND** its auxiliary verb is null
Then return (0).
**Step2.** If object of *x* **OR** object of ξ(*x*) is null
Then return (0).
**Step3.** If the SPAC of object of *x* **AND** the SPAC of object of ξ(*x*) are same
Then return (0).
**Step4.** If SPAC of the object of *x* is NP and SPAC of the object of ξ(*x*) is PP
Then return (1)
Else return (0).

**Figure 1. Algorithm for Identification of Structural Divergence**

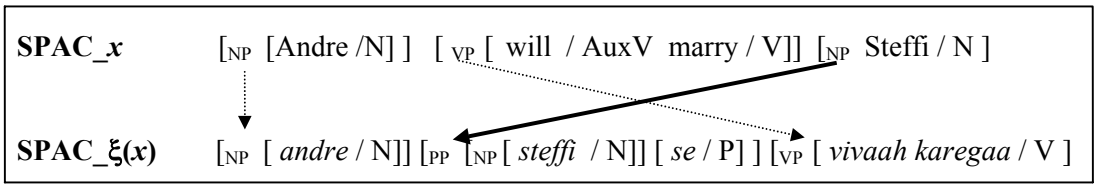


**Figure 2. Correspondence of SPAC_x and SPAC_ξ(*x*)**

main verb is not "*ho*", and there is no auxiliary verb then categorial divergence cannot occur. Otherwise depending upon whether the SC phrase structure is a NP, or AdjP or ADVP or PP, the method returns sub-type 1,2,3 or 4, respectively, indicating occurrence of the corresponding sub-type of categorial divergence. Our algorithm has been designed by taking care of the above observations. Figure 3 provides a schematic view of the proposed algorithm. We illustrate the algorithm with two sub-types of this divergence.

Let *x* be the sentence "The meeting is over", and let its translation ξ(*x*) be "*sabhaa samaapt ho gayii hai*". The corresponding SPACs for both the sentences and the term correspondences are given in Figure 4 below.

One may observe that in this example no tense correspondence is established. In *x* (the English sentence) the tense is simple present, whereas in ξ(*x*) it is present continuous. The root form of the main verb of *x* is "be", and no auxiliary verb is present in this sentence. Further, the root form of the main verb of ξ(*x*) is not "*ho*". Therefore, one can check the conditions for categorial divergence. Here, the SC of *x* is "over", while the SC of ξ(*x*) is null. SPAC of the SC of *x* is [ADVP [ADV]]. Therefore, it satisfies the condition of categorial divergence. It implies that above sentence pair has categorial divergence of sub-type 3.

In another example, we consider *x* to be "She is in tears". Its Hindi translation ξ(*x*) is "*wah ro rahii hai*". The SPACs of these sentences and their term correspondence are given in Figure 5 below. Note that in this example too tense correspondence is not established. In case of English sentence *x* the tense is simple present, whereas for the Hindi sentence ξ(*x*), it is present continuous.

The main verb root form of x is "be" and its auxiliary verb is null. In ξ(*x*) the root form of main verb is "*ronaa* (to cry)", and also the auxiliary verb is not null. Therefore, all the conditions for categorial divergence have been satisfied.

Here, SC of *x* is "in tears", and its SPAC is [PP[P][NP[N]]]. SC of ξ(*x*) is null. This implies that the above sentence pair has a categorial divergence of sub-type 4.

Due to lack of space we cannot illustrate other sub-types of categorial divergences. The same reason precludes us from presenting algorithms for identification of other divergence types in this paper.

### 3.4 Some critical comments

The technique proposed in this work provides a systematic way of checking presence of divergence for a sentence-translation pair. Efficiency of the

| | | |
|---|---|---|
| **Step 1.** | If tense and form of *x* and ξ(*x*) are same | Then return (0). |
| **Step 2.** | If main verb root word of *x* is not **"**be" | Then return (0) |
| **Step 3.** | If auxiliary verb of *x* is not null | Then return (0) |
| **Step 4.** | If main verb root word of ξ(*x*) is "*ho*" **AND** its auxiliary verb is null | Then return (0) |
| **Step 5.** | If SC of ξ(*x*) is not null | Then return (0) |
| **Step 6**. | If SPAC of SC of *x* is NP | Then return (1) |
| **Step 7.** | If SPAC of SC of *x* is AdjP | Then return (2) |
| **Step 8.** | If SPAC of SC of *x* is ADVP | Then return (3) |
| **Step 9.** | If SPAC of SC of *x* is PP | Then return (4) |

**Figure 3. Algorithim for Identification of Categorial Divergence**

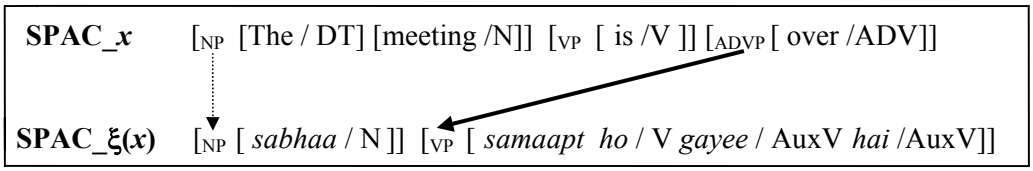


**Figure 4. Correspondence of SPAC_x and SPAC_ξ(*x*)**

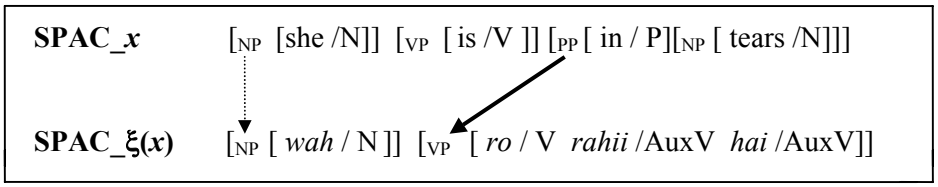


**Figure 5. Correspondence of SPAC_x and SPAC_ξ(*x*)**

algorithm however, is heavily dependent on the following three points:

- Cleaned and aligned parallel corpus of both the source and the target languages should be built.
- An on-line bi-lingual dictionary should be available. For the present work, we have used English - Hindi on-line dictionary "Shabdanjali" [http://www.iiit.net/ltrc/Dictionaries/Dict_Frame .html].
- Appropriate parsers have to be designed for SL and TL. The parsers should be able to provide the FT and SPAC information for both the languages. Note that presently, no such parser is available for Hindi. For our experiments we are using manually annotated Hindi corpora.

## 4 Conclusions

This paper deals with divergences in English to Hindi translation. Identification of divergence is very important from Example-Based Machine Translation point of view. Since an EBMT system works on imitating similar past examples, occurrence of divergence imposes a great challenge for successful machine translation.

This work presents an algorithm to identify translation divergence from English to Hindi. Although divergence often leads to different semantics for the source and target languages, we have found that comparison of the SPACs of the two sentences can identify most of the divergences. The technique is designed by considering different divergence examples that we have discovered in our example base comprising about 3000 sentences obtained from different sources, such as, children's divergence examples that we have discovered in books, translation books, advertisement material, official documents etc.

Once divergences are identified, the focus of a system designer should be on the following:

- whether a *separate* database is needed for divergence examples;
- given an input sentence what should be the *retrieval* policy so that the most appropriate examples are picked up for carrying out the adaptation.
- how to design appropriate *adaptation strategies* for modifying sentences that may involve divergence. Since translations having divergence do not follow any standard patterns, their adaptations may need specialized handling that may vary with the type/sub-type of divergence.

Consideration of these aspects is paramount for development of EBMT system for any pair of languages in general. Our experiments with different English to Hindi translation systems that are currently available on-line (Sinha et al., 2002) (Sangal et. al., 2003), (Rao, 2001) reveal that none of them is able to deal with divergences properly. Outputs of these systems are often found to be semantically incorrect. The work presented in this work should provide new insight into English to Hindi machine translation, which has gained popularity in India in recent years. Since many languages of Indian subcontinent have grammar rules and sentence structures similar to Hindi, the same approach should be useful for identification of translation divergences from English to these languages as well. Evidently, application of this technique requires target language parsers. Since MT in general is in its infancy in India, such resources are not yet available freely. MT research in Indian subcontinent should be directed to fulfil these requirements as well.

Our present work aims at developing software for identification of all the different types of divergences covering all their sub-types. We intend to build separate example bases for storing regular and divergence examples. We are also working on developing algorithms for an efficient similarity measurement scheme that will be able to decide from which of the databases past examples are to be retrieved in order to generate the translation of a given input sentence.

## 5 Bibliographical References


R.D. Brown. 1996. Example-Based Machine Translation in the Pangloss System. In *Proceeding of COLING-96:* Copenhagen, pp. 169-174.

Bonnie J. Dorr. 1993. *Machine Translation: A View from the Lexicon.* MIT Press, Cambridge, MA.

Bonnie J. Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. DUSTer: A Method for Unraveling Cross-Language Divergences for Statistical Word Level Alignment. In *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002,*Tiburon, CA.

Deepa Gupta and Niladri Chatterjee. 2003a. Divergence in English to Hindi Translation: Some Studies. *International Journal of Translation*, Bahri Publications, New Delhi. ( In print ).

Deepa Gupta and Niladri Chatterjee. 2003b. Divergence in English to Hindi Translation. Submitted to *International Conference on Natural Language Processing, ICON-2003*, Mysore, India.

Nizar Habash. 2002. Generation-Heavy Hybrid Machine Translation. In *Proceedings of the International Natural Language Generation Conference (INLG'02)*, New York.

Nizar Habash and Bonnie J. Dorr. 2002. 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.

Chung Hye Han, Benoit Lavoie, Martha Palmer, Owen Rambow, Richard Kittredge, Tanya Korelsky, Nari Kim, and Myunghee Kim. 2000. Hadling Structural Divergences and Recovering Dropped Arguments in a Korean/English Machine Translation System. In *Proceedings of the Fourth Conference of the Asoociation for Machine Translation in the Americas, AMTA-2000*, Cuernavaca, Mexico.

Yamuna Kachru. 1980. *Aspects of Hindi Grammar*. Manohar Publications, New Delhi.

Rev. S.H.Kellog and T. G. Bailey 1965. *A Grammar of the Hindi Language*, Routledge and Kegan Paul Ltd., London.

Durgesh Rao. 2001. Human Aided Machine Translation from English to Hindi: The MaTra Project at NCST. In *Proceedings Symposium on Translation Support Systems, STRANS-2001*, I.I.T. Kanpur.

Rajeev Sangal et. al. 2003. Machine Translation System "Shakti", http://gdit.iiit.net/~mt/shakti/.

R. M. K. Sinha et. al. 2002. An English to Hindi Machine Aided Translation System based on *ANGLABHARTI* Technology "ANGLA HINDI", I.I.T. Kanpur, http://anglahindi.iitk.ac.in/.

Hideo Watanabe, Sadao Kurohashi, and Eiji Aramaki. 2000. Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation. In *Proceeding of COLING-2000*, Saarbrucken, Germany.