Eric Schmidt

Prof. Suzanne Flynn

Linguistic Study of Bilingualism

December 13, 2013

## A Minimalist Approach to Code-Switching

In the field of linguistics, the topic of bilingualism is a broad one. There are many interesting phenomena to be studied, but one of particular interest in this paper is the interaction between languages during language production. One manifestation of this is code-switching, which is the "alternate use of two or more languages among bilingual interlocutors" (MacSwan, 2010). Code-switching is very fascinating because although some may believe it is simply the use of words from different languages in the same utterance, previous research on the subject has shown that it is a consistent, rule-governed phenomenon (cf. Poplack, 1980). Naturally, the next steps are to begin uncovering these rules and understanding exactly how mixed sentences are formed. There have been a number of investigations devoted to this, ranging in style from empirical study to theoretical proposals. This paper will take a more theoretical approach, using code-switching data to propose a model of language production that can account for the observed patterns of code-switching.

Before continuing, it is imperative to understand the nature of the different approaches that can be taken to this topic. Most systems which attempt to account for code-switching follow two main paradigms, which are the constraint-based and constraint-free research programs. These two approaches are fundamentally different. Constraint-based systems are generally very descriptive in nature, and posit that the observed constraints on code-switching are embedded as rules in the grammar. Thus, constraint-based models are often comprised of many rules specific

to code-switching that describe the environments in which switching may and may not occur. In contrast, constraint-free models formulate no such rules about code-switching in the grammar. The goal of the constraint-free approach is to place as few explicit constraints as possible on language production, instead opting to use the preexisting features of grammar to account for code-switching. The main principle, as MacSwan notes, is that "nothing constrains [code-switching] apart from the requirements of the mixed grammars" (2010). This principle not only encourages an elegant explanation of code-switching, but also prompts the formulation of more robust models of grammar.

While a number of different ideas have been proposed under each approach, the most popular systems have followed the constraint-based paradigm. To further motivate the present work, some of the previous constraint-based options will be evaluated. In one of the first syntactic approaches to code-switching, Di Sciullo *et al*. proposed the Government Constraint, which essentially states that "when a government relation holds between elements, there can be no mixing" (1986). An illustration of the government relation is shown in Figure 1, where in (1a) we see that in the English sentence "they have given the lectures," the verb *have* governs the verb phrase [*given the lectures*], and in the French-Italian mixed production given in (1b) we see that the verb *hanno* governs the verb phrase [*donné des cours*].

Figure 1. The government relation

(a)

1a. V 'have' governs VP 'given the lectures'

1b. V 'hanno' governs VP 'donné des cours'
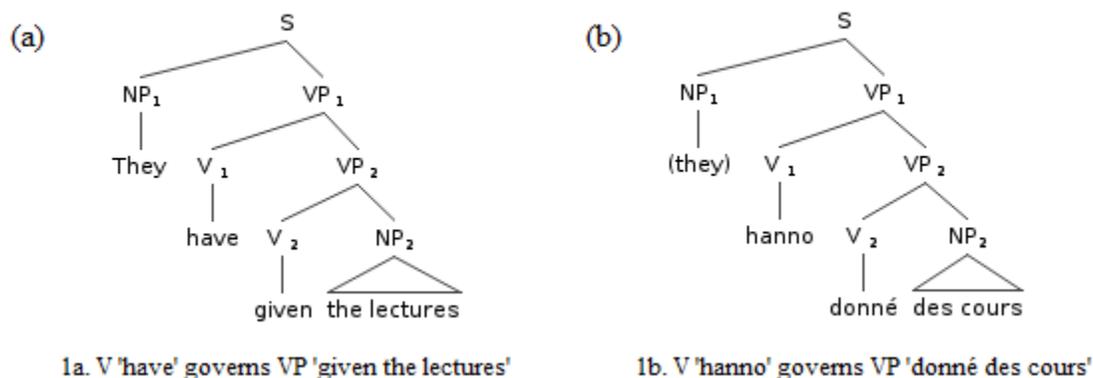
The Government Constraint, despite its initial appeal as an attempt to capture the patterns of code-switching in one constraint, unfortunately fails in the case of (1b) above, which is reportedly a grammatical construction (Di Sciullo *et al.*, 1986). The verb *hanno* is from Italian, while the governed verb phrase [*donné des cours*] is French. This is a clear violation of the proposed constraint, which Di Sciullo *et al*. attempt to explain by stating that modal and auxiliary verbs do not govern their complement verb phrases. This leads to questions about the definition, and in fact the entire notion, of syntactic government. Basing language production rules on seemingly arbitrary relations in a syntax tree does not seem intuitive, and so this approach is not ideal.

The next approach to be popularized was the Matrix Language Frame Model, introduced by Myers-Scotton (1993). This model maintains that all sentences exhibiting code-switching have a matrix language and an embedded language, and their production is based on two basic principles – the Morpheme Order Principle, which states that the word order of the sentence is defined by the matrix language, and the System Morpheme Principle, which states that all "system morphemes" must come from the matrix language. A "system morpheme" is defined as a morpheme which is able to restrict its possible referents; thus determiners are a canonical example (consider *cats* vs. *the cats* vs. *the cat*). The model also allows for "embedded language islands," which are sub-sentential phrases entirely in the embedded language. These are in effect exceptions to the two principles, but we need not even appeal to this fact to show why the Matrix Language Frame Model is inadequate. Consider the following constructions:

(2a) Los teachers

(2b) *The casa

(English/Spanish; MacSwan, 2010)

Interestingly, (2a) is generally judged as grammatical by English/Spanish bilinguals in most contexts while (2b) is not. If Spanish is chosen to be the matrix language, then (2) is accounted for because (2b) violates the System Morpheme Principle; however, because this finding holds in both English and Spanish contexts, English could just as well be the matrix language, which would render (2a) a violation. The vagueness of the definition of the matrix language leads to this issue, and is a main concern with the model. Overall, it is clear that the application of the principles of the Matrix Language Frame Model is somewhat subjective, due to the loose nature of the model's formulation. This is an unideal characteristic of a grammatical model, and further motivates a different approach.

One point of note is that both of the discussed systems, and many others, necessitate the notion of different "languages." The Government Constraint does this by stating that sentence elements in a government relation must have the same "language index," which is a marker attached to words "drawn from a particular lexicon" (Di Sciullo *et al.*, 1986). The Matrix Language Frame Model does the same by explicitly referring to "languages." The issue with this is that the existence of discrete languages cannot be taken for granted; much current work acknowledges the existence of dialect continua, and thus different "languages" can be seen as different manifestations of a common underlying system (cf. Chambers & Trudgill, 1998). Further, a distinction between different languages would support the idea that multiple separate lexicons are represented in the brain. However, there is no empirical evidence for this, and intuition may guide us in the opposite direction – the mere existence of code-switching is enough to illustrate that lexical items from supposedly different lexicons can come together, so the idea that there is only one master lexicon is in fact quite sensible. This is all the motivation necessary to undertake the formulation of a new model of language production.

The basis for this model will be the grammatical framework set forth by Chomsky in his Minimalist Program (1995). One of the goals of the program is to account for as much of language production as possible using the fewest number of explicit rules. This is thus a very constraint-free approach to the matter, and serves as a good starting point from which a grammatical model can be developed. Before building a new model, however, it is helpful to understand some of the groundwork that has been done within the program. General systems that follow the minimalist approach hold that the language faculty is comprised of two main elements – the computational system for human language and the lexicon. There are then a number of functions in the computational system that can act to build phrases, perhaps the most important of which is MERGE. This function combines elements recursively in order to generate tree-like sentence structures; an example is shown in Figure 3. An important characteristic of this function is that the combined result is marked by the category of one of the elements – in the example, the adjective *lazy* combined with the noun *dog* results in the *noun* phrase [*lazy dog*]. This apparent domination of one element by the other is due to the nature of the lexical items themselves. In particular, all information necessary for the use of a word is contained in its lexical entry.

This leads to the basic idea upon which our model is formulated: language production is a result of operations selectively applied to a collection of data, and nothing else. There are no *constraints* on what can and cannot be produced, only functions which choose and combine lexical data in order to produce phrases. If the lexicon is taken to be a collection of objects, each storing information about phonology, semantics, and morphosyntactic properties, then the basic

Figure 3. MERGE
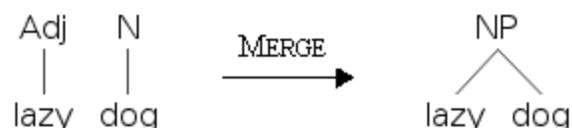
idea states that language production is a consequence of operations on these lexical objects. As such, a model which describes this situation is necessary.

The first point to address is the lexicon. In this model, there is only one lexicon; all items said to come from different "languages" are contained within this one lexicon. The lexical items themselves, being the basic unit of language production, are then objects containing all information relevant to their use. A simplified breakdown of a lexical element is shown in Figure 4, using the Spanish *casa* 'house'. The main properties of a lexical object are *category*, *dominates*, *features*, *phonology*, and *semantics*. The *category* is quite simply the same as the already established notion of lexical category, also termed 'word class' or 'part of speech'. The *dominates* property serves to describe the types of objects that the current object may dominate when acted upon by MERGE. In the example, the noun *casa* can dominate adjectives, prepositional phrases, and complementizer phrases; combining it with another type of object will result in an ungrammatical construction. The *dominates* field may also hold any subcategorization information, so lexical objects may select for specific properties of the objects they dominate. The *features* of a lexical object are the grammatical features already understood to be part of lexical items, such as person, number, gender, case, etc; as Figure 4 illustrates, *casa*

Figure 4. Sample breakdown of a lexical object

```
casa : {
        category : N ,
        dominates : {Adj, PP, CP} ,
        features : { number : singular ,
                     gender : feminine } ,
        phonology : /kasa/ ,
        semantics : 'house'
    }
```

is singular in number and feminine in gender. The *phonology* property of a lexical object simply contains information about its pronunciation, and may additionally be posited to contain any phonological rules associated with the item (such as vowel harmony or devoicing in certain situations). Lastly, the *semantics* field stores information about the meaning of the object (which, in the example, has been rendered as 'house'; this of course depends on the true representation of meaning in the brain but is not the current focus). As phonology and semantics are not pertinent to this work, we will assume that the *phonology* and *semantics* fields store all information necessary to physically produce and understand the lexical objects.

Now that the nature of our data has been made clear, the available operations must be detailed. This model uses two primary functions to produce phrases, SELECT and MERGE. More functions may exist which can manipulate the produced syntax trees, but for an analysis of the grammaticality of code-switching the two presented operations are sufficient. The first is SELECT, whose function is to choose objects from the lexicon. Use of this operation is decided by the speaker, who selects objects based on their properties and the meaning that is to be conveyed. The chosen objects are placed in an array which shall be called the *selection*; from this point, further functions used in the production of a phrase may only operate on objects in the selection. To construct phrases from objects, MERGE must be applied. MERGE, in this formulation, is a binary operation which, based on the properties of its arguments, creates a combined object with a derived set of properties; an example is shown in Figure 5. In particular, one of the objects

Figure 5. MERGE under the object operation model

(a) *la* : {
    category : Det ,
    dominates : {N, NP} ,
    features : { number : singular ,
           gender : feminine } ,
    phonology : /la/ ,
    semantics : 'the'
  }

(b) MERGE(la, $casa$) = *la casa* : {
    category : DetP ,
    dominates : $\varnothing$,
    features : { number : singular ,
           gender : feminine } ,
    phonology : /la kasa/ ,
    semantics : 'the house'
  }

must be able to dominate the other, and all of the features of the dominated object must be matched by features of the dominating object. We assume that lexical objects contain information about the order in which they combine with the elements they dominate, and MERGE places them in that order. The resulting object has the category *x phrase*, where *x* is the category of the dominating object. If any features of the dominated object are unmatched, the resultant construction is ungrammatical.

Let us now apply this model to some of the reported English/Spanish code-switching data to see how it accounts for the observed patterns. Returning to the example of (2), reproduced as (6) below, we find a relatively simple explanation.

(6a) Los teachers

(6b) *The casa

(English/Spanish; MacSwan, 2010)

These two constructions are both determiner phrases, and thus the determiners are the dominating objects in each pair. The properties of each lexical object must be determined in order to apply the new model, and the feature sets are given below in (7).

(7a) *los*: {number: plural, gender: feminine}

(7b) *the*: {number: singular} [*]

(7c) *teachers*: {number: plural}

(7d) *casa*: {number: singular, gender: feminine}

Under the newly proposed model, [*los teachers*] is grammatical because all features of *teachers*, the dominated object, are matched by *los*. However, in (6b), the dominated *casa* contains an unmatched gender feature, and thus the construction is ungrammatical, which is consistent with the grammatical judgment of (6).

---

[*] The lexical object *the* can be thought of as having two variants; this is the singular variant.

For another example, consider the following sentences:

(8a) *El no wants to go

(8b) *He doesn't quiere ir

(English/Spanish; MacSwan, 2004)

Both are judged as ungrammatical by English/Spanish bilinguals; let us now understand why. In (8a), the feature set of the verb *wants* is {person: third, number: singular}, while the negator *no* has no features. Although we see that the features of the dominated element *no* in the verb phrase are matched, we may posit under our model that *wants* has the property that it may dominate a negator on the left. Thus when the Spanish negator *no* is combined on the left, the resultant construction is ungrammatical. In (8b), the English verb *does* (and thus, under our model, the derived phrase [*does not*]) can be thought of as subcategorizing for a type of infinitive form[*] when it dominates another verb. Thus, the combination [*doesn't quiere*] is ungrammatical because the Spanish verb *quiere* is conjugated and not infinitive.

There are of course instances of code-switching that are difficult to account for in the current state of the model. One particularly interesting example is:

(9a) Mi hermano bought some ice cream.

(9b) *El bought some ice cream.

(English/Spanish; MacSwan, 2010)

The construction in (9a) is found to be grammatical while (9b) is not; our model can only account for (9a). If we assume that verbs are dominated by their subjects, then the feature set {person: third, number: singular} of the English *bought* is matched by the feature set {person: third, number: singular, gender: masculine} of the Spanish [*mi hermano*], and the sentence is

---

[*] "A type of infinitive form" rather than just "an infinitive" because the construction *doesn't querer* is also quite deviant. Additionally, the English infinitive *to want* has a slightly different notion than the plain *want*; this is a topic for a different investigation.

grammatical. We may then suppose that the Spanish pronoun *el* has the same feature set as the phrase [*mi hermano*], but even though the features appear to match, (9b) is ungrammatical. This leads us to believe that perhaps there are more features at work; perhaps *bought* and [*mi hermano*] both have a feature that *el* does not. Further work under the current approach may elucidate such information.

Clearly, this model, and in fact the entire constraint-free approach to code-switching, is still in its infancy. Much more investigation is needed to build a stronger grammatical model, but hopefully it has at least been made evident that a constraint-free approach to the topic is generally more robust than a constraint-based approach. Along with this, due to the inadequacies of previously proposed models, an approach based on the principles of the Minimalist Program would be the ideal way to account for the observed patterns of code-switching. The model of lexical object operations, though not completely formulated, is a promising framework founded on these principles. Hopefully, it can be regarded as a step in the right direction towards understanding the interactions between languages during language production, and ultimately, the underlying function of the human language faculty.

References

Chambers, J. K. & Trudgill, P. (1998). *Dialectology* (2nd ed.). Cambridge: Cambridge

University Press.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Di Sciullo, A.-M., Muysken, P., & Singh, R. (1986). Government and code-switching. *Journal of

Linguistics*, 22, 1-24.

MacSwan, J. (2004). Code switching and grammatical theory. In T. K. Bhatia & W. C. Ritchie

(Eds.), *The Handbook of Bilingualism* (283-311). Oxford: Blackwell Publishing.

MacSwan, J. (2010). Plenary address: Unconstraining codeswitching theories. *Proceedings from

the Annual Meeting of the Chicago Linguistic Society 44*. Chicago: University of Chicago

Press.

Myers-Scotton, C. (1993). *Dueling languages: Grammatical structure in code switching*.

Oxford: Clarendon Press.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish *y termino en español*. *Linguistics*,

18, 581-618.