The production of code-mixed discourse

David SANKOFF
Centre de recherches mathématiques, Université de Montréal
CP 6128 Succursale Centre-Ville
Montréal, Québec H3C 3J7
sankoff@ere.umontreal.ca

Abstract

We propose a comprehensive theory of code-mixed discourse, encompassing equivalence-point and insertional code-switching, palindromic constructions and lexical borrowing. The starting point is a production model of code-switching accounting for empirical observations about switch-point distribution (the equivalence constraint), well-formedness of monolingual fragments, conservation of constituent structure and lack of constraint between successive switch points, without invoking any "code-switching grammar". Code-switched sentence production makes alternate reference to two virtual monolingual sentences, one in each language, and is based on conservative conditions on language labeling of constituents, together with a constraint against real-time "look-ahead" from one code-switch to the next. Selective weakening of model conditions can produce (i) the type of palindromic (or portmanteau) construction occasionally occurring e.g., in switches between prepositional and postpositional languages, (ii) the switching by "insertion" of very specific kinds of constituent reported e.g., for French noun phrases in switching with Arabic and, most important, (iii) lexical borrowing. Borrowing can create ambiguity as to language membership of sentence items, but the model predicts where this can be resolved, and the confirmation of these predictions, based on empirical studies of inflectional morphology, validates key aspects of the model.

Introduction

Communities of bilinguals tend to evolve a conversational mode where elements of both languages appear in the same interaction and even in the same sentence despite the fact that all participants may be competent in either of the two languages. Whether this mode is used in preference to monolingual discourse depends on the type of interaction, the participants, the subject of conversation and many other factors. The grammatical nature of code-mixed discourse, however, tends to be very specific to the community and varies widely among bilingual communities, even among communities which share the same pair of languages. Empirical research has isolated four clearly distinct processes which may be responsible for mixing to different extents in different communities — code-switching, nonce borrowing, specialized incorporation and interference. None of these processes requires the deformation, alteration or convergence of either of the two constituent languages at the syntactic, lexical, morphological, phonological, or semantic levels at the moment the mixing occurs. Except for code-switching, however, they may all lead in the long term to lexical expansion in one or both of the languages.

This paper is a contribution to a coherent formal account of code-mixing which integrates all of these processes[1], though only code-switching and borrowing will be considered here. This is based on a series of empirical studies which now allows us to distinguish between them structurally and quantitatively. Our starting point will be a recent formal characterization of a equivalence-point code-switching (Sankoff,

[1] The analysis of code-mixing is a controversial subject with respect to several aspects: Is all code-mixing — borrowing, switching, interference — really the same process? Do languages involved in code-mixing tend to converge syntactically? Are patterns of code-mixing predictable or explicable by theories of (monolingual) grammar? We assume a negative response to all these questions and refer to the literature for more detailed discussion.

1998). We will then extend this to two rarer code-switching mechanisms, and finally to lexical borrowing, the most frequent type of code-mixing.

1 Code-switching

1.1 "The facts"

The modern motivation for studying code-switching was initially to explain the observation that in bilingual communities, speakers tend to switch from one language to another intrasententially at certain syntactic boundaries and not at others (Gumperz & Hernandez, 1969). The first general explanation to account for this distribution was Poplack's (1978, 1980) argument that switching should be favored at the kinds of syntactic boundaries which occur in both languages, thus avoiding word order that might seem unnatural according to one or both grammars: *the equivalence constraint* (see also Lipski, 1977; Pfaff, 1979). Despite criticism of this approach (Rivas, 1981; Woolford, 1983; Di Sciullo *et al.*, 1986; Pandit, 1990; Myers-Scotton, 1993; Belazi *et al.*, 1994; Mahootian & Santorini, 1994), it has been successfully used to account for code-switching in Spanish-English (Poplack, 1978,1980), Finnish-English (Poplack *et al.*, 1987b), Arabic-French (Naït M'Barek & Sankoff, 1988), Tamil-English (Sankoff *et al.*, 1990), Fongbe-French (Meechan & Poplack, 1995), Wolof-French (Poplack & Meechan, 1995), Igbo-English (Eze, 1997) and many other bilingual communities.

Other fundamental facts about code-switched sentences include the *well-formedness of monolingual fragments* within such sentences — true whether a fragment constitutes a complete constituent or stretches across two or more (possibly incomplete) constituents, the *conservation of constituent structure*, and the *unpredictability* of switching — even if we can determine where a code-switch can occur and where it cannot, there is no way of knowing in advance for any site whether a switch *will* occur there or not. In particular, if a switch occurs at some point in a sentence, this does not constrain any potential site(s) later in the sentence either to contain another switch or not to — there are *no forced switches*.

1.2 A production approach

We do not assume that the mechanisms of switching from one language to another can be deduced entirely from the general principles of monolingual grammars[2]. Thus we do not analyze the distribution of intrasentential switch points in terms of a grammar of any of the types ordinarily used for accounting for single languages, but by means of a left-to-right process that refers to two well-formed monolingual sentences (i.e. each satisfying the constraints of an "ordinary" monolingual grammar for one of the two languages) in producing monolingual sentence fragments and in evaluating potential switch points between these fragments.

Our model is based on the assumption that bilinguals are fully competent in their two languages, that there is no convergence of the two monolingual codes even during bilingual discourse, and that a code-switched sentence consists of fragments of two monolingual sentences (each one a translation of the other) pieced together. This is done first through an otherwise unconstrained production model that simply copies part of one monolingual process followed by part of the other in such a way that constituent structure is conserved. The resulting process satisfies neither equivalence nor unpredictability. By adding two very simple rules for labeling some constituents according to their language, the existence of a consistent labeling — which can be easily monitored during real-time linear production — turns out to guarantee both equivalence and, for most situations, unpredictability.

[2] While formal theories of grammar may well account for monolingual language in terms of general linguistic principles, there is no reason to believe that processes which juxtapose two languages can be explained in exactly the same way. The reasons implicit or explicit in attempts to do so have to do with explanatory economy, either of individual linguistic competence or of linguistic theories. This seems specious since both are based on a notion of a "hard-wired" human linguistic faculty evolving in prehistoric monolingualism. Experience in widely diverse speech communities suggests instead that code-mixing strategies, including code-switching, evolve in the life-time of particular communities, are only partly dependent on linguistic typology of the two languages and exhibit widely different patterns of adapting monolingual resources for incorporating linguistic innovation.

1.3 Hierarchy and linearity.

That monolingual fragments are not coextensive with entire constituents is problematic for any model relying on hierarchical relations for deciding well-formedness, since such models are designed primarily to ensure well-formedness of entire constituents, monolingual or bilingual. They cannot ensure that adjacent same-language parts of neighboring constituents are compatible (i.e. yield a well-formed fragment when juxtaposed), since these parts may not even be in the same language as the rest of the constituents that contain them (Muysken, 1995). For example, an earlier model (Sankoff & Mainville, 1986), using the context-free grammars of two languages to account for code-switched sentences satisfying the equivalence constraint, could not ensure the well-formedness of monolingual fragments, for the very reason Muysken has pointed out.

This problem is at the core of the conflict between hierarchical and linear modes of explanation. We will resolve it by ascribing ultimate responsibility for the well-formedness of monolingual fragments to production-level processes. These fragments, arbitrary substrings of pre-constructed well-formed monolingual sentences, are pieced together during linear production in a way which corresponds to the other general observations about code-switching and which is essentially neutral with respect to theories of monolingual grammar.

1.4 The syntactic model

We are interested in seeing how two hierarchically structured languages resolve their word-order differences during intrasentential code-switching, without the confounding effects of other linguistic phenomena. Thus we will construct a model where the only differences between two languages have to do with word order and the (phonological) form of lexical items, and work out the logical consequences for code-switching of various assumptions and constraints on the production of bilingual sentences. Linearity and hierarchy are the key structural aspects here[3], and the simplest class of recursive grammars accommodating these properties is that of context-free grammars.

Consider a context-free grammar consisting of a set C of "categories" or non-terminal symbols, including one distinguished symbol s (for "sentence"); a set of terminal symbols T ("lexical slots"), none of which are also in C, a lexicon L which is basically a set of words and an indication of the kind of lexical slot each word can fill, and a set of rewrite rules R of form $c \rightarrow v_1 \cdots v_n$ where c is a symbol in C, and the string on the right hand side $v_1 \cdots v_n$ consists of one or more symbols in T or C. A sentence is derived by writing s, then rewriting s by the string $u_1 \cdots u_m$ on the right hand side of any rule in R of form $s \rightarrow u_1 \cdots u_m$, then rewriting any u_i which is non-terminal by some rule of form $u_i \rightarrow w_1 \cdots w_p$ and so on. Whenever a lexical slot appears in the string it can be filled with words of the appropriate category from L. When there are no more non-terminal symbols (and R must be such that this is always possible), the derivation stops, and the current string is just a sentence generated by the grammar.

We will make use of phrase structure tree representation for sentences and their constituents. Each symbol appearing in the derivation is represented by a node of the tree; it dominates the constituent of which it is the highest node, and may be used to represent that constituent. (Each terminal node is itself a constituent.)

Note that the words in each constituent form a (contiguous) substring of the sentence. We define these, and all other contiguous substrings of a well-formed sentence string to be *well-formed fragments.*

In order to speak of code-switching between two different grammars, it is necessary to have some connection between the categories of one and the categories of the other[4]. We make

[3] In focusing on the relationship between word order and hierarchy, we are choosing a model not adapted to the treatment of tags and moveable elements such as many adverbials (whose "switchability" is uncontroversial), nor of remote relationships such as discontinuous constituents, internal co-reference, etc., whose effects on code-switching, if any, have never been systematically documented. On the other hand, context-sensitive phenomena such as subcategorization (Bentahila & Davies, 1983), cliticization or certain deletion processes also escape the scope of context-free modeling, as do other null elements, agreement rules and other features which may, in particular communities, be important to understanding switch sites (Muysken, 1995).

[4] In natural languages, such correspondences will usually be imperfect (Muysken, 1995), but this is peripheral to our interest in word order.

the strong assumptions of *lexical translatability* and *categorial congruence*, meaning that there is a one-to-one correspondence between the lexicon L_A of language A and the lexicon L_B of language B, though the words are all recognizable as coming from one language or the other. We use the same categories C and lexical slots T for both languages. Furthermore we assume *grammatical congruence*: there is a one-to-one connection between the rulesR_A of language A and R_B of language B — if R_A contains a rule $c \to v_1 \cdots v_n$, then R_B must contain a rule $c \to u_1 \cdots u_n$, where each symbol in $v_1 \cdots v_n$ has its counterpart in $u_1 \cdots u_n$, and vice-versa, though the order of the terms in one string will not in general correspond to the order of the terms in the other. Finally, for convenience, we will assume *fixed word order*, that is if $c \to v_1 \cdots v_n$ in a given grammar, then there may be other rules rewriting C in that grammar, but none where the right hand side contains exactly the same set of symbols $v_1, \cdots, v_n$.

To simplify our presentation in this article, we do not allow ambiguity in our grammars. Not only must each monolingual sentence be derivable in exactly one way, but each rule may contain any one symbol only once on its right hand side. These conditions may be relaxed, as long as there is a way of identifying corresponding symbols in corresponding rules in the two grammars.

In comparing the structure of two sentences, we say that they have the same constituent structure if there is a one-to-one correspondence between their constituents such that if x in one sentence corresponds to y in the other, then the (unordered) set of subconstituents of x corresponds to the set of subconstituents of y.

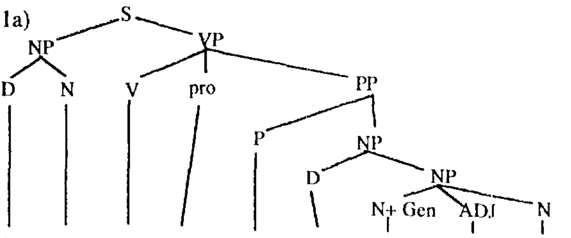
The consequences of our assumptions are summarized in:

Theorem 1 [5] *Every sentence in language A has a unique counterpart in language B with the same constituent structure and whose lexical items are translations of those in the sentence of language A.*

Examples (1a,b) are two (fictitious) sentences in English and French which we may imagine to be counterparts of each other according to some grammatical analysis of the two languages, in the sense of Theorem 1:

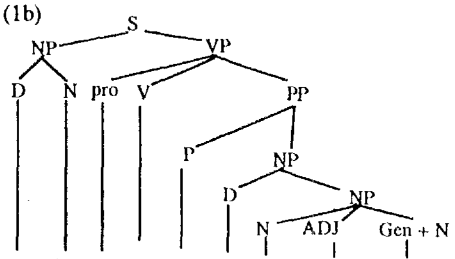
Despite differences in word order, the constituent structure is identical and the lexical items are word-for-word translations (without quibbling about the questionable lexical status of the reflexive clitic and the genitive particle and the somewhat different internal structure of determiner in the PP).

(1a)



The brothers wash themselves with some of Grandma's remarkable soap

(1b)



Les frères se lavent avec du savon remarquable de grand-maman

1.5 The production model.

In the model, the production of a code-switched sentence presupposes the existence of two virtual sentences, one in language A and one in language B, counterparts of each other as in Theorem 1. For each[6] of (2a,b,c,d) the pair of virtual sentences is the one illustrated in (1a,b).

(2a) The brothers wash themselves | avec du savon remarquable de grand-maman
(2b) Les frères se lavent | with some of Grandma's remarkable soap
(2c) The brothers | se lavent avec | some of Grandma's remarkable soap
(2d) The | frères | wash themselves | avec | some of Grandma's remarkable soap

Given the two virtual sentences in languages A and B, the code-switched sentence is produced by taking part of one of them, followed by part

[5] Proofs of all theorems are given in Sankoff (1998)

[6] Our examples of English/French mixing in this paper are fabricated, and their well-formedness (or not) asserted, solely to illustrate our arguments; they do not constitute empirical data.

of the other, and so on, without using any word (or its translation) more than once, until every lexical element (or its translation) has been used up.

The idea of using virtual sentences is an extension of concepts implicit in Poplack's original discovery (1978, 1980) of the importance of equivalence sites to code-switching and is what distinguishes it from attempts to account for code-switching using purely distributional data — examples of sentences thought to be either well-formed or not. It implies the comparison of the sentence actually produced with what "could have been said" in either of the monolingual modes. Though the comparative data are of course not directly accessible, since only one sentence is uttered, controlled inference about unrealized possibilities is consistent with rigorous methodology (cf the notion of the linguistic variable (Labov, 1969; Sankoff, 1988).

Postulating two complete virtual sentences, however, is an analytical convenience. All that would really be needed in a more realistic (and complicated) analysis, are the parts of each sentence that are actually used plus some additional details about the constituent within which a switch occurs. The consequences of this device, and its realism, lie largely in the way the monolingual fragments are produced "on the fly" by the monolingual grammars, however these grammars are conceived in theory. Indeed, we need not refer to any particular linguistic theory for this aspect.

The above process produces not only plausible code-switched sentences such as those in (2), but also any combination of elements in any order as in (3).

(3) de grand-maman | the with | remarquable frères | soap themselves wash some of

To arrive at an empirically and conceptually satisfactory model, we must add constraints. The first constraint is motivated by the empirical observation that each monolingual fragment in bilingual discourse tends to be well-formed in its lexifier language. We will assume that the production of the sentence starts with a word in (either) one of the virtual sentences, and copies successive words from left to right in that sentence without skipping any until there is a code-switch to some word in the other virtual sentence. From this point in the other virtual sentence, production continues from left to right, and so on. When the left-to-right production arrives at a word (such as *se* after *frères* in example 4) which has already been used in the current or the other language (also *some of* after *with*) or at the end of one of the virtual sentences, there must be a switch to the other virtual sentence or, if all the words have been used (after *remarkable*), the production must stop.

(4) du savon | wash themselves with | les frères | Grandma's remarkable

Theorem 2 *The monolingual fragments in a code-switched sentence produced by left-to-right copying are well-formed.*

The *left-to-right* assumption ensures that monolingual fragments are well-formed, as illustrated in (4), as well as (2). By itself, however, it does not constrain how the alternating fragments in one language and the other are related; indeed it allows them to be juxtaposed in any order, as long as the fragment languages alternate; sister elements in the same constituent in a virtual sentence may find themselves remote from each other in the code-switched sentence, as with *remarkable* and *savon* in example (4). As mentioned in Section 1.1, however, empirical research confirms that constituent structure, insofar as content and embedding or nesting relations are concerned, is conserved even if the constituent contains a code-switch.

To conform to this observation, we make a second assumption, that once the production process enters or switches into a constituent, it must exhaust all the lexical slots in the constituent, in one language or both, before returning into a higher-level constituent or entering a sister constituent. In other words, each time it enters a deeper, or more nested, subconstituent, it cannot exit from it until that subconstituent is exhausted. Note that this assumption is independent of whether or not the production follows the left-to-right process described above, so that the sentence (5) satisfies *nested first*, but not *left-to-right*.

(5) (wash((some of (Grandma's soap remarkable)) with) themselves) | (frères les)

Theorem 3 *Lexicalizing constituents according to* nested-first *is a sufficient condition for* *conserving the same constituent structure in the code-switched sentence as in the two virtual sentences.*

The *nested first* condition and the *left-to-right* assumption are independent, in the sense that neither implies the other, as is clear from (4) and (5). Neither excludes the other, and together, as in (6) and (2), they produce code-switched sentences with well-formed monolingual fragments and the same constituent structure as the virtual sentences.

(6a) (se lavent | (with (some of (Grandma's remarkable soap)))) | (les frères)
(6b) (frères | the) | ((avec | (some of (Grandma's | savon remarquable))) | wash themselves)

1.6 The language of constituents and subconstituents.

The model in Section 1.5, though it produces sentences like those in (2) and (6) with some desirable properties, is not complete. With only the two conditions, certain configurations may occur, such as those in (6), that are clearly unrealistic. For example, if the monolingual fragment being copied includes a word which must be positioned finally in a constituent, like *frères* in (6b), and if the words of the constituent in one virtual sentence or the other have not yet been used up, then an immediate code-switch, to *the* in this instance, is obligatory to satisfy *nested first* — the monolingual fragment cannot continue, even though it may have a natural continuation into another constituent. This, and other instances of forced code-switching, are clearly not phenomena observed in real bilingual discourse.

Another type of construction not found in natural bilingual corpora but permitted in the simple production model might include a code-switched sentence which begins with a fragment of the virtual sentence in language A which would never occur in sentence-initial position in monolingual discourse in language A, such as se lavent in (6a). More generally, if there are several sister subconstituents in a constituent, they may be permuted in any order, as long as there is a switch between each adjacent pair. Still another anomalous output from this model, even if the two monolingual grammars contrast completely, is a code-switch between every adjacent pair of words in the sentence.

Thus the output of the production process as is hitherto formulated seems unduly constrained from the production point of view (by forcing switches) and not constrained enough from the perspective of the output structure. We thus come to the main point of this section: short of the equivalence constraint itself, can we motivate some constraint to account for the observation that switching occurs almost exclusively at equivalence points (notion to be formalized in Section 1.6.1) and virtually all equivalence points seem to be eligible switch points? And can this be done in such a way that during production, the model speaker avoids any switching which will obligatorily require compensatory switches later on in the sentence construction (switch planning or forcing)? Furthermore, can we do this without referring to facts of particular languages, properties of particular grammatical categories, or even the mechanisms of particular theories of monolingual grammar?

Our approach to this problem is to postulate a limited degree of structural monitoring. Monitoring the monolingual fragments for well-formedness is uncontroversial, and we need not enter into the details. What we propose for monitoring at the switch points is the "language label" of the constituent in which the switch occurs and of its immediate subconstituents.

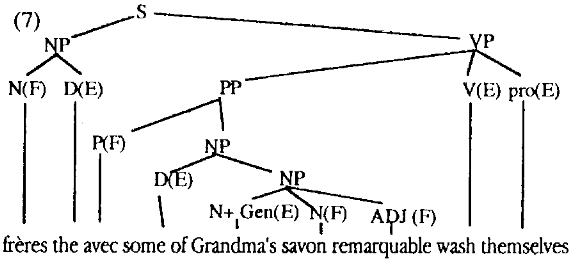
We first ask: what parts of the constituent structure of the code-switched sentence may be with certainty ascribed to one language or the other only, and hence should be labeled accordingly? Certainly (i) all terminal symbols — lexical slots — are labeled according to whether the word filling the slot comes from L_A or L_B, since all words are identifiable as to their language, by definition in our model of congruent grammars. (ii) At the constituent level, any constituent, all of whose immediate subconstituents have the same label, should itself have this label. Anything else would be inconsistent. We will propose a third criterion for labeling constituents, but we first prove the following:

Theorem 4 *Any non-terminal node carrying a label must have at least one immediate descendant node which has this same label or is unlabeled, and one descendant (possibly a lexical slot) which has the same label.*

Different instantiations of this model may actually specify that certain subconstituents "inherit" the label — in one theory the determiner may inherit the label of the noun phrase, in another theory it may be the noun itself.

Requirement (ii) depends on constituent content, but not on constituent order, so that it applies meaningfully to any sentence satisfying *nested first* such as (5) or (6b). The labeling of (6b) is illustrated in (7).

(7)


S
NP
VP
N(F)
D(E)
PP
V(E)
pro(E)
P(F)
NP
D(E)
NP
N+Gen(E)
N(F)
ADJ (F)
frères the avec some of Grandma's savon remarquable wash themselves

To constrain the order, we are motivated to try to exclude situations such as a declarative sentence which begins with a well-formed verb phrase entirely in English followed by a subject noun phrase in another language. More generally, (iii) any subconstituent which is out of rank order position among its sister subconstituents according to one of the languages, must receive the label of the other language. For example, if the languages A and B are SVO and VSO, respectively, and the code-switched sentence has order VSO, then the labeling should be $V^B S^B O$; if the code-switched sentence were SOV, then condition (iii) could not be satisfied, since O is out of order according to both languages, as is the V.

If conditions (i)-(iii), all of which are well-motivated, cannot be simultaneously satisfied, the code-switched sentence cannot be considered well-formed. Thus in our hypothetical example with a sentence-initial English verb phrase, requirements (ii) and (iii) conflict, so the sentence is not well-formed. And in example (7) condition (iii) cannot be satisfied with respect to any of the categories lexicalized by *frères, the, savon, remarquable, wash* and *themselves*, nor the PP node — according to the extremely constrained grammars responsible for examples (1a,b). On the other hand, each of examples (2a,b,c,d) satisfy all of the conditions (i)-(iii).

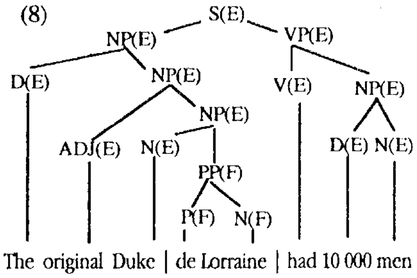
All that a speaker monitors is the language label of the constituent in which a potential switch occurs and that of its subconstituents.

No additional labeling is warranted within this framework, though other treatments of code-switching all have their own particular way of assigning a label to each and every constituent (Rivas, 1981; Woolford, 1983; Joshi, 1985; Di Sciullo *et al.*, 1986; Myers-Scotton, 1993). In particular, we would claim that there is no conceptual justification or need for postulating an underlying (or "matrix") language for the entire sentence itself when this is not motivated by criteria (ii) and (iii).

1.6.1 The equivalence constraint.

In the string of words which constitute a code-switched sentence in our model, there is no problem in identifying where a fragment in language A stops and one in language B starts. On the constituent level, however, it is not as obvious where this switch should be located and sometimes even whether or not there is an inter-constituent switch, as in (8).

(8)



This is an example of string-level code-switching with and without corresponding constituent-level switches. The rule NP→ ADJ+NP in English and NP→ NP+ADJ in French results in the lowest NP being labeled E, by requirement (iii). The higher NPs, the VP, the object NP and the S are labeled E, and the PP labeled F because all of their subconstituents are (requirement (ii)). The code-switch between *Duke* and *de* is also a constituent-level switch between *Duke* and the PP, but the switch between *Lorraine* and *had* is not reflected by a switch between the highest NP and its sister VP since they are both labeled E.

In general, what constitutes a code-switch between two adjacent sister constituents? The only reasonable answer is that one constituent is labeled A and the other B. What happens if two differently labeled sister constituents are separated by one or more unlabeled constituents? Once again it is clear that there has been a code-switch at the constituent level, but the site cannot be pinned down, other than by saying that it occurred in the interval between the two labeled constituents. Note that there also must be switches at some lower levels within each of the intervening unlabeled constituents; otherwise they could not be unlabeled, by criterion (ii).

We can now state the equivalence constraint. Consider the corresponding rules for ordering the n subconstituents of a given constituent in language A and language B. If the sets of the first i symbols on the right-hand side of the rules are different in the two grammars (and hence the set of the last $n - i$ symbols are also different, since the two rules are congruent), then the equivalence constraint prohibits a code-switch between the i-th and $i + 1$-st subconstituents. Otherwise the boundary between the two subconstituents is an equivalence point and a code-switch is permitted. In the case $n = 2$, this reduces to the prohibition of a code-switch between the two subconstituents unless the two languages order them in the same way.

1.7 Proof of the equivalence constraint.

The production model in Section 1.5 and the labeling rules in Section 1.6 ensure that monolingual fragments are well-formed, constituent structures are correct and that no constituent labeled X appears within a higher constituent in a rank order position not permitted in language X. This does not mean that the equivalence constraint holds. Consider for example the rules $c \dashrightarrow xyz$ and $c \dashrightarrow zyx$ in languages A and B respectively. Then the model as it is now constituted would permit the constituent order $x^A y^B z^A$, with two constituent-level code-switches, both of which violate the equivalence constraint. (cf *Grandma's* | *remarquable* | *soap* or *savon* | *remarkable* | *de grand-maman.*)

In one important case, however, equivalence always holds. Monolingual grammars have *binary* constituent-subconstituent structure if there are at most two symbols on the right-hand side of every rule. Then the following holds:

Theorem 5 *Given a well-formed code-switched sentence where the monolingual grammars have binary constituent-subconstituent structure. If two sister subconstituents are labeled A and B, respectively, the code-switch between them satisfies the equivalence constraint.*

As we have seen, however, the theorem may not hold if rules may have more than two terms on their right-hand sides. The counter-example shown above, for example, represents the insertion of a language B subconstituent into an otherwise language A constituent. This requires two code-switches, one before and one after the inserted subconstituent. If a code-mixing strategy were to be based on the insertion of constituents in this way, every code-switch before an insertion would require the speaker to plan for an appropriate second code-switch later on in the sentence.

While many types of relatively complex forward planning must be incorporated into monolingual production models, the distribution of code-switches in bilingual corpora is more consistent with a hypothesis of the independence of successive code-switches: *no forced switches* (or no planning). Where a switch takes place in between two constituents, well-formedness of the code-switched sentence cannot depend further switches later on in the left-to-right order[7].

Theorem 6 *Given a well-formed code-switched sentence. If two sister subconstituents are labeled A and B, respectively, and there is no labeled subconstituent between them, then under* no forced switches, *there must be a code-switch satisfying the equivalence constraint in the interval between the labeled subconstituents.*

2 Relaxing the constraints.

2.1 Repetition-translation.

The model in Section 1 precludes forms such as **se lavent* | *themselves* where corresponding items from *both* virtual sentences (in this case *se* and *themselves*) appear. Such repeat-translations (also called *portmanteau* or *palindromic* constructions do occur, albeit rarely, in some corpora.

Example (9) is drawn from the Finnish-English code-mixing corpus of Poplack *et al.* (1987b). We assume, following these authors' arguments, that this sentence consists of three fragments, with code-switches immediately before and after the English preposition *to*. The

[7] There are some exceptions: these will be discussed in Section 2.2.

ellative case-marked *kidneystä* and illative *aortaan* are formed of borrowings from English and behave as native items (e.g. there is no English determiner preceding them as would be expected within English fragments; rather they manifest null determiners and case-marking characteristic of Finnish — see Section 3).

(9) Mutta se oli kidneystä | to | aortaan.
but it was kidney -el. *aorta*-il.
'But it was from the kidney to the aorta.'

The interesting aspect of this example is that *to* and the ellative marker *-an* play identical roles and only one should have appeared in the adpositional phrase containing *aorta* according to our model. The same, highly bilingual, speaker produced a similar example in (10).

(10) Ja sitten, uh, missä hän n- | at | yliopistossa
and then where she n- university-in.

otti, niin kuin, | art history.
took-3p., *like,*

'And then, uh, where did she- at university
she took, like, art history.'

Again, the inessive marker *-ssa* has the same function as the English preposition *at* and only one of them should have appeared, according to our model.

This type of construction, involving the redundant use of functionally identical words, is rare — examples (9) and (10) are the only two in the Finnish-English corpus — and, as can be seen in (10), tends to evidence production-level difficulties (hesitations, autocorrection, etc.). Nishimura (1986) presents several similar examples involving adpositional phrases for Japanese-English code-mixing.

It is also possible to find occasional instances of redundant verb use in SVO/SOV mixing — producing a SVOV structure. Examples (11) and (12) below are drawn from the Tamil-English code-mixing corpus of Sankoff *et al.* (1990).

(11) They gave me a research
grant | kodutaa.
gave-3p.-pl.-past
'They gave me a research grant.'

(12) I was talking to |
oru orutanooda peesindu iruntein.
one person-com. *talk*-cont. *be*-1p.-sg.-past
'I was talking to a person.'

Still other examples from the same corpus combine redundant verb plus complementizer for propositional complements:

(13) I think it's the European
influence | nu ninaikirein.
that think-1p.-sg.-pres.
'I think that its the European influence.'

Precisely how does this violate the conditions of our model, and can they be relaxed to accommodate it? Clearly the first postulate violated (Section 1.5) is that the "...code-switched sentence is produced by taking part of one of [the virtual sentences], followed by part of the other, and so on, without using any word (or its translation) more than once, until every lexical element (or its translation) has been used up." What if we changed the latter part of this to "... and so on, until every lexical element (and/or its translation) has been used once?" This general extension allows for the use of any element as well as its translation in the same sentence. More limited extensions, where only specified lexical classes may occur in both languages, would be more consistent with actual speech behaviour.

With these kinds of changes, there is no difficulty in retaining the *left-to-right* and *nested first* assumptions, as well as consituent labeling criteria (i) and (ii). Condition (iii), formulated as it is in terms of the rank order of constituents, must be worded differently to capture the basic idea that a constituent is labeled according to language A whenever it is out place according to a rule of language B, and vice-versa. "Out of place" can no longer be detected by simply checking the rank order, but by ascertaining whether any sister constituent preceding the candidate constituent is prohibited from preceding it by the appropriate rule of language B, and similarly for constituents following the candidate constituent. An analogous procedure can be used to verify the equivalence constraint.

Typically, one sister constituent of the repeated translated constituents will receive two conflicting labelings from criterion (iii) in this situation. The noun in the Finnish adpositional phrase, the object in SVOV constructions, the proposition complement of both the English and Tamil verbs, all receive two labels this way. To weaken the model in order to accept such sentences, we must discard conflicting criterion (iii) labelings due to "out of place" configurations with respect to the repeated translated constituent. Criteria (i) and (ii) still operate. The equivalence constraint will not hold with respect to this sister constituent, but can be verified elsewhere.

2.2 Insertional code-switching.

In some bilingual communities, the code-mixing mode of discourse may include the possibility of inserting one specific type of constituent into positions where it would not occur monolingually. In the Tamil-English corpus in Section 2.1, examples (14) and (15), as well as (13), illustrate the placement of an English proposition *preceding* the Tamil propositional complementizer, instead of in its obligatory English position following *that.*

(14) Even there, I am really lucky | nu collanum.
that say must
'Even there, one must say that I am really lucky.'

(15) It corrodes your confidence |
nu, enakku oru feeling[8]
that I-dat. *a*
'I have a feeling that it corrodes your confidence.'

It is important to note that in this community this pattern is confined to the very particular category of propositional complements. All other code-switches satisfy the equivalence constraint; none of the numerous other word-order conflicts between Tamil and English give rise to an insertional code-switching possibility.

A different type of constituent insertion has been characterized quantitatively by Naït M'Barek & Sankoff (1988). This involves the insertion, by bilingual Moroccans, of a full French noun phrase, including determiners and quantifiers, in all contexts where an Arabic noun phrase would be appropriate. This includes, among other contexts, post-verbal subjects as in (16), which are not possible in French. Determiner-initial French noun phrases also appear after demonstratives as in (17) and (18), producing demonstrative-determiner sequences, and after the indefinite *waḥd* as in (18), producing determiner-determiner sequences, neither of which is a French pattern.

(16) γaw | les demandes
arrived the applications
'The applications arrived.'

(17) miši bḥal duk | les avions légers
This is not like these the airplanes light
'It's not like these light planes.'

(18) ça s'adresse surtout | l'waḥd | une
this is targeted mostly at one a
certaine classe | ḥant walla | le luxe |
certain class because has become luxury
bezzaf f' | les hôtels.
much in the hotels
'This is targeted mostly at a certain class because the hotels have become too luxurious.'

The sentences we have shown containing examples of constituent insertion are all excluded from the model based on the equivalence constraint. For example, the English propositional complement preceding the Tamil complementizer *nu* would be labeled for Tamil by criterion (iii) because of its non-English position, but this would conflict with the English label it would receive from criterion (ii) since it is a normal English sentence containing only English lexical items. Similarly, the post-verbal subject consisting entirely of a normal French noun phrase receives conflicting labels from its position corresponding to Arabic rules and from its own constituent elements.

It must be stressed that not all bilingual communities that develop code-mixing modes of discourse make use of constituent insertion; those that do (e.g. Tamil-English or Arabic-French), use it very sparingly in the sense that typically only one type of constituent (English propositions or French noun phrases) may be inserted in contexts (before *nu*, postverbally) where it would not be found in monolingual discourse.

It is not difficult in this case to relax the model conditions so that such sentences are permitted. As in Section 2.1, the specified category

[8] *feeling* is treated as a loanword, for reasons discussed in Section 3.

is simply allowed to escape labeling by criterion (iii), and can be labeled by its own subconstituents (criterion (ii)). There is no danger that this will result in anomalous labeling higher in the phrase structure, since a sister constituent (the Tamil complementizer, the Arabic verb) will already have the contrary label, and the constituent containing them is thus prevented from receiving a label by way of criterion (ii). Note, however, that only those insertions which are not in conflict with the fundamental production conditions *left-to-right* and *most nested* can be considered well-formed. In addition, the equivalence constraint, wherever the specified category is not one of the constituents directly involved, still holds.

3 The borrowing process.

The equivalence constraint formalized in Section 1.6.1 has been verified as a general tendency in several communities – Puerto Rican Spanish and English in New York (Poplack, 1980), Finnish and English (Poplack *et al.*, 1987b), Tamil and English (Sankoff *et al.*, 1990), Wolof and French, and Fongbe and French (Meechan & Poplack, 1995; Poplack & Meechan, 1995), Igbo and English (Eze, 1997), and others. However, there are actually relatively few data on which it has been independently tested, since most of the voluminous literature on code-switching, especially that on insertional switching, is based on data which represent, we would claim, lexical borrowing (e.g. Eliasson, 1991; Mahootian & Santorini, 1994; Backus, 1996). While code-switching essentially involves the reconciliation of the word orders of both languages, only the word-order of the recipient language is pertinent to borrowing. Thus attempts to understand code-switching based on a mixture of borrowing and true switching data are likely to be misleading.

In the model constructed above, the borrowing process is not relevant. Loanwords, including those are *ad hoc*, "nonce", or momentary, uses, are not excluded, but simply considered to be syntactically integrated, i.e. to behave as native lexical items with respect to word order. How can this working hypothesis be validated? In Sections 3.1 and 3.2, we demonstrate an answer to this problem.

3.1 Properties of loanwords.

Many loanwords have long histories in the recipient language, are used by monolinguals (often with no consciousness of their foreign etymology), are widespread and are accepted by monolingual dictionaries and other linguistic arbiters. None of these non-structural characteristics, however, are necessary to the borrowing process. In many communities, bilinguals have access to essentially the entire content-word lexicon of one language as potential loanwords into the other, perhaps for a single usage only. What is important is that when these words are borrowed the structural linguistic characteristics of their usage are the same as with established loanwords. What are these characteristics? Some of them are: integration into the recipient language at the syntactic, morphological, semantic and phonological levels, use as a single item independent of other donor language material, and restriction to nouns, verbs, adjectives, etc., to the exclusion of determiners, pronouns, prepositions and other grammatical words.

Often, during the study of a bilingual corpus, we discover a pattern of words from a specific lexical category in language A appearing in mixed discourse in contexts where they seem to violate the equivalence constraint, but when considered as language B words, i.e. borrowings, there is no violation, e.g. *kidney* in (9), *feeling* in (15). Thus these words seem *syntactically integrated* into language B. To confirm their borrowed status, we verify the other properties of loanwords.

Phonological integration turns out to be an unreliable indicator, for two reasons (cf Poplack *et al.*, 1987a). One is that bilinguals, in contrast to monolinguals, tend to be aware of the etymology of loanwords and, in some communities, will often reflect this knowledge in their pronunciation of borrowed items. Second, in some communities, the learning context results in phonologies for languages A and B which converge in unpredictable and diverse ways from speaker to speaker.

Semantic integration refers to a shift in function or meaning of a loanword from donor language characteristics to recipient language characteristics. This may often be documented for established loanwords that have had time to evolve within the host language, or for borrowings between languages whose functional categories are structured very differently, but in general, where bilingual borrowings are most frequently from and into the category of nouns, the criterion of semantic integration, though satisfied, may not always revealing.

The criterion of isolated occurrence of loanwords in recipient language contexts is more universal and is usually relatively easy to apply. The main difficulties come from compound words and other multi-word forms whose status as single lexical items is not always clear. Statistically, these should not constitute a major problem. There is also the possibility of coincidence. If nouns are often borrowed and adjectives are often borrowed, then occasionally a noun-adjective combination will appear to have been borrowed together, when this is just the result of chance. This will also be relatively rare.

The lexical/grammatical (or content/ function) contrast is also useful. It is true that among the world's languages, loanwords have on occasion included prepositions, pronouns, determiners, and other grammatical categories, but these are exceptional, and the overwhelming tendency is for borrowing, and especially one-time borrowing by bilinguals, to affect nouns, and to a lesser extent, verbs, adjectives and adverbs. Moreover, in specific communities, bilingual borrowing may be focused on particular categories more than in other communities, and these patterns may be useful for analytical purposes.

Finally, it is the criterion of morphological integration which is of great interest. Loanwords, established or momentary, are inflected exclusively through recipient language morphological rules. Insofar as such marking is non-null and is different for language A and language B, words borrowed from the former into the latter should display exclusively language B inflectional morphology.

3.2 Case marking of English-origin nouns in Tamil

In the Tamil corpus referred to in Sections 2.1 and 2.2, many English-origin nouns occur in preverbal position, where the verb is an inflected Tamil form. Tamil being a SOV language, this is just where Tamil direct (and indirect) objects appear. Examining these English-origin nouns, we first note that these occur most frequently in isolation, and occasionally as compounds, or as familiar adjective-noun combinations, but never preceded by English prepositions, articles, quantifiers or demonstratives as would frequently be the case if these were parts of well-formed English fragments resulting from code-switching.

	marker	marker	
	present	absent	N
ACCUSATIVE			
English origin	29%	71%	108
Native Tamil (no pronouns)	39%	61%	51
DATIVE			
English origin	86%	14%	91
Native Tamil	99%	1%	230

Table 1: Variable accusative and dative marking on English-origin and native Tamil objects.

Second, whereas the preponderance of preverbal native Tamil objects are actually pronouns, from 45-70% depending on the case, no English pronouns whatsoever appear in this context, as would be expected from borrowings, but not if these were code-switches into English fragments — which would normally include at least the occasional pronoun.

Third, it is the inflectional morphology on these nouns which is the most revealing. They either have null morphology or Tamil inflections. Since in Tamil the numerically frequent accusatives and datives are prescribed to take non-null case-marking, we examine marking rates quantitatively. In fact, as in Table 1, many (non-pronominal) Tamil forms are unmarked, especially accusatives. The English-origin forms show remarkably parallel rates, especially when the accusative-dative contrast is considered. This morphological integration into Tamil is exactly what would be expected of borrowings, and certainly not of well-formed English fragments produced by code-switching.

In summary, the criteria of syntactic integration, isolation, lexical category, and morphological integration all confirm the loanword status of the preverbal English-origin nouns, and justify our considering them as Tamil nouns for the purposes of applying our model of code-switching.

3.3 Formal considerations

Formally, the only adaptation of our model necessary to allow for borrowing is, for specified terminal categories in T, the list of words in language A in the category is added in to the pre-existing list for language B, so that there are now two possible translations in B for each word in A in this category. The uniqueness statement in Theorem 1 has to be modified to take this into account, but this leads to no difficulty. We presume of course that the borrowed word in B can be distinguished from its "etymological" origin in A by the tests and criteria illustrated in Section 3.2.

Discussion.

The core of this work is our model of equivalence-point code-switching. This avoids issues of grammatical theory by focusing on the "real-time" production of a code-mixed sentence drawing on the output of two monolingual grammars.

Our model is built on an earlier formulation of the equivalence constraint (Sankoff & Mainville, 1986). It is the production aspect here, however, that allows us to achieve the all-important well-formedness of monolingual fragments, not strictly guaranteed in the earlier work, and to model the essential unpredictability of code-switching. The present version (Sankoff, 1998) has a more economical protocol for constituent labeling and a more complete account of the coincidence (or lack thereof) between word-level switching and constituent-level switching. We have shown here how to weaken the strong conditions leading to Theorems 5 and 6 to account for other types of code-switching that have been reported.

We have allowed some degree of asymmetry in the model. Borrowing can be unidirectional with respect to specific categories. The same is true for constituent insertion. Further development will require weakening the one-to-one correspondence of the sets of rules, and of the grammatical and lexical categories of the two languages. For example, the use of specialized incorporation devices, like inflection-carrying dummy verbs, or techniques for marking borrowed adverbs, typically belong to one language and not the other.

We have not treated the topic of interference in this presentation. Interference differs from borrowing in several respects, in particular on the level of intentionality — interference is more likely to be avoided or corrected by speakers, and is more likely to show up in communities where the domain of monolingualism, and frequency of use of the affected language are restricted. Nevertheless, interference has interesting consequences for our model in that it tends to affect pre-sentential discourse markers, tags, conjunctions, prepositions and other grammatical morphemes rather than lexical items as is the case for loanwords. Both borrowing and interference can lead to the long-term establishment of lexical items, so that when interference is frequent, the distinction between the two becomes of interest.

Acknowledgements

This work supported in part by grants from the Natural Sciences and Engineering Research Council and the Social Sciences and Humanities Research Council. The author is a Fellow of the Canadian Institute for Advanced Research.

References

Backus, A. (1996). Two in One. Bilingual Speech of Turkish Immigrants in the Netherlands. Tilburg: Tilburg University Press.

Belazi, H. M., Rubin, E. J., and Toribio, A. J. (1994). Code switching and X-bar theory: The functional head constraint. Linguistic Inquiry, 25, 221-237.

Bentahila, A., and Davies, E. E. (1983). The syntax of Arab-French code-switching. Lingua, 59, 301-30.

Di Sciullo, A.-M., Muysken, P., and Singh, R. (1986). Government and code-mixing. Journal of Linguistics, 22, 1-24.

Eliasson, S. (1991). Models and constraints in code-switching theory. Papers from the Workshop on Constraints, Conditions and Models, pp. 17-50. Strasbourg: European Science Foundation.

Eze, E. (1997) Aspects of language contact: a variationist perspective on code-switching and borrowing in Igbo-English bilingual discourse. Ph.D. dissertation, University of Ottawa.

Gumperz, J., and Hernandez, E. (1969). Cognitive aspects of bilingual communication. Working Paper Number 28, Language Behavior Research Laboratory. University of California, Berkeley.

Joshi, A. K. (1985). Processing of sentences with intrasentential code-switching. In D. R. Dowty, L. Karttunen and A.M. Zwicky (eds.), Natural Language Parsing, pp. 190-205. Cambridge: Cambridge University Press.

Labov, W. (1969). Contraction, deletion and inherent variability of the English copula. Language , 45, 715-762.

Lipski, J. (1977). Code-switching and the problem of bilingual competence. In M. Paradis (ed.), Aspects of Bilingualism, pp. 250-263. Columbia, SC: Hornbeam Press.

Mahootian, S., and Santorini, B. (1994). Adnominal adjectives, codeswitching and lexicalized TAG. In A. Abeille, S. Aslanides and O. Rambow (eds.), 3e colloque international sur les grammaires d'arbres adjoints (TAG+3), Technical Report TALANA-RT-94-01, pp. 73-76.

Meechan, M., and Poplack, S. (1995). Orphan categories in bilingual discourse: Adjectivization strategies in Wolof-French and Fongbe-French. Language Variation and Change , 7, 169-194.

Muysken, P. (1995) Code-switching and grammatical theory. In L. Milroy and P. Muysken (eds.), One speaker, two languages, pp. 177-198. Cambridge: Cambridge University Press.

Myers-Scotton, C. (1993). Dueling Languages. Oxford: Clarendon Press.

Naït M'Barek, M., and Sankoff, D. (1988). Le discours mixte arabe/français: des emprunts ou des alternances de langue? Revue Canadienne de Linguistique, 33(2), 143-154.

Nishimura, M. (1986). Intrasentential codeswitching. The case of language assignment. In J. Vaid (ed.), Language Processing in Bilinguals: psycholinguistic and neuropsychological perspectives, pp. 123-43. Hillsdale, NJ: Lawrence Erlbaum.

Pandit, I. (1990). Grammaticality in codeswitching. In R. Jacobson (ed.), Codeswitching as a worldwide phenomenon, pp. 33-69. New York: Peter Lang.

Pfaff, C. (1979). Constraints on language mixing: intrasentential code-switching and borrowing in Spanish/English. Language, 55, 291-318.

Poplack, S. (1978). Syntactic structure and social function of code-switching. In R. Duran, (ed.), Latino Discourse and Communicative Behavior, pp. 169-184. New Jersey: Ablex,

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching. Linguistics, 18, 581-618.

Poplack, S., and Meechan, M. (1995). Patterns of language mixture: Nominal structure in Wolof-French and Fongbe-French bilingual discourse. In L. Milroy and P. Muysken (eds.), One speaker, two languages, pp. 199-232. Cambridge: Cambridge University Press.

Poplack, S., Sankoff, D., and Miller, C. (1987a). The social correlates and linguistic processes of lexical borrowing and assimilation. Linguistics, 26(1), 47-104.

Poplack, S., Wheeler, S., and Westwood, A. (1987b). Distinguishing language contact phenomena: Evidence from Finnish-English bilingualism. In P. Lilius and M. Saari (eds.), The Nordic languages and modern linguistics 6, pp. 33-56. University of Helsinki Press.

Rivas, A. (1981). On the application of transformations to bilingual sentences. Manuscript. Amherst, MA: Dept. of Spanish and Portuguese, University of Massachusetts.

Sankoff, D. (1988) Sociolinguistics and syntactic variation. In F. Newmeyer (ed.), Linguistics: the Cambridge Survey. IV Language: the socio-cultural context, pp. 140-161. Cambridge: Cambridge University Press.

Sankoff, D. (1998) A formal production-based explanation of the facts of code-switching. Bilingualism, Language and Cognition 1 (in press).

Sankoff, D., and Mainville, M. (1986). Codeswitching of context-free grammars. Theoretical Linguistics, 13, 75-90.

Sankoff, D., Poplack, S., and Vanniarajan, S. (1990). The case of the nonce loan in Tamil. Language Variation and Change, 2(1), 71-101.

Woolford, E. (1983). Bilingual code-switching and syntactic theory. Linguistic Inquiry, 14, 52-536.