COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

PROCESSING OF SENTENCES WITH INTRA-SENTENTIAL CODE-SWITCHING [1]

Aravind K. Joshi

Department of Computer and Information Science
R. 268 Moore School
University of Pennsylvania
Philadelphia, PA 19104

Speakers of certain bilingual communities systematically produce utterances in which they switch from one language to another, suggesting that the two language systems systematically interact with each other in the production (and recognition) of these sentences. We have investigated this phenomenon in a formal or computational framework which consists of two grammatical systems and a mechanism for switching between the two systems. A variety of constraints apparent in these sentences are then explained in terms of constraints on the switching mechanism, especially, those on closed class items.

1. INTRODUCTION

Speakers of certain bilingual communities systematically produce utterances in which they switch from one language to another (called code-switching), possibly several times, in the course of an utterance. Production and comprehension of utterances with intrasentential code-switching is part of the linguistic competence of the speakers and hearers of these communities. Much of the work on code-switching is in the sociolinguistic framework and also at the discourse level. Recently there have been few studies of code-switching within the scope of a single sentence. (See Sridhar (1980) for a good review, also Pfaff (1979)). Also until recently, this phenomenon has not been studied in a formal or computational framework. (See Sankoff and Poplack (1980), Woolford (1980), Joshi (1980), and Doron (1981). Space does not permit a detailed comparison.)

The discourse level of code-switching is important, however, it is only at the intrasentential level that we are able to observe with some certainty, the interaction between two grammatical systems. These interactions, to the extent they can be systematically characterized, provide a nice framework for investigating some processing issues both from the generation and parsing points of view.

There are some important characteristics of intrasentential code-switching which give hope for the kind of work described here. These are as follows. 1. The situation which we are concerned with involves participants who are about equally fluent in both languages. 2. Participants have fairly consistent judgements about the "acceptability" of mixed sentences. (In fact it is amazing that participants have such acceptability judgements at all.) 3. Mixed utterances are spoken without hesitation, pauses, repetitions, corrections, etc., suggesting that intrasentential code-switching is not some random interference of one system with the other. Rather, the switches seem to be due to systematic interactions between the two systems. 4. The two language systems seem to be simultaneously active. 5. Intrasentential code-switching is sharply distinguished from other interferences such as borrowing, learned use of foreign words, filling lexical gaps, etc. all of which could be exhibited by monolingual speakers. 6. Despite extensive intrasentential switching, speakers and hearers usually agree on which language the mixed sentence is "coming from". We call this language the matrix

language and the other language the embedded language. These interesting characteristics of the mixed sentences suggest that the two language systems are systematically interacting with each other in the production (and recognition) of the mixed sentences.

Our main objectives in this paper are (1) to formulate a system in terms of the grammars of the two languages and a switching rule, (2) to show that a variety of observable constraints on intrasentential code-switching can be formulated in terms of constraints on the switching rule. The main result of this paper is that a large number of constraints can be derived from a general constraint on the switchability of the so-called closed class items (determinizers, quantifiers, prepositions, tense morphemes, auxilliaries, helping verbs, complementizers, pronouns, etc.). This result is of interest because the differential behavior of closed class items (as compared to the open class items) has been noted in various aspects of language processing (in the monolingual case), for example, (1) certain types of speech errors which strand the closed class items, (2) resistance to change as well as resistance to incorporate new items as closed class items, (3) frequency indpendent lexical decision for closed class items (as compared to open class items for which lexical decision is frequency dependent), (4) the absence of frequency independence for closed class items in certain types of aphasia, (5) closed class items aiding in comprehension strategies, etc. (This list is based on a talk given by Mary-Louise Kean at the University of Pennsylvania). It is not clear what the relationship is between the behavior of closed classes in intrasentential code-switching and the other behaviors (in monolingual situations) described above. We believe, however, that investigating this relationship may give some clues concerning the organization of the grammar and the lexicon, and the nature of the interface between the two language systems.

The examples in our paper are all from the language pair, Marathi (m) and English (e). Marathi (m) is the matrix language and English (e) is the embedded language. (The coincidence of the abbreviation m for the matrix language, which is Marathi and e for the embedded language, which is English, is accidental, but a happy one!). A few facts about Marathi will be useful to note. It is an Indo-European language (spoken on the west coast of India near Bombay and in parts of central India by about 60 million people). It is an SOV language. Adjectives and relative clauses appear prenominally and it has postpositions instead of prepositions. It uses a rich supply of auxilliary or helping verbs. Other facts about Marathi will become apparent in the examples. (See Section 3).

2. FORMULATION OF THE SYSTEM

Let L_m be the matrix language and L_e be the embedded language. Further let G_m and G_e be the corresponding grammars, i.e., G_m is the matrix grammar and G_e is the embedded grammar. A "mixed" sentence is a sentence which contains lexical items from both L_m and L_e. Let L_x be the set of all mixed sentences that are judged to be acceptable. Note that a mixed sentence is not a sentence of either L_m or L_e. However, it is judged to be "coming from" L_m. The task is to formulate a system characterizing L_x. Our approach is to formulate a system for L_x in terms of G_m and G_e and a 'control structure' which permits shifting control from G_m to G_e but not from G_e to G_m. We assume a 'correspondence' between categories of G_m and G_e, for example, NP_m corresponds to NP_e (written as $NP_m \approx NP_e$). Control is shifted by a switching rule of the form

(2.1) $A_m \times A_e$, where A_m is a category of G_m, A_e is a category of G_e, and $A_m \approx A_e$.

At any stage of the derivation, (2.1) can be invoked, permitting A_m to be switched to A_e. Thus further derivation involving A_m will be carried out by using rules of G_e, starting with A_e. The switching rule in (2.1) is asymmetric i.e., switching a category of the matrix grammar to a category of the embedded grammar

is permitted but not vice versa. This asymmetry can be stated directly in the rule itself, as we have done, or it can be stated as a constraint on a more generalized switching rule which will permit switching from A_m to A_e as well as the other way round. We have chosen to state the asymmetry by incorporating it in the rule itself because the asymmetry plays such a central role in our formulation. This asymmetric switching rule together with the further constraints described in Section 3 is intended to capture the overpowering judgement of speakers about a mixed sentence "coming from" the matrix language L_m. The switching rule in (2.1) is neither a rule of G_m nor a rule of G_e. It is also not a rule of a grammar, say G_x for L_x. As we have said before, we will construct a system for L_x in terms of G_m and G_e and a switching rule and not in terms of a third grammar, say G_x. Although formally this can be done, there are important reasons for not doing so. Using this general framework we will now show that the system for L_x can be formulated by specifying a set of constraints on the switching rule (besides the asymmetry constraint). These further constraints primarily pertain to the closed class items.

3. CONSTRAINTS ON THE SWITCHING RULE

Our hypothesis is that L_x can be completely characterized in terms of constraints on the switching rule (2.1). The types of constraints can be characterized as follows;

3.1 Asymmetry: We have already discussed this constraint. In fact, we have incorporated it in the definition of the switching rule itself. The main justifications for asymmetry are as follows. (a) We want to maintain the notion of matrix and embedded languages and the asymmetry associated with this distinction. (b) Arbitrarily long derivations would be possible, for example, by allowing back and forth switching of A_m and A_e along a nonbranching path. There appears to be no motivation for allowing such derivations. (c) The asymmtry constraint together with certain constraints on the non-switchability of closed class items seem to allow a fairly complete characterization of L_x.

3.2 Constraint on switchability of certain categories: Rule (2.1) permits switching any category A_m to A_e if $A_m \approx A_e$. However certain categories cannot be switched. Although all major categories can be switched, we must exclude the root node S_m. Obviously, if we permit S_m to be switched to S_e, we can derive a sentence in L_e starting with S_m in a trivial manner. Hence, we need the following constraint.

(3.2.1) Root node S_m cannot be switched.

Constraints on closed class items: (3.2.2) Certain closed class items such as tense, aux, and helping verbs when they appear in main VP cannot be switched.

Examples: (underlined items in the examples are from L_m).

(3.1) mula khurcyā rangawtāt.
boys chairs paint

(3.2) mula khurcyā paint kartāt.
do(+tense)

In (3.2) the root verb has been switched from Marathi to English. The closed class item tat is not switched, however it is attached to an auxilliary or helping verb kar, since it cannot be stranded. This phenomenon appears in mixed sentences of other language pairs (see Pfaff(1979).)

It is not possible to switch both the V and the tense in (3.1), and also not the entire VP.

(3.3) *mula khurcyā paint. (3.4) *mula paint chairs.

Note that (3.4) could be derived by starting with S_e (i.e., by starting to derive a sentence of L_e) and then switching VP_e to VP_m, but this is not permitted by the asymmetry constraint. Of course, one cannot start with the S_e node because this requires switching S_m toS_e which is blocked by the constraint on the switchability of the root node.

(3.2.3) Closed class items (e.g., determiners, quantifiers, prepositions, possessive, aux, tense, helping verbs, etc.) cannot be switched. Thus, for example, DET_m cannot be switched to DET_e. This does not mean that a lexical item belonging to DET_e cannot appear in the mixed sentence. It can indeed appear if NP_m has already been switched to NP_e and then NP_e is expanded into DET_eN_e according to G_e.

Examples

(3.5) kāhi khurcyā DET_mN_m (3.6) some chairs DET_eN_e
some chairs

(3.7) kāhi chairs DET_mN_e (3.8) * some khurcyā * DET_eN_m

Adjectives are not closed classes; hence all four combinations below are possible.

(3.9) unca peṭi (3.10) unca box (3.11) tall peṭi (3.12) tall box
tall box

Note that (3.12) is a Marathi NP_m in which both the A_m and N_m have been switched. It is not derived from NP_e, if it were, it would have a determiner. (Determiner is optional in Marathi).

Prepositions and postpositions are closed class items. Marathi has postpositions while English has prepositions.

(3.13) kāhi khurcyāwar (3.14) kāhi chairswar (3.15)? some chairswar+
some chairs on

(3.16)* some chairs on (3.17)* kāhi khurcyā on (3.18) on some chairs

(3.19) *on kāhi khurcyā (3.20)* war kāhi khurcyā)3.21)* war some chairs

(3.2.3) Constraints on Complementizers: Complementizers are closed class items and therefore cannot be switched in the same sense as in (3.2.2) above. However, often we have a choice of a complementizer. This choice depends both on the matrix verb V_m and the embedded verb V_e ($V_m \approx V_e$) to which V_m has been switched. Let the complementizers of V_m be $COMP_m = \{C_1, C_2, C_3\}$ and the complementizers of V_e ($\approx V_m$) be $COMP_e = \{C_1', C_2', C_4'\}$ where $C_1 \approx C_1'$, $C_2 \approx C_2'$. Now if V_m is switched to V_e i.e., the verb is lexically realized in the embedded language, then the choice of the complementizer is constrained in the following manner. Since complementizers are closed classes, they cannot be switched. Hence, the choice is C_1, C_2, or C_3; however only C_1 and C_2 are permitted, as the equivalent lexical verb V_e permits C_1' and C_2' which are the equivalents of C_1 and C_2 respectively. C_3 is not permitted because its equivalent C_3' is not permitted for V_e, and C_4 which is the equivalent of C_4' is not permitted because it is not allowed by V_m. Thus the only complementizers that are permitted, if V_m is switched to V_e, are those that are permitted by V_m and the equivalents of which are permitted by V_e ($V_m \approx V_e$). Thus the choice is constrained not only to the complementizers of V_m (because of non switchability of complementizers) but it is further constrained by the choice of complementizers of V_e as explained above.

+This is a problematic case which is discussed in detail in the longer version of this paper.

Examples:

(3.22) tō parat jāyca tharawtō. ca:ing
he back going decides

(3.23) * tō parat jāyla tharawtō. lā:to
he back to go decides

The Marathi verb tharaw (decide) takes the complementizer ca(ing) but not the complementizer lā(to). The corresponding English verb decide takes both the complementizers to and ing (after on). We now switch the Marathi verb V_m (tharaw) to V_e (decide) in both (3.22) and (3.23). Since the tense in the main VP cannot be switched (as we have seen in (3.1) and (3.2) earlier) a helping verb kar (do) has to be introduced so that the tense can be attached to it. Thus we have

(3.24) tō parat jāyca decide kartō. ca:ing
he back going do(+tense)

(3.25) * tō parat jāyla decide kartō. la:to
he back to go do(+tense)

Note that although decide takes both the complementizers to and ing, only (3.24) is allowed. (3.25) is blocked because the Marathi verb tharaw does not allow the complementizer to. Thus the only complementizer that appears in the mixed sentence is ing.

There are several interesting issues concerning the generation and recognition of sentences such as (3.24) and (3.25). For example, at what point the decision to switch the main verb is made? (We could have raised this issue earlier when we discussed (3.1) and (3.2)). Since a new helping verb has to be introduced when the switch is made, does it mean that some 'local' structural change has to be made along with the switching of the verb? Another point is that the choice of the complementizer (which comes before the matrix verb) also determines whether the verb can be switched or not. The machinery we have provided so far may have to be augmented to provide systematic answers to these questions. Thus for example, we may have to introduce additional constraints on the switching rules.

4. PARSING CONSIDERATIONS

In this paper, we have given an account of the constraints on intrasentential code-switching in a generative framework. The formal model and the constraints on switching that we have proposed clearly have implications for the kind of parser we may be able to construct. We will not pursue this aspect in this paper. However, we would like to point out that by adopting some parsing strategies, we can account for some of the constraints described earlier. A preliminary attempt was made in Joshi (1981) by proposing a strategy involving a so-called left corner constraint. This strategy has some serious drawbacks as was pointed out by Doron (1981). She has proposed an alternate strategy called 'early determination strategy', according to which the parser tries to determine as early as possible the language of the major constituent it is currently parsing. Thus upon encountering a Marathi (m) determiner i.e., DET_m the parser would predict a Marathi NP_m. The Marathi N_m could be then realized lexically in Marathi or the N_m would be switched to N_e and then lexically realized in English. NP_m is expanded into $DET_m Nom_m$ where Nom_m is expanded into $A_m N_m$.
Note that A_m and N_m could be independently switched to A_e and N_e respectively, thus giving four possible sequences:

$DET_m\ A_m\ N_m$, $DET_m\ A_m\ N_e$, $DET_m\ A_e\ N_m$, $DET_m\ A_e\ N_e$, all of which are permissible.

If the parser encountered an English determiner, i.e. DET_e then it would predict NP_e, but now N_e or $A_e N_e$ into which NP_e can expand cannot be switched to N_m or A_m

because of the asymmetry constraint. Thus the only permissible sequence is DET_e (A_e) N_e, and the following are excluded, i.e., *DET_e N_m, * DET_e A_e N_m, * DET_e A_m N_e, * DET_e A_m N_m, which checks with the data.

Of course, so far we have the same predictions as we had with the constraint on the nonswitchability of closed class items. However, there is some evidence to the effect that a parsing strategy as described above may be in effect. The following distribution is correctly predicted by the above strategy:

(5.1) * tall petya (5.2) tall boxes (5.3) unca petya (5.4) unca boxes.

(5.1) is disallowed, because upon encountering an English adjective, A_e, the parser predicts Nom_e, which is expanded into A_e N_e. However, N_e cannot be realized lexically in Marathi, unless N_e is switched to N_m, which is disallowed. Note that (5.1) cannot be disallowed by invoking nonswitchability of adjectives, because these are not closed classes. This early determination strategy does not help however in accounting for the distribution of phrases involving postpositions (see Section 3).

Our conclusion at present is that the framework described in Section 3 along with the constraints on closed class items is the proper way to formulate the code-switching system. A parsing strategy as discussed above is perhaps also operative (see Examples (5.1) - (5.4)) and when a closed class item is the leftmost constituent of a major category then the two formulations made the same predictions.

5. CONCLUSION

We have presented a formal model for characterizing intra-sentential code-switching. The main features of this model are that 1) the model treats the two grammars (languages) asymmetrically, 2) there is no third grammar, and 3) the constraints on the switchability of closed class items. We believe that further investigation of code-switching in the proposed framework will be very productive, as it captures some essential aspects of intrasentential code-switching. Another interesting result concerns the role of closed class items. Since several important characteristics of closed class items are well-known in the context of processing of monolingual utterances, we think that further investigation of the role of closed class items in the context of code-switching will give us some insights into the processing of monolingual utterances. Our investigation of intra-sentential code-mixing can also be considered as a small contribution towards the larger problem of determining the nature of the interface between the two language systems of a bilingual speaker or hearer.

REFERENCES

Doron, E., "On formal models of code-switching", MS., U. of Texas at Austin (1981).

Joshi, A.K., "Some problems in processing sentences with intrasentential code switching", Extended Abstract of paper read at the U. of Texas Parsing Workshop, March 1981.

Pfaff, C., "Constraints on language switching", Language, Vol. 55, 1979.

Sankoff, D. and Poplack, S., "A formal grammar of code-switching:, Working Paper, 1980.

Sridhar, S.N., "The syntax and psycholinguistics of bilingual code-mixing", Studies in the Linguistic Science, Spring 1980, University of Illinois.

Woolford, E., "A formal model of bilingual code-switching", Working Paper, M.I.T., 1980.