सार्थक
A Unified Computational Lexicon for Hindi-English Code-Switching

Achla M Raina, Amitabha Mukerjee, Pankaj Goyal, Pushpraj Shukla,
Indian Institute of Technology,
Kanpur 208016,
Uttar Pradesh, India.
Email: {*pankajgo,amit,achla,praj*}*@iitk.ac.in*

Abstract

We investigate how lexicons of languages in contact are merged to generate a fused lexicon for the code-mixed variety. Using the HPSG formalism, we develop computational lexicons for Hindi and English, and explore how these can be merged to obtain a fused-lect lexicon. We consider the Hindi-English Code Switching variety (HECS), a stable variety that has resulted from contact between these languages. HECS uses words and larger phrasal constituents from one language with the syntax of the other, with the matrix language being predominantly Hindi. The grammar developed here captures this mixing of the two languages in terms of a unified lexicon that mixes pure English, pure Hindi, and cross-referenced lexical structures based on synset information for the entries. The construct of a *hinge word* is proposed to capture the cross-linguistic linkages which preserve the HPSG-based head-subcategory schema of the source lexicons. The claim is that the code-switching structures in a bilingual repertoire are triggered by cross-linguistic lexical representations that unify the matrix and embedded lexicons, and that computational mechanisms for handling this mixing can be constructed using the same principles.

1 Introduction

As computational processes are increasingly brought to bear on language varieties such as speech, email, SMS, and other informal communications, there is an increasing need to handle code-switching varieties that commonly occur in the speech interactions between multilinguals. Indeed, such varieries, especially those involving English as one of the languages, are likely to expand significantly in an increasingly bilingual world:

> [The idea that] English will become the world language to the exclusion of all others . . . is past its sell-by date. English will indeed play a crucial role in shaping the new world linguistic order, but its major impact will be in creating new generations of bilingual and multilingual speakers across the world. [Graddol, 2004]

Although [Labov, 1971] spoke of code switching as an "irregular mixture of two distinct systems," it is generally accepted today that code switching is actually quite regular. Peter Auer [Auer, 1998] has attempted to define the degree of regularity in terms of three phases in this process – code-switching, Language Mixing and Fused Lects – which progress along a continuum of what he calls "grammaticalization." In this work, we restrict our computational lexicon to relatively stable varieties in which the speakers are fluent in both languages, and where grammaticality is quite well-defined.

Consider a situation where independent lexicons are available for handling constructs in languages L1 and L2. What can then be said about using these lexicons to analyze the fused lect resulting from contact between L1 and L2? The computational issues hinge on whether it is possible to construct a fused lexicon from the two independent lexicons, and whether this merging process would require structures from outside these two lexicons? Here we explore these issues and develop a computational lexicon for *one* stable code-switching variety from the lexicons of its source languages, and in the process attempt to determine some of the algorithms that would be needed to form such fused-lect lexicons.

We consider the Hindi-English Code Switching variety (HECS) that has resulted from over two centuries of contact between these languages in the Indian context. Spoken by Hindi-English ambilinguals in northern India, HECS is regarded as a prestige dialect by the educated elite. Words and larger phrasal constituents from one language are used with the syntax of the other, with the matrix language being predominantly Hindi.

For the computational model, we implement a multilingual parser सार्थक [Saarthaka] based on the HPSG framework, which works on strings from Hindi and English. Whereas the grammatical description for English is drawn from the available work on this language [Pollard and Sag, 1994], an HPSG grammar of Hindi has been developed here [Sharma *et al.*, 2002]. In order to extend this grammar to handle code-switching structures, we depend on inter-lexical cross-referencing of synsets and a set of derivative lexical structures in a merged lexicon that is largely obtained from the two base lexicons. The fused lexicon mixes pure English, pure Hindi, and cross-referenced lexical structures while preserving the basic HPSG Head-Subcategory schema of the source lexicons. Apart from the implications that this work has for machine and human understanding systems, it makes a specific claim about the bilingual processing of mixed language varieties, where the unification of two firmly grounded lexicons is shown to be a crucial construct.

1.1 Hindi-English Code-Switching (HECS)

HECS is part of the speech repertoire of Hindi-English ambilinguals, who switch codes in different speech contexts for a variety of linguistic or socio-pragmatic reasons (e.g. to fill a lexical or even a pragmatic gap in the matrix language, or to signal social distance/proximity in interactions). In India where English is a prestige language, code-mixed varieties drawing upon English as the embedded language assume a degree of prestige: ". . . the mixed language can be said to have prestige, since the amount of mixing corresponds with the level of education and is an indicator of membership in the elite group." ([Annamalai, 2001]).

HECS is a stable variety consisting of inter- and intra-sentential mixing which demonstrates a certain regularity in observing constraints on structure. Let us consider some examples from HECS:

(1)	student	ने	teacher	के लिये	library	से	book	issue	की.
	student	PP	teacher	PP	library	PP	book	issue	operator+tense.

(The student issued a book for the teacher from the library.)

The sentence (1) is constituted almost entirely of English lexical items, yet the grammatical structure is that of Hindi. The usage is perfectly normal in most North Indian contexts. The verb "issue" is drawn from English to fill a lexical gap in Hindi. The code-switching verb "issue की(kii)", consisting of the English verb "issue" and a form of the Hindi operator "कर" (kara), observes grammatical constraints as defined for Hindi: each of the noun phrases occuring in the string is followed by postpositions "ने(ne)","के लिये(ke liye)", and "से(se)". These grammatical constraints do not apply when the English verb "issue" is used as the verbal head in a code-switching context. Thus strings such as

(2)* student ने issued a book teacher के लिये library से.

(3)* छात्र issued एक किताब अध्यापक के लिये पुस्तकालय से.

both of which use the English verb, "issue", are unacceptable in HECS. It may be noted here that the verb in (1) is a complex predicate consisting of the English verb "issue" and the Hindi operator "कर" (kara).

1.2 Grammar of Code-Switching Structures

Code-mixing has been recognized as an important issue not only in terms of sociolinguistic and historical linguistic approaches to language change, but also in psycholinguistic studies of the bilingual mind, and in other applications such as language pedagogy. Although much work has been done on grammatical analyses of different code-mixed varieties used in South Asia [Joshi, 1985; Bhatt, 1997; Annamalai, 2001], and around the world [MacSwan, 1997; Mahootian and Santorini, 1995], implemented grammatical analyses of code-mixed varieties are scarce.

Constraints on code-switching have been a subject of discussion ever since the earliest proposals regarding grammatical properties of code-mixed varieties began to appear in the 70's. Many researchers suggest that code-mixing is governed by a "third grammar" which constrains the interaction of the two language systems. The important question to investigate is whether machine implementation would require explicit construction of a completely separate "third" grammar (and its associated lexicon), or whether this can be achieved by appropriate extensions to the existing grammars of the matrix and embedded languages, with a unified lexical structure triggering the code-mixed forms.

Some constraints on code-mixing cited in the literature (following [MacSwan, 1997]) include

- Free Morpheme Constraint: There may be no switch between a bound morpheme and a lexical form unless the lexical form has been phonologically integrated into the language of the bound morpheme [Sankoff and Poplack, 1981].
- Equivalence Constraint: Code switching can occur at points in discourse when the juxtaposition of L1 and L2 elements does not violate a syntactic rule of either language [Poplack, 1980].
- Dual Structure Constraint: Structure of the embedded constituent need not conform to the constituent structure rules of the matrix language, so long as the placement in the matrix language obeys the rules of the matrix language [Sridhar and Sridhar, 1980].
- the Word Grammar Integrity Corollary proposed by Belazi et al [Belazi *et al.*, 1994] which stipulates that "A word of language X, with grammar GX, must obey grammar GX".

Joshi (1985) proposes the Closed-class constraint whereby Closed class items (det, Q-word, preposition, auxiliaries etc.) from one language cannot be mixed with open class items from the other [Joshi, 1985]. Mahootian

and Santorini (1995) propose alternate accounts that focus on the head-complement relation in the sentence. In our work, we adopt an approach similar to Mahootian and Santorini, and constrain the grammar such that the lexical properties of heads determine the grammatical structure in code-switching strings. In other words, lexical heads determine the syntactic frames of the subcategorized elements, regardless of the language from which these elements are drawn. We posit the construct of the "hinge" which is a lexical unit in L1, and can take the subcategorization frame of a synonymous lexical unit from L2. These *hinge words* can be selected as both L1 and L2 heads in the mixed-code.

Many researchers challenge the constraint-based approach to code-mixing. They also question constructs such as matrix and embedded languages, claiming these to be largely a matter of perspective [Agnihotri, 1998]. Agnihotri cites data that show violations of the constraints proposed in the literature, and goes on to suggest that "the concept of 'a language' with its attendant codified grammar may not be adequate for characterizing the constraints that condition the nature of mixed codes . . ." [Agnihotri, 1998](p. 228).

Our work Hindi-English code mixing attempts to capture some of the fuzziness inherent in code-mixed varieties within the broader objective of developing a machine implementable parser for the mixed code, HECS. We adopt a very general structural constraint invoking the HPSG-based subcategorization restrictions on lexical heads, which may be hinge words drawn from either language. The grammar works on the basis of a merged lexicon with cross-linkages between English-Hindi synsets. As such, in our system, notions like matrix and embedded language are simply a matter of terminological convenience.

2 Parser Implementation

While many parsers in use today are general enough to incorporate more than one language [Copestake and Flickinger, 2000; Erbach, 1991; Chaitanya *et al.*, 1997], the main contribution of this work is that it investigates the process of merging the lexicons of two languages to create a fused grammar for the code-mixed variety HECS. This effort is made particularly challenging by the structural differences between the two languages under consideration. सार्थक [Saarthaka] implements a Hindi-English bilingual parser using two separate lexicons for Hindi and English. In handling pure Hindi or pure English input strings, one or the other lexicon is chosen depending on the words of the input string. The question that arises while handling HECS is whether to introduce a new third lexicon, or whether to add appropriate structures to an union of the existing lexicons.

Existing parsers based on commonly employed grammatical formalisms like the ATN, TAG, CFG, and HPSG rarely deal with multilingual syntactic and semantic issues. सार्थक is an advance to the extent that it simultaneously operates on a bilingual input, and can be used as a testbed for investigating the process of creating merged lexicons for code-switching structures. In addition, unlike other implementations of HPSG which are built upon logic languages such as Prolog, our Java-based parsing engine is designed for portability.

Saarthaka's larger aim, that of generating graphics animation from multilingual stories, requires the use of a restricted domain within which object graphic models and verb action procedures can be instantiated. At present, the work is restricted to stories involving a typical family environment, with about 1000 words in English and a little less in Hindi.

3 Grammar Generation

Since the HPSG grammar remains relatively unexplored for South Asian languages, we have developed our own grammar for Hindi within the broad HPSG framework. Even for English, owing to the larger objectives that require a semantic treatment, we have constructed a grammar and lexicon of our own. The needs of the code-mixed variety HECS have led to further modification of the lexical structures.

Entries in the lexicon are *Head-driven* in the spirit of HPSG, which implies that all grammatical properties of a phrase are a function of its head. In order for the same parsing engine to handle Hindi and English inputs, the design of the lexicon needs to specify constrained word-order grammars (as in English) as well as free word-order grammars (as in Hindi). As an example, consider the following entries in the pre-code-mixed English lexicon,

`dog:dog` N(-,S,-,C) { D(S) | *J | ! } , `dogs:dog` N(-,P,-,C) { ∼D(P) | *J | ! }

The word following the colon is the root word, used in semantic processing (here it is the same as `dog`). This is followed by the lexical category, with a list of the case, number, gender, property. The second list is the set of features which can exist in the sub-category of this head word - here it means that the word `dog`, preceded necessarily by one and only one (this flexibility in `dogs` is indicated by the ∼) determiner(singular), and any number of adjectives would form a phrase with `dog` as the head. Owing to the differences between Hindi and English, some feature sorts have been added to the HPSG set of features. For example, a noun in Hindi can occur either in its root form or in the oblique form (e.g. when it takes a case marker). Thus, the noun कुत्ता [kutta, *dog*] can be used in the oblique form कुत्ते [kutte], which is captured by an additional feature in the fifth subfield in its entry:

कुत्ता:कुत्ता N(N,-,S,M,1,V) , कुत्ते:कुत्ता N(N,-,S,M,0,V)

In the same spirit, verbal agreement for gender, which is a feature of Hindi, is implemented by adding a field in verb entries. The description of Hindi nouns consists of six features as compared to four in English; the additional features in Hindi are Obliqueness and Person. For the Hindi verb entry khaata, consider:

खाता : खा T(-,-,S,M,1,V) { / N(N,-,S,M,1,V) | [N(A,-,-,-,1,V) |a(V)] |∼i|∼p|∼s|∼l|∼Y\ | !}

The delimiters '/' and '\' indicate that all features occuring inside them are freely ordered in the structure of which this word is the head. The notation '[' and ']' indicate complimentarity - i.e. only one of the features inside them can occur in the sub-category. Note that the subcategorization also differs on a few parameters which are unique to each language.The lexicon has entries for expressions that might consist of more than one word. For instance the expression '`in front of`' in English and 'के लिये' in Hindi have independent entries.

3.1 HECS Grammar Generation

Multilingual parsers permit text in one of several languages to be input, and discriminate between them at runtime. Such systems typically employ different lexicons for each language, which are identified based on the lexical items used.

Mixed-code parsers could be built using a new lexicon for the mixed code, or by merging the source language lexicons. The merged lexicon option has an obvious advantage over a new lexicon in terms of economy and elegance. Among others, Macswan [MacSwan, 1997] has claimed the merged lexicon to be a more viable option for mixed-code grammars.

The merged lexicon approach, as adopted here, is based on cross-linked synset mappings which permit a lexical unit in L1 to access its synonyms in L2. However, implementing grammars based on merged lexi-

cons has the obvious problem that the syntactic constraints of one language may get carried over to lexical units from the other, thus generating incorrect parses in specific instances. Consider the following example:

(4) raama who lives in the house books बेचने के लिये कानपुर गया

(raama who lives in the house books {becane ke liye kaanpura gayaa}) - {to sell books went to Kanpur}

Here the parser based on a merged lexicon, permits the word “books” to accept “in the house” as a subcategory based on its english lexical structure. This results in two parses, including the incorrect Parse 2 shown below:

```
+-  raama who lives in the house books
    becane ke liye kaanapura gayaa
    +-raama who lives in the house
    | +-raama N(N,3,S,M,-,V,H)
    | +-who lives in the house
    |   +-who W(-)
    |   +-lives in the house
    |     +-lives H(P,S,V,N)
    |     +-in the house
    |       +-in P(N)
    |       +-the house
    |         +-the D(-,-)
    |         +-house N(-,S,-,-,E)
    +-books becane ke liye
    | +-books becane
    | | +-books N(A,3,P,F,1,V,H)
    | | +-becane Z(P,-)
    | +-ke liye p()
    +-kaanapura N(A,3,S,M,-,V,H)
    +-gayaa A(-,-,S,M,-,V)
```

Parse 1: Correct Parse

```
+-  raama who lives in the house
    books becane ke liye kaanapura gayaa
    +-raama who lives
    | +-raama N(N,S,M,-)
    | +-who lives
    |   +-who W(-)
    |   +-lives H(P,S,V,N)
    +-in the house books becane ke liye
    | +-in the house books becane
    | | +-in the house books
    | | | +-in the house
    | | | | +-in P(N)
    | | | | +-the house
    | | | |   +-the D(-,-)
    | | | |   +-house N(-,S,-,-)
    | | | +-books N(-,P,-,-)
    | | +-becane Z(P,-)
    | +-ke liye p()
    +-kaanapura N(A,3,S,M,-,V)
    +-gayaa A(-,-,S,M,-,V)
```

Parse 2: Incorrect Parse

English being a head-first language, the noun phrase in the subject position - in this case “raama who lives in the house”- should be followed by a verb phrase, introduced by a verbal rather than a nominal element. But in the code-switching string above, it is the Hindi purposive clause “books becane ke liye”- beginning with a nominal element “books” - that follows the first noun phrase. Now since “books” has features specified by the English grammar, the parser sees it as a nominal head subcategorizing the prepositional phrase “in the house”, as in a sring like “in my house books are kept in the study”, thus generating the incorrect parse above.

One obvious though inelegant solution to this problem would be to abandon the merged lexicon option altogether, and create a third grammar with an associated lexicon for the code-mixed lect. A possible solution to such overgeneration in the merged lexicon approach is to introduce an additional feature field indicating language - ’H’ or ’E’. The merged lexicon is based on cross-linking of English-Hindi synsets. For example, in (4) above, ’books’ is cross-linked with किताबें (kitaabe), and takes an ’H’ feature, and as such, it does not unify with the preceding phrase “in the house”. The original entry of “books”, marked ’E’ is not operative since its head ’becane’ is marked ’H’. Thus only one coorect parse ($Parse1$, shown earlier) is generated. Based on these considerations, we have, in this work, used a merged lexicon with cross-linking between English-Hindi

synsets, as well as source language tags in the cross-linked entries.

3.2 HECS Parsing

The structure in mixed-code is governed by the lexical head, i.e. the verbal head for the sentence, and other heads for other phrasal constituents. Thus, if the head is from Hindi, the structure observes constraints imposed by the Hindi grammar, and similarly for English. The lexical units which in the merged lexicon are represented as carrying both H and E tags are critical to the code-mixing process. These units, termed here as *hinge words*, refer to a lexical unit in L1, which takes the subcategorization frame of a synonymous lexical unit from L2. Given the broad head-subcategory constraint on code-mixing, the hinge words are involved in HECS generation and parsing in several ways. We discuss some of these next.

3.2.1 Substitution

Noun substitutions are handled as equivalence relations between Hindi and English words that function as synsets in HECS. Thus an entry for the *hinge word* "books" is equated to the list of fields for किताबें, किताबों [kitaabe, kitaabo] which "books" can now replace in code-switching sentences such as:

[उसने books/किताबों को पढ़ा] or

[raama books/किताबें पढ़ता है]

The two additional entries for books are:

किताबें,books : kitaab,book N(A,3,P,F,1,V,H)

किताबों,books : kitaab,book N(N,3,P,F,0,V,H)

We note also that the entry for the *hinge word* "books" with feature 'H' inherits the PNG and oblique features of the hindi noun "किताबों" which then constrain its usage. Similarly, the Hindi word किताबें is equated with the entry of the English word 'books' as [books,किताबें : book,किताब N(-,P,-,-,E).] This enables us to parse sentences such as [I have issued these किताबें]. Since gender in Hindi is grammatical, inanimate common nouns are arbitrarily assigned masculine or feminine gender features. Thus the Hindi synonym for "novel", "upanyaas", is masculine, whereas book ("kitaab") is feminine, and this leads to structures such as:

राम का novel market में बिका.

राम की book market में बिकी

Finally, consider again the sentence (4):

(4) raama who lives in the house books बेचने के लिये कानपुर गया

Here, "raama" as the head of a subcategory of गया (gayaa) is marked H, yet in the phrasal constituent "raama who lives in the house" it appears as E - and thus serves as the *hinge*.

Lexical categories other than the noun are also subject to cross-linkages similar to those under discussion. Code-mixing structures which have the verb as the head undergo what we term below as verb incorporation. This is indeed a very productive code-mixing pattern in our cross-referenced lexicon, and merits detailed discussion.

3.2.2 Verb incorporation

An English verb can be incorporated into a Hindi grammatical structure. This is done by a class of complex predicates consisting of an English root verb in its nominalised form, together with the Hindi operator कर

(kara) or हो (ho) and their morphological variants such as किया (kiyaa) , की (kii) , करता (karataa) , हुआ (huaa) , हुई (huii) , होता (hotaa) , etc. For example - the sentence {उसने किताब पढ़ी [usane kitaaba paDi]} could be code mixed to {उसने किताब read की [usane kitaaba read ki]}. Here पढ़ी has been replaced by "read की", which takes on the subcategorization features of the synonym पढ़ी [paRhi]. Again,{किताब बिकी [kitab bikii]} may be realised as {किताब sell हुई[kitaab sell hui]}.

So the original entry for की (kii) - की:की Aux(kii,F){/n(V)|[N(A,-,S,F,1,V)|a(V)]|∼i|∼p|∼s|∼l|∼Y\|!}
now changes into - की:की Aux(kii,F){/n(V)|[N(A,-,S,M,1,V,H)|a(V)]|∼i|∼p|∼s|∼l|∼Y|\| **N(V)** |!}
where English root verb entries are tagged as "N(V)" to refer to this specific N+V construct where the verb now occurs in a nominalised form. Now, the operator 'kara' can take them as subcategories: - `read:read N(V){!}` We note that many N+V combinations of this kind may not have a synonym in L2 (e.g., "book kiyaa" as in "ticket book kiyaa", or "issue kii"). In such cases the subcategorisation restrictions are to be stated.

A class of English lexical units comprising action verbs is thus incorporated into the mixed code as a subcategory of the Hindi action operator 'kara' as the *hinge word*. Likewise, the stative verbs get incorporated into the mixed code as a subcategory of the stative operator 'ho' yielding complex predicates of N+kara and N+ho type. The complex predicate formed as a result of verb incorporation is cross-linked with its Hindi synonym, with which it shares its lexical structure. In other words, the mixed code verb observes the subcategory restrictions stated for its Hindi synonym. To what extent is the class of verbs undergoing incorporation constrained by the semantic type of the verb is a matter which deserves further investigation.

We consider now some of the HECS data discussed in the literature on Hindi-English code-mixing [Pandit, 1990]. a. vo hameshaa daftar me on samay aataa hai 'He always comes to the office on time.' b. he always comes to the office time par c. * vo hameshaa daftar me samay on aataa hai. d. * he always comes to the office par time.

Assuming our Head-driven constraint on code-switching, we need to take note of the fact that Hindi postpositional heads are preceded, in terms of head directionality, by their NP subcategories, whereas English prepositional heads are followed by their NP subcategories. (a) and (b) are unproblematic: the preposition 'on' in (a) is a head with subcat properties of the English prepositional heads, the subcat 'samay', therefore follows the head, even though it is drawn from the Hindi lexicon. In (b), 'par' is a Hindi postpositional head whose subcat 'time', drawn from the English lexicon, precedes it. This is what our constraint predicts - the language of the head determines the subcat properties. (c) is bad because the English preposition 'on' is preceded, rather than followed, by its subcat 'samay'. Likewise, (d) is bad because, the Hindi postposition 'par' is followed, rather than preceded by its subcat 'time'. This is exactly what we expect.

3.3 Lexicon Merger

The lexicon for HECS is thus a unified lexicon, built using the following two constructs:

1. Cross-linkages which add a feature to *hinge words* - nouns and other lexical categories from L1 are duplicated with a language tag L2, and taking the subcategorization features of an appropriate lexical unit from the L2 synset.
2. Additional entries for the operators kara and ho from hindi, which yield N+V entries (with subcategorisation features of corresponding Hindi synsets.

The main tool needed for automating the process of generating the code-mixed lexicon is a crosslinguistic synset mapping that accounts for contrastive polysemies. To generate the cross linkages through an algorithm, a semantically rich lexicon in which kitaab chooses only the appropriate synset and no inappropriate ones - given nearly ten or more entries for the noun book - would be needed. The synset model in WordNet [Fellbaum, 1998] provides a possible framework for this exercise, provided that the Hindi Wordnet is available with cross-indexed synset labels. Given the availability of such tools, it would become possible to automate the procedure for cross-linking the appropriate entries from the synset in L1 into the entry for a word from its corresponding synset in L2.

As of now, since the Hindi Wordnet is still under development [Bhattacharya, 2002], the above devices are handled manually. All cross linkages between synonymous nouns are introduced manually. As such kitaab is manually cross-linked with the appropriate entry for book.

Similarly, with the availability of synset cross-linking, it would be possible to automate the process for creating new entries of the the N+V kind. Such an algorithm will again work with only a semantically rich lexicon, given the rich polysemy of verbs in both the languages. However, substitutions involving synset mismatch, for which no corresponding subcategorization frame is available in the other language, would clearly defy automated handling. The same applies to verb incorporation.

4 Summary and Conclusion

The main contribution of this work is the formalisation of the process for merging two lexicons. This involves the availability of a crosslinguistic synset mapping, based on which substitutable hinge words are cross-referenced using source language tags. The general constraint governing code-mixing is that phrasal heads determine the syntactic properties of the subcategorized elements, regardless of the source language these elements are drawn from. The parser Saarthaka has been implemented to test the merged Hindi-English lexicon on its ability to generate parses for strings from Hindi, English, and from the code-mixed variety HECS.

In the coming decades, with emphasis shifting away from formal textual sources to speech and other informal text, determining computationally efficient processes for constructing mixed-code lexicons will remain a challenging problem. Whether the processes outlined in this work can generalize to other code-switching varieties is an important issue that remains to be addressed. Some of the mechanisms outlined here, such as hinge word substitutions, appear to be applicable in a more general context, whereas the particular mechanism of verb incorporation adapting English verbs into Hindi may be specific to a class of code-mixed varieties.

References

[[Agnihotri, 1998]] Rama Kant Agnihotri. *Social Psychological Perspectives on Second Language Learning*. Sage Publications, New Delhi, 1998.

[[Annamalai, 2001]] E. Annamalai. *Managing multilingualism in India - Political and Linguistic manifestations*. Sage Publications, New Delhi, 2001.

[[Auer, 1998]] Peter Auer. From code-switching via language mixing to fused lects: Toward a dynamic typology of bilingual speech. Technical Report InLiSt No. 6, Interaction and Linguistic Structures, Freiburg i. Br., September 1998.

[Belazi *et al.*, 1994] Hedi M. Belazi, Edward J. Rubin, and Almeida J. Toribio. Code switching and x-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2):221–237, 1994.

[Bhatt, 1997] Rakesh Mohan Bhatt. Code-switching, constraints, and optimal grammars. *Lingua*, 102(4):223–251, August 1997.

[Bhattacharya, 2002] Pushpak Bhattacharya. Hindi wordnet, 2002.

[Chaitanya *et al.*, 1997] Vineet Chaitanya, Amba P. Kulkarni, and Rajeev Sangal. Anusaaraka: Machine translation in stages. *Vivek*, 10, 1997.

[Copestake and Flickinger, 2000] Ann Copestake and Dan Flickinger. An open source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

[Erbach, 1991] Gregor Erbach. A flexible parser for a linguistic development environment. In O. Herzog and C.-R. Rollinger, editors, *Natural Language Understanding in LILOG*. Springer, Berlin, Germany, 1991.

[Fellbaum, 1998] C. Fellbaum. Wordnet:an electronic lexical database, 1998.

[Graddol, 2004] David Graddol. The future of language. *Science*, 303(5662):1329–1331, 27 Feb 2004.

[Joshi, 1985] Aravind Joshi. Processing of sentences with intrasential code switching. In *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*. Cambridge University Press, Cambridge, 1985.

[Labov, 1971] William Labov. The notion of 'system' in creole languages. In Dell Hymes, editor, *Pidginization and Creolization of Languages*. Cambridge University Press, Cambridge, 1971.

[MacSwan, 1997] Jeffrey MacSwan. *A Minimalist Approach to Intrasentential Code Switching: Spanish-Nahuatl Bilingualism in Central Mexico*. PhD thesis, University of California Los Angeles, 1997.

[Mahootian and Santorini, 1995] Shahrzad Mahootian and Beatrice Santorini. Codeswitching and the syntactic status of adnominal adjectives. *Lingua*, 95:1–27, 1995.

[Pandit, 1990] Ira Pandit. Grammaticality in code switching. In Rodolfo Jacobson, editor, *Codeswitching as a Worldwide Phenomenon*. Peter Lang, New York, 1990.

[Pollard and Sag, 1994] Carl Pollard and Ivan E. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, 1994.

[Poplack, 1980] Shana Poplack. Sometimes i start a sentence in english y termino en espanol: Towards a typology of code-switching. *Linguistics*, 18:581–618, 1980.

[Sankoff and Poplack, 1981] David Sankoff and Shana Poplack. A formal grammar for code-switching. *Papers in Linguistics*, 14(1):3–46, 1981.

[Sharma *et al.*, 2002] Deepak Sharma, K. Vikram, Manav R. Mital, Amitabha Mukerjee, and Achla M Raina. Saarthaka: A generalized hpsg parser for english and hindi. In *Recent Advances in Natural Language Processing - Proceedings ICON-2002*, Mumbai India, December 2002. Vikas Publishing House.

[Sridhar and Sridhar, 1980] K.N. Sridhar and S.N. Sridhar. Psycholinguistics of bilingual code-mixing. *Canadian Journal of Psychology*, 34(4):409–418, 1980.