ON A FORMAL MODEL OF CODE-SWITCHING

Edit Doron

1. Introduction

This paper is concerned with a linguistic skill that manifests itself in multilingual communities.*

Producing and understanding utterances with intra-sentential alternating of two languages -- code-switching -- appears to be part of the linguistic competence of members of communities where these languages are in contact.[1] Anyone interested in proposing a model for this competence is confronted first of all with the question of how to interrelate the grammars dealing with monolingual sentences so as to get a system that can deal with mixed sentences.

In this paper I first present and evaluate three existing treatments of code-switching: Sankoff and Poplack (1979), Woolford (1980) and Joshi (1981). I then suggest that constraints on code-switching can be explained by a general parsing strategy together with considerations about the syntactic characteristics of the particular languages.

---

*I wish to thank my consultant Terry Ovalle for patiently going through my examples and providing me with clear judgements. I am indebted to John Baugh and Aravind Joshi for reading through an earlier version of this paper and offering me valuable comments, some of which I have incorporated in the text. I am especially grateful to Aravind Joshi for kindly answering my questions on Marathi.

[1]See the references to Poplack's work, where she convincingly argues that "...code-switching is a verbal skill requiring a large degree of linguistic competence in more than one language, rather than a defect arising from insufficient knowledge of one or the other." (Poplack (1979) p.72).

2. Stating the problem

Let G^1 and G^2 be two grammars generating the languages L^1 and L^2 respectively. I assume the lexicons of G^1 and G^2 are disjoint - but for borrowings.[2] On the other hand, it is reasonable to assume that G^1 and G^2 have some category symbols and phrase structure rules in common. For example S, NP, VP, N, A, Det, might be common categories, and S ---> NP VP, NP ---> Det N, might be common rules.

Let L be the set of mixed sentences that are acceptable to members of the community where L^1 and L^2 are in contact.[3] There are two basic ways of viewing L:

A. L is a language generated by a single grammar G that is symmetrically built from G^1 and G^2.

B. Each string in L is "assigned" to either L^1 or L^2, even though it has had some substring replaced by a substring from the other language.

The problem of choosing between the two treatments can be formulated in the jargon of grammars or of parsers. In the jargon of grammars, A and B can be reformulated as A' and B' respectively:

[2] On the distinction between borrowing and code-switching see for example Pfaff (1979).

[3] As we shall see later, L is not well defined, since the acceptability of a mixed sentence depends on whether the "host language" is L^1 or L^2 (cf. the discussion below of Joshi's work). But I leave it at that for now, since Sankoff and Poplack (1979) and Woolford (1980) implicitly assume that L is well defined. Notice also that I do not specify whether L contains L^1 and L^2. Again the different authors differ on this point: as we shall see below Sankoff and Poplack's system sets L to contain L^1 and L^2, whereas in Woolford's system L is set to be disjoint from L^1 and L^2.

A'. L is generated by a grammar G such that the set of categories of G is the union of the sets of categories of G^1 and G^2, and likewise for the rules. What this implies is for example that a category like NP, that is common to G^1 and G^2, and therefore is also a category of G, will not be marked in G for being an NP of G^1 or an NP of G^2.

B'. L is generated by a grammar G that is defined in the following way: The sets of categories and rules of G^1 and G^2 are first made disjoint - e.g. by marking NP from G^1 as NP^1 and NP from G^2 as NP^2. Unions are then taken over these disjoint sets to create the sets of categories and rules of G. Since S^1 and S^2 are different category symbols, we can "assign" even the mixed sentences they dominate to L^1 and L^2 respectively - in this sense we have kept the two grammars separate. Notice that G also needs to have rules "relating" the corresponding categories of the two grammars.

In the jargon of parsers the problem is whether the parser should or should not mark differently category symbols and rules it draws from the two grammars.

The problem is basically the same in the two jargons. The jargon of parsers favours the approach that keeps the two grammars disjoint, since the marking of the categories from the two grammars could be made differently in different discourses.

3. On Three Existing Approaches to Code-Switching

3.1. Sankoff and Poplack's approach

Sankoff and Poplack (1979) oppose keeping the two grammars disjoint:

> The evidence which seems most pertinent to this issue is the finding that code-switching generally does not entail pauses, hesitations, repetitions, corrections or any other interruption or disruption in the rhythm of speech. This is distinct from many bilingual situations marked by language interference, for example, and is at least partial justification for treating code-switched discourse as being generated by a single grammar based on the two monolingual ones.(p. 5)

However, evidence about the smoothness of speech cannot bear on the question of the existence of a single grammar. Poplack's evidence only demonstrates that mixing codes is not an erratic performance phenomenon, and that we should expect there to be a cognitive mechanism in charge of the competence for code-switching.

Poplack(1979) offers further argument for the mono-grammar model:

> ... segments are about as likely to be switched into English as into Spanish, providing further evidence for the suggestion that the code - switching mode proceeds from a single grammar." (p. 47)

Poplack seems to be suggesting that it is an empirical finding that the two languages play a symmetrical role in code-switching. In fact her data does not show whether the two languages indeed play a symmetrical role: rather she assumes it. For her, a segment "is switched" into one language if the segment preceding it was in the other language. For example, if the sentence consists of segments X-Y-Z where X and Z are in language L^1, and Y is in language L^2, Poplack reports two switches - which is already assuming symmetry, since it's conceivable to analyse the same data as one switch from L^1 to L^2.

Poplack (1979) further suggests that there are two constraints on code-switching. One is the free morpheme constraint: no switch may occur between a bound morpheme and the rest of the word. The other is the equivalence constraint: no switch may occur between two constituents whose relative order is different in the two languages.

To incorporate these constraints into G, Sankoff and Poplack (1979) build G in the following way:

To the union of the sets of category symbols from G^1 and G^2 they also add all category symbols marked for their origins, e.g. the category symbols of G will include NP, NP^1 and NP^2. Whenever the set of rules of G includes a rule of the form **A ---> ... B C ...** where a switch between B and C might violate one of the two constraints mentioned above, they mark B and C in the rule as B^1 and C^1 (or B^2 and C^2 as the case may be), which ensures that the terminals generated from B and C will be in the same language. The lexicon of G is the union of the lexicons of G^1 and G^2. Notice that because of the way G is built, the language L it generates contains L^1 and L^2.

For example, let L^1 be Spanish and L^2 be English. Then G^1 and G^2 contain the rule S ---> NP VP, whereas only G^1 contains the rule **S ---> VP NP**. Example (1) below violates the equivalence constraint, and indeed is not acceptable in L:

(1) * Arrived yesterday **mi mamà**.
my mother

To prevent generating (1), Sankoff and Poplack enter in G the rules **S ---> NP VP** and **S ---> VP^1 NP^1**. But then they realize that the following sentences, where a switch occurs between VP and NP, are nevertheless acceptable:

(2) a. **Llegò** **ayèr** my mother.
arrived yesterday

b. **Llegò** yesterday **mi** **mamà**
arrived my mother

Their solution is to adopt a superscript notation: $VP^{1,V}$, to ensure that only the V of VP^1 is obligatorily Spanish. The final form of the relevant rule is **S ---> $VP^{1,V}$ NP**.

As shown in Woolford (1980), this rule and others suggested in Sankoff and Poplack (1979) are set in each specific case to describe the relevant data, but they do not follow in any pricipled way from the two constraints. Sankoff and Poplack actually fail to show how to build the new grammar from the two monolingual ones.

Moreover, it seems to me hopeless to try to account for code-switching exclusively with strictly "local" constraints, that can only refer to the point of transition. For me there is a clear difference in acceptability between the following two sentences, both mixes of English and Hebrew:

(3) a. **ha** faculty **yictarxu** **kol** **exad** to give a talk
the will-have-to each one
'The faculty will each have to give a talk.'

b. *the faculty **yictarxu** **kol** **exad** to give a talk

Notice that (3b) has less switches than (3a), but it is (3a)

that is acceptable, not (3b)! (3a) has a switch between the determiner and noun in the subject NP that (3b) does not have. Both (3a) and (3b) have a switch between the subject and the VP, and another one inside the VP following the verb. It seems that the acceptability of these two switches depends on there being a switch inside the subject NP. But according to Poplack's constraints, the acceptability of a switch can only depend on the constituents immediatly adjacent to it. So we see that Sankoff and Poplack's account does not even meet observational adequacy.

Aravind Joshi has pointed out to me an additional problem that makes Sankoff and Poplack's approach unappealing on different grounds. His argument demonstrates that Sankoff and Poplack's approach predicts unnecessary complications in the processing of monolingual sentences. Assume G has the rules **A ---> B C D** and **A ---> B^2 C^2 E^2** (i.e. G^1 contains only **A ---> B C D** whereas G^2 contains both rules **A ---> B C D** and **A ---> B C E**). Sankoff and Poplack predict that in parsing a constituent of category A, even if it is part of a monolingual sentence of L^2, there would be an indeterminacy as long as D or E have not been reached, as to whether the correct parse is $[_A$ B C ...] or $[_A$ B^2 C^2 ...]. This is counter-intuitive, since there is no reason to believe that the processing of monolingual sentences should be affected just because the speaker or hearer can also process mixed sentences.

3.2. Woolford's approach

Woolford (1980) is in favour of keeping at least some components of the two grammars separate:

> The phrase structure rules from both grammars are freely mixed in the construction of trees; but, the lexicon and word-formation component of each grammar remain entirely autonomous. Terminal nodes created by a phrase structure rule unique to English cannot be filled from the Spanish lexicon and vice-versa, but lexical insertion is free in nodes created by phrase structure rules common to both languages. (p. 1)

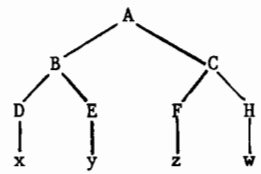
Woolford's proposal amounts to suggesting three autonomous grammars: the two monolingual grammars G^1 and G^2, and a new grammar G. G includes the phrase structure rules common to both languages and a lexicon that is the union of the two lexicons. In this system, unlike in Sankoff and Poplack's, L does not necessarily contain L^1 and L^2.

Notice that this system is not open to Joshi's criticism of

Sankoff and Poplack that I discuss above. If G^1 again has the rule **A ---> B C D** whereas G^2 has the two rules **A ---> B C D** and **A ---> B C E**, G will include this time only the rule **A ---> B C D.** Therefore the only possible parse for a constituent of type A, before D or E have been reached, is $[_A$ B C ...].

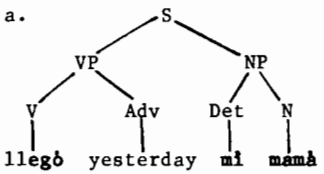
Unfortunately there are other problems. As Woolford herself notices, there is a problem with the interpretation of her expression: "terminal nodes created by a phrase structure rule" (see the quotation above from Woolford's paper). For example, if **A ---> B C** is a rule unique to G^1, whereas **B ---> D E** and **C ---> F H** are rules common to both G^1 and G^2, what lexicons can x, y, z and w in the derivation below be drawn from?

(4)

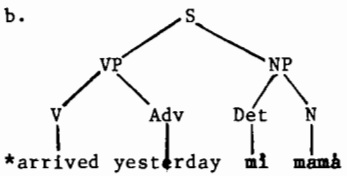


The question is relevant in examples like (1) and (2) above, where L^1 is Spanish and L^2 is English, and the observation was that y may be drawn from the English lexicon (even though **S ---> VP NP** is a rule unique to Spanish), and **x** may be drawn only from the Spanish lexicon (even though **VP ---> V Adv** is presumably a rule both of Spanish and of English).

(5) a.



b.



To counter the general problem, Woolford proposes to introduce constraints on lexical insertion that will percolate downward in the derivation. But their exact nature is left for future

research.

Woolford also tries to account specifically for examples such as (5), by arguing that S ---> VP NP is not a phrase structure rule of Spanish. She suggests that Llegó ayèr mi mamà is derived by a stylistic inversion rule from Mi mamà llegó ayèr. This inversion rule is part of the grammar of Spanish but not of English. The problem is of course that this rule should be allowed to apply to some mixed sentences -- since we want to derive (5a), but not to all mixed sentences -- since we want to block (5b). At this point Woolford says:

> If this analysis is correct, it appears that a sentence qualifies as Spanish for the purpose of the application of stylistic rules if its verb is Spanish. (p. 10)

So even though her analysis basically views the role of the two languages, following Sankoff and Poplack, as symmetrical, Woolford acknowledges here that there is a sense in which a mixed sentence qualifies as "belonging" to one of the languages, and not the other. But she does not make this sense explicit, nor does she give evidence to justify her assigning to the verb a major role in this asymmetry.

3.3. Joshi's approach

Joshi (1981) presents a model where this asymmetry is made explicit, and where no new grammar is introduced at all - neither to replace the two monolingual grammars, as in Sankoff and Poplack, nor in addition to them, as proposed by Woolford.

The burden of producing mixed sentences is assigned to the processor (parser or generator). The processor has access to both monolingual grammars, but "treats" them non-symmetrically. Within the domain of a sentence, the processor treats one language as the "host language" and the other language as the "guest language". While building a derivation tree, when a rule is drawn from the grammar of the guest language, every category symbol A appearing in that rule is entered in the tree as A', to distinguish it from the corresponding symbol in the host language.

The asymmetry is introduced by the rules that "relate" the corresponding categories from the two grammars. The processor has a meta-rule AxA', that allows it to "switch" from the host language L to the guest language L', but it does not have rules of the form A'xA. Notice that AxA' is the only kind of rule that is

allowed to mention category symbols from both grammars.

Now that the notion of asymmetrical "treatment" has been made precise, I shall refer to it as the Asymmetry Principle.

Joshi's paper discusses constraints on the meta-rule AxA'. In all his examples, L (the host language) is Marathi, and L' (the guest language) is English. The first constraint he discusses I will dub the Left Corner Constraint (LCC). Joshi suggests that this constraint is due to the nature of the parser, if we assume a left corner parsing strategy.

His relevant observations have to do with the distribution of acceptable mixes listed under (6). (It is important to keep in mind all along the discussion of Joshi's examples, that primed categories belong to English (the guest language), whereas the others belong to Marathi (the host language).)

(6)

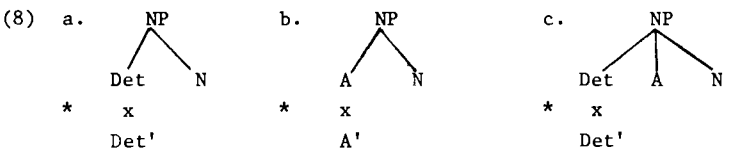
Det		N'	*Det'		N
A		N'	*A'		N
Det	A'	N	*Det'	A'	N
Det	A	N'	*Det'	A	N'
Det	A'	N'	*Det'	A'	N'
			*Det'	A	N

For example:

(7) a. **kā**hi chairs 'some chairs'
b. **ū**ncə boxes 'tall boxes'
c. **kā**hi tall **pē**tyā 'some tall boxes'
d. *some **khū**rčyā 'some chairs'
e. *tall **pē**tyā 'tall boxes'

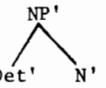
The generalization is that switches are not allowed on leftmost daughter nodes, for example:

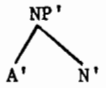
(8)

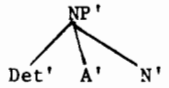


This is explained by the fact that a left corner parser, upon encountering A' will look for all rules of L' of the form

B' ---> A'... and therefore any subtree with a primed category at the left corner will be of the form:

(9) a. 

b. 

c. 

There are several problems with this explanation.[4] First, there is no reason why the parser would not also have access to rules of the form AxA' upon encountering A'.

The second problem was noted by Ken Ross.[5] The LCC predicts that if the first category parsed has a prime, then the whole sentence belongs to L' (intuitively, this is so since the first category is a "left corner" for the whole sentence). This prediction is clearly false. I give here an example (that I discuss below in more detail) where L is Hebrew and L' is English:

(10) [$_S$ [$_{\overline{N}}$ salesman] **Se oved kaSe** can make a lot of money]
who works hard
'A salesman who works hard can make a lot of money.'

Moreover, there are problems with the rule **NP ---> Det A N** that Joshi uses. For example it is hard to account with such a rule for the construction some old men and young women. Constructions of this type exist in Marathi just as they do in English.[6] It is preferable to treat A N as a constituent - Nom. We would therefore need the following rules (both for English and for Marathi):

---

[4]Joshi also presents an interesting alternative explanation, namely that categories corresponding to closed classes (e.g. determiners, quantifiers, prepositions etc.) are nonswitchable. This is a very plausible explanation from the point of view of the organization of the lexicon. But I shall not linger upon it, since it does not account for the nonswitchability of the adjectives in the examples above.

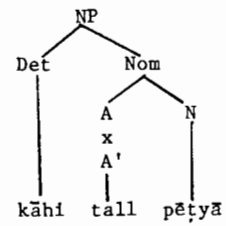
[5]In his discussion of Joshi's paper at the Workshop.

[6]The parallel structure in Marathi is the following: (I owe this information to Aravind Joshi.)

kāhi	**mhatare**	**purush**	**ani**	**tarun**	**bayka**
some	old	men	and	young	women

NP ---> Det Nom and Nom ---> A N. We also need NPpl ---> Nom for plural NPs of Marathi or English.

But now the example: kāhi tall pēṭyā becomes a problem for Joshi, since it is derived in the following way:

(11) 

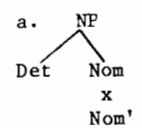
The problem is that there has to be a switch on the leftmost daughter of Nom.

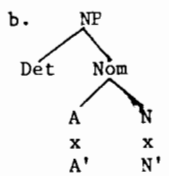
4. An Alternative Account for the Constraints on Code-Switching

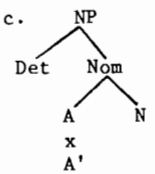
4.1. The Early Determination Strategy

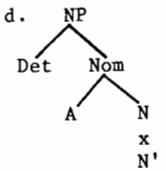
I now show how the problems that Joshi's account ran into can be solved if we make the following reasonable assumption: just as there is a meaning to saying that a sentence is in one of the languages even though it uses lexical items from the other, there is a meaning in saying the same for other constituents too. The language of major constituents such as S, NP, VP, PP is determined as early as possible in the course of parsing the constituent itself. The language of a mixed Nom (Nom is not a major constituent) is determined by the language of the NP, if that has already been determined.

Thus in (11), upon parsing the determiner kāhi, the language of the NP is determined to be the host language L, and consequently the language of Nom is set to be L. Any subsequent switch on Nom is free, since its language has already been determined. In (12) I list the permissible switches inside an NP that unequivocally belongs to the host language:

(12) a. 

b. 

c. 

d. 

On the other hand, in an example like (7e): * tall **pēṭyā** 'tall boxes', where the language of NP has not yet been determined upon encoutering Nom, the language of Nom will be determined by its leftmost element, i.e. by A, since the strategy is to establish the language of NP as soon as possible. In this particular case the language of Nom is determined to be English - but since English is the guest language, there is no way an English Nom can dominate a Marathi noun. (7e) is therefore unacceptable.

Notice that in major constituents, switches such as those in the Noms of (12b,c) could not take place. For assume **A ---> B C** is a rule of L, and **A' ---> B' C'** is a rule of L', where A is a major constituent. Upon encountering B', the parser fixes the language of the node dominating B' to be L', thereby achieving the same result as Joshi's LCC.

I will call the parsing strategy that I am advocating the "Early Determination Strategy" (EDS), and I informally formulate it as follows:

The Early Determination Strategy:
The parser seeks to determine as early as possible the language of the major constituent it is currently parsing.

Notice that there are testable differences between the predictions of Joshi's LCC and the predictions of the my EDS. I will now present a relevant test-case.

Let L now be Hebrew, and L' still be English. We have the following rules:

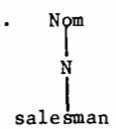
Hebrew	English
NP ---> (Det) Nom	**NP ---> Det Nom**
	NPpl ---> (Det) Nom
NP ---> NP R	**NP ---> NP R**
Nom ---> N A	**Nom ---> A N**

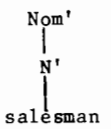
Consider examples (10), (13) and (14), all acceptable mixes in L:

(13) salesman **dinami** can make a lot of money
dynamic
'A dynamic salesman can make a lot of money.'

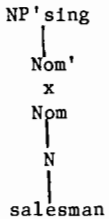
(14) salesmen **dinamiyim** can make a lot of money
dynamic(pl)
'Dynamic salesmen can make a lot of money.'

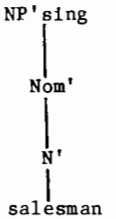
salesman is a noun of English; therefore the parser may, according to both the LCC and the EDS, build either of the subtrees below:

(15) a. 

b. 

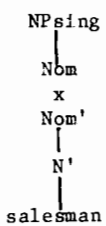
At the next stage, (16a) is blocked by the Asymmetry Principle, and (16b) is blocked by the grammar of English:

(16) a. * 

b. * 

(In (15a) and (16a) the parser is trying out the hypothesis that English is the host language, and in (15b) and (16b) - that it is the guest language.)

At this point Joshi's parser has to discard the sentence, since the LCC prohibits it from building a sub-tree like (17).

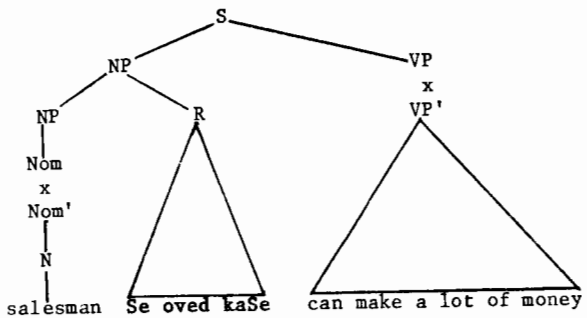
(17) 

But this outcome is observationally wrong, since (10) is a perfectly acceptable mix where the host language is Hebrew.

My parser, on the other hand, determines exactly at this point that NP is in Hebrew, therefore that Nom is in Hebrew, and -- since the language of the major constituent is now determined -- that (17) is an acceptable subtree.

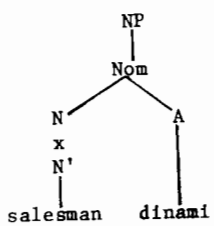
At this point it can also determine that the language of S is Hebrew, and produce (18) as a parse for (10):

(18)



In (13) there is an additional reason why the subject NP has to be Hebrew. Even its Nom can be only Hebrew,since in an English Nom the order would be A N rather than N A:

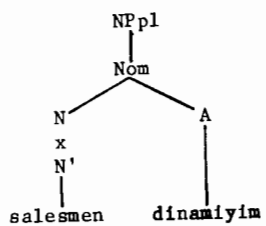
(19)



Again my parser gets the correct result since the EDS allows (19). Joshi's LCC prohibits (19).

In (14), the word order inside Nom is the only reason why the subject NP is Hebrew rather than English, since the omission of Det here does not distinguish between Hebrew and English:

(20)



Again the EDS yields the right result, whereas Joshi's LCC fails to do so.

4.2. Additional constraints

4.2.1. The problem of Hebrew and Spanish adjectives

Observe next that other constraints are needed in addition to the EDS. Notice the difference in acceptability between (21a) -- where the host language is Hebrew, and (21b) -- where the host language is English.[7]

(21) a. **iS mexirot** dynamic can make a lot of money
man sales

b. *a **dinami** salesman can make a lot of money.
dynamic
both:'A dynamic salesman can make a lot of money.'

According to the EDS both should be equally acceptable, since both involve the rule AxA' inside a Nom.

I would like to suggest that the difference between (21a,b) may be due to differences between agreement rules of Hebrew and English. According to this approach, Hebrew As can be plugged only into positions specified for gender and number; therefore, they cannot be plugged into English Noms. English As, on the other hand, can be inserted into any A position in the derivation tree.

Since the same differences in types of N and A agreement exist between Spanish and English, my approach could also explain the following:

(22) a. ? **una pelicula** interesting
a movie
'An interesting movie.'

b. * seven **chiquitas** houses
small(fem.pl.)

---

[7]It may be that I am unqualified to give judgements where the host language is English, since I am not a balanced bilingual, and the only code switchings that occur in my speech have English as the guest language. It has been pointed out to me by Aravind Joshi that my finding (21b) unacceptable could be an artifact of this biographical accident.

I am suggesting that what blocks switches such as (22b) are not considerations about differences of the order of constituents, as proposed by Sankoff and Poplack (1980), but considerations about agreement. The fact that word-order is not the only thing that distinguishes grammars of different languages seem to be neglected by these authors. Even Woolford attributes too much to word-order, as might be inferred from her treatment of the following examples from Labov (1971):

(23) a. **Yo no estoy** proud of it.
'I am not proud of it.'

b. * **Yo no** am proud of it

c. * **Yo estoy** not proud of it

Woolford's explanation is as follows: The order of the two elements in the English Aux is different from the order of these elements in the Spanish Aux. Depending on which Aux rule was used, only words from that language can be inserted.

I would like to suggest that again it is not the order of elements that is problematic, but the fact that Spanish may have no Aux node at all. Notice that any verb that takes an adjectival complement can appear in the position of estoy in (23a):

(24) **El no parece** proud of it.
'He does not seem proud of it.'

If there were an Aux node in Spanish, Woolford would have to predict that (25) is acceptable. ((25) is derived by using the rule **S ---> NP Aux VP** common to English and Spanish according to Woolford, and then using the English Aux rule.)

(25) * **El** does not **parece** proud of it.

4.2.2. The problem of PPs

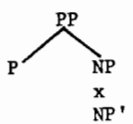
Another kind of unacceptable mixed sentence that the EDS (or the LCC) cannot account for is examplified in (26). In (26a) L is Hebrew and L' is English. In (26b) L is English and L' is Spanish.

(26) a. ***samti et ha-sefer al*** the table
I+put the book on

b. *I put the book on **la mesa**
the table

The structure of the PPs that occur in the sentences of (26) is shown in (27):

(27) *



Nothing in the EDS or LCC (and for that matter, nothing in Poplack's constraints) blocks (27).

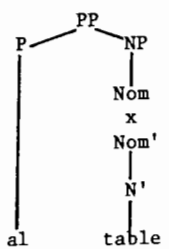
I do not have an articulated account of the unacceptability of (27) at the moment, but it seems to me that we can make sense of this phenomenon if we view prepositions as grammatical markers of case. From this point of view we would not expect (27) to be an acceptable mix, since we do not expect case markers in one language be coupled with NP from another language.

It is interesting to notice in this context that (28) is acceptable, where (26a) was not. In (28), as in (26a), L here is Hebrew and L' is English.

(28) **samti et ha-sefer al** table
I+put the book on

The reason is that the English N table in (28) is part of a Hebrew NP. The structure of the PP in (28) could be the following:

(29)



We now come across another indication that word-order is not

the interesting factor when accounting for code-switching. Marathi has postpositions rather than prepositions, but when L is Marathi, we get the same results as in (26a) and (28). Joshi (personal communication) has provided me with the following judgements:

(30) a. *some books var
on

b. land var
on

5. On the Notion "corresponding categories"

5.1. Relative clauses

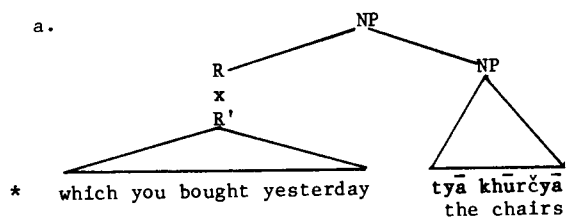
We now turn to another phenomenon described by Joshi. In Marathi, relative clauses precede the head NP, whereas in English it is the other way round. Joshi proposes rules like the following:[8]

(31)

Marathi	English
NP ---> R NP	NP ---> NP R
R ---> JO S/NP	R ---> WH S/NP

He then notes the following distribution of acceptability of mixed NPs:

(32) a.



* which you bought yesterday tyā khūrčyā the chairs

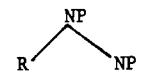
b. tyā khūrčyā which you bought yesterday
the chairs

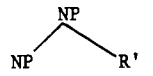
In Woolford's system, (32 a,b) come out as acceptable to the

[8]Where S/NP is a sentence with a NP gap (à la Gazdar).

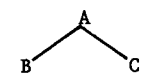
same extent. Sankoff and Poplack can of course add superscripts to the English rule to allow for a Marathi head NP, but (32b) remains a counterexample to their equivalence constraint.

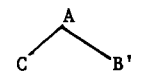
Joshi's LCC accounts for the unacceptability of (32a), and to account for (32b) he proposes a structural change: switching RxR' in the subtree (33a) results in subtree (33b).

(33) a. 

b. 

Ken Ross has suggested that there might be a generalization here to every case where L has a rule A ---> B C and L' has the rule A' ---> C' B'. In those cases BxB' would result in subtree (34a) changing into (34b):

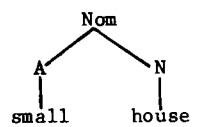
(34) a. 

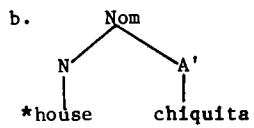
b. 

Pfaff (1979) has a couterexample to this tentative generalization:

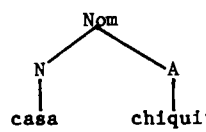
(35) * I went to the house chiquita
small

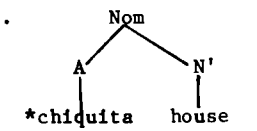
Ross predicts that (36a) changes into (36b), but his prediction turns out to be incorrect:

(36) a. 

b. 

We have already seen that Spanish adjectives may not plug into English Noms because of agreement problems. My consultant found the parallel case where L is Spanish just as unacceptable:

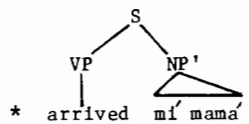
37) a. 

b. 

Ross's generalization again makes the wrong prediction in the

case of the English rule **S ---> NP VP** and the Spanish rule **S' ---> VP' NP'**. It wrongly predicts that we get the following:

(38)



It may be that Ross's generalization is valid only in cases where the relative order of the two constituents is strict in both languages. That would explain why it failed above: Spanish also has **S' ---> NP' VP'**, and there are cases where the A N order is reversed in both English and Spanish, e.g. the page above, una grande error.

Rather than trying to save Ross's generalization, I will propose here a different explanation for (32 a,b).

Notice what is entailed by the fact that in English, the head-NP precedes the relative clause: In parsing an English relative clause there has to be an NP already on hold to fill in the gap in the relative clause. In Marathi, on the other hand, the head-NP follows the relative clause; therefore in parsing Marathi relative clauses, the gap is expected to be filled later.[9]

This suggests to me that English relative clauses and Marathi relative clauses are not corresponding categories. I therefore propose the rules in (39) to replace Joshi's rules in (31).

(39)

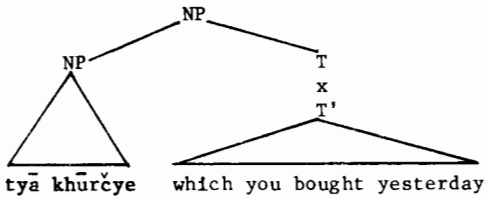
Marathi	English
NP ---> R NP	**NP ---> NP T**
R ---> JO S/NP	**T ---> WH S/NP**

The reason why (32a) is not generated is simply that there is no rule RxT'. To account for (32b) we need the assumption that a bilingual Marathi-English speaker, unlike a monolingual Marathi speaker, also has in his grammar of Marathi a rule **NP ---> NP T**. Intuitively, what this means is that the bilingual realizes that if Marathi had relative clauses that were parsed similarly to English relative clauses (Ts), they would have to follow the head

[9] I disregard here the fact pointed out to me by Aravind Joshi, that in Marathi, deletion of NP in the relative clause is optional.

NP. Notice that T is not expanded in Marathi, and therefore **NP ---> NP T** is never used in monolingual speech. But in code-switching, the rule TxT' can be used, and we get the following structure for (32b):

(40)



Further predictions of this explanation can be tested, for cases where the host language is English:

In the English rule **NP ---> NP T**, both the LCC and the EDS prohibit applying NPxNP' to the lower NP, whereas applying TxT' will block the derivation, since T is not expanded in Marathi. A bilingual also has the rule **NP ---> R NP** in his grammar of English. But because R is not expanded in English, and RxR' is permitted neither by the LCC nor by the EDS, **NP ---> R NP** will never yield permissible mixed sentences.

The prediction therefore is that no mixing can occur in an NP with a relative clause, when the host language is English. (The whole NP can of course be switched, if it is not itself the leftmost element of an S, say).

For example, I predict:

(41) a. ***jyā tu kāl vikat ghētlyās** the chairs are very beautiful
JO you yesterday bought

b. ***tyā khūrčye** which you bought yesterday are very beautiful the chairs

Aravind Joshi (personal communication) indeed finds (41a,b) unacceptable.[10]

[10] With the same qualification as in note 7, since Joshi reports that all mixed utterances that occur in his speech have Marathi as the host language.

5.2. On S and S'

The last constraint on switching discussed by Joshi has to do with lexical control. In constructions involving Vs and their complements, where V determines what COMP is appropriate in the S or VP complement, Joshi notices that in mixed sentences

> ...we have the resultant COMP as the COMP corresponding to the host verb 'plus' those realizations of COMP that correspond to the equivalent guest verb, excluding those which are in conflict with the host verb. We will call the resultant COMP as the 'minimal' union of $COMP_{host}$ and $COMP_{guest}$.(Joshi (1981) p.15 of abstract)

For example, the COMPs corresponding to the English verb feel are that and 0, as in (42a,b) respectively:

(42) a. I feel that he will return tomorrow.

b. I feel he will return tomorrow.

that happens to be "in conflict" with the corresponding Marathi verb, but 0 is not:

(43) a. ***malā wāṭat āhē** that he will return tomorrow.
to me feel

b. ?**malā wāṭat āhē** he will return tomorrow.

Therefore that is not an acceptable COMP for feel as a guest verb, whereas 0 is:

(44) a. ***malā** feel **hotə āhē** that he will return tomorrow

b. **malā** feel **hotə āhē** he will return tomorrow

Joshi also mentions that COMP by itself is not switchable; this follows, for example, from his LCC:

(45) ***malā wāṭat āhē** that **tō udyā parat yēil**
to me feel he tomorrow return will

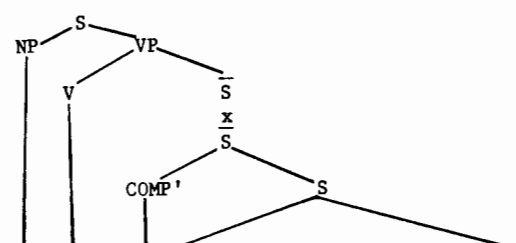
But notice that you don't really need the LCC to block (45), since it is already blocked by the minimal union principle. Pfaff (1979) has an example from English-Spanish code switching that is parallel to (45), but is acceptable:

(46) It goes without saying I think **que** along with the picketing we are doing a boycott.

que is probably not in conflict with the guest verb think. But there is another problem: It follows from the EDS (or the LCC) that L in (46) is English. The complement of think is switched to L' (Spanish), which accounts for que. But then why is what follows que switched back to English? This contradicts the Asymmetry Principle.

The answer, I think, lies in the fact that what follows que is a sentence. We have to allow for the roles of guest and host to be redefined at every new S. It is plausible to say that every sentence is by definition a sentence of the host language, and what changes is the assignment of the roles "host" and "guest" to the two languages. Accordingly, I shall from now on use only the symbol S for sentence, even when it appears in a rule of the guest language.[11] So the derivation tree for (46) is the following:

(47)



Poplack (1978) has a similar example:

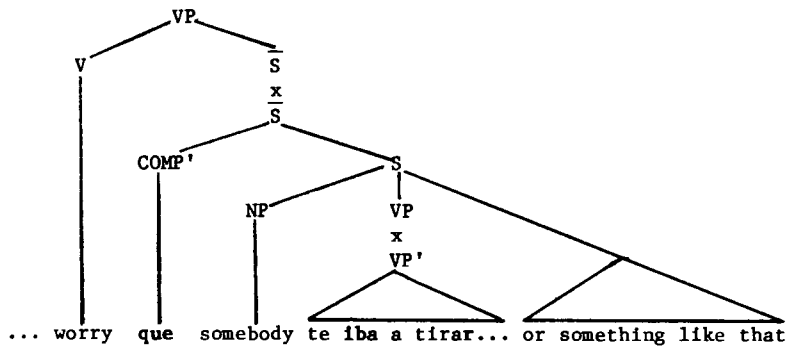
(48) You didn't have to worry **que** somebody **te iba a tirar con cerveza o una botella** or something like that.

(49) is a possible derivation tree for (48):

---

[11] Instead of doing that we could allow S'xS.

(49)



6. Conclusion

This paper argues that Joshi's "asymmetrical" approach to the languages participating in code-switching is superior to previous "symmetrical" approaches. Joshi proposes the notions of "guest language" and "host language", together with a meta-rule that "switches" from the host language to corresponding categories in the guest language. Constraints on the application of these rules are motivated by general processing strategies, such as left corner parsing.

In the framework of Joshi's system I have contributed the following observations:

I. The parsing of mixed sentences is guided by the need to determine as soon as possible the language of major constituents.

II. Grammatical processes characteristic to a language (e.g. gender and number agreement, case marking etc.) are factors that may explain unacceptable mixed sentences.

III. The notion of "corresponding categories in two languages" is not always clear.

There are, of course, many additional interesting open questions on code-switching. I think that discussing them in Joshi's framework will turn out to be extremely fruitful.

REFERENCES

Joshi, A.K. (1981) Some Problems in Processing Sentences with Intrasentential Code Switching. Paper read at the University of Texas Parsing Workshop, 1981.

Labov, W. (1971) The Notion of 'system' in Creole Languages. In D. Hymes (ed.) Pidginization and Creolization of Languages. pp. 447-72. Cambridge: University Press.

Pfaff, C. (1979) Constraints on Language Mixing. Language 55:291-318.

Poplack, S. (1978) Syntactic Structure and Social Functioning of Code Switching. In R. Duran (ed.) Latino Discourse and Communicative Behavior. New Jersey: Ablex Publishing Corp.

Poplack, S. (1979) Sometimes I'll Start a Sentence in Spanish Y TERMINO EN ESPANOL. Working Paper 4. New York: Centro de Estudios Puertorriquenos.

Sankoff, D. and S. Poplack (1979) A Formal Grammar of Code-Switching. Working Paper 8. New York: Centro de Estudios Puertorriquenos.

Woolford, E. (1980) A Formal Model of Bilingual Code-Switching. Unpublished ms. (MIT)