A universal model of code-switching and bilingual language processing and production

Carol Myers-Scotton
and
Janice Jake

19.1 Introduction

The Matrix Language Frame (MLF) model (Myers-Scotton 1993a, 1997), augmented by the 4-M model of morpheme classification (Myers-Scotton and Jake 2000; Myers-Scotton 2002a), provides a major linguistic theory of language contact dedicated to bilingual processing and production. This model has inspired many studies of bilingual speech within diverse language pairings and accounts for a variety of bilingual behaviors, principally code-switching (hereafter CS). Unlike most other approaches to CS, the MLF model enjoys widespread appeal among linguists and psycholinguists alike.

19.1.1 No chaos allowed: the Uniform Structure Principle

This chapter elaborates and illustrates the research program framed by the MLF model. What is new is that it emphasizes how a principle of uniform structure drives the explanation of what does and does not occur in CS. The first goal of this chapter is to show how the Uniform Structure Principle (USP) implies a particular view of processing and production in bilingual speech, especially in CS. Bilingual speech is defined as surface level morphemes from two or more language varieties in the same clause. With the 4-M model, the USP clarifies and strengthens the Matrix Language Frame (MLF) model, as a model of CS. A succinct way of viewing the USP is the phrase, ”no chaos allowed.” This may be obvious for monolingual speech; that it applies to bilingual speech is not so obvious. A priori, the ways in which languages participate in bilingual speech are unconstrained. In particular, the source of grammatical structure within a bilingual clause could be shared in any number of ways. But this does not happen.
For bilingual speech, "no chaos allowed" means a particular asymmetry between the participating languages. This is formalized in the USP, as follows:

A given constituent type in any language has a uniform abstract structure and the requirements of well-formedness for this constituent type must be observed whenever the constituent appears. In bilingual speech, the structures of the Matrix Language (ML) are always preferred. Embedded Language (EL) islands (phrases from other varieties participating in the clause) are allowed if they meet EL well-formedness conditions, as well as those ML conditions applying to the clause as a whole (e.g. phrase placement).

(cf. Myers-Scotton 2002a)

The second goal here is to make more explicit how specific morphemes are classified under the 4-M model and to show how differences in morpheme type explain their distribution in CS. These distributions will be shown to reflect the USP. In doing this, the chapter focuses on CS in general, but gives special attention to prepositions, complementizers, and pronouns.

First, the MLF model of CS and its relation to the USP is summarized. Next, the view of language production motivated by empirical CS data and the MLF model is outlined, as is the 4-M model and how it relates to the USP and the MLF model. Finally, the descriptive sections of the chapter are shown to support the theoretical goal of explaining the asymmetries that pattern CS data. With the USP as an overarching framework in which the MLF and 4-M models add specific hypotheses, a set of principled predictions emerges about what does and does not occur in CS. These predictions should have relevance to other types of contact phenomena as well.

19.2 Summary of the MLF model

The key feature of the MLF model is that it differentiates both the participating languages and morpheme types at a number of abstract levels. It emphasizes asymmetry, claiming crucially a dominant role in the bilingual clause for only one of the participating languages, the ML. That is, reflecting the USP, the MLF model limits the EL's main role to providing either content morphemes in mixed constituents or EL phrase-level constituents (EL islands), or both. Asymmetry under the model also differentiates content and system morphemes and their participation in CS. The model assumes that these two asymmetries apply universally in Classic CS and empirical evidence largely supports this. Classic CS is defined here as CS in which empirical evidence shows that abstract grammatical structure within a clause comes from only one of the participating languages. Which of the participating languages is the ML is determined.
for each corpus. Within a corpus, the ML may vary from clause to clause, although this is unusual.

The MLF model has always defined content and system morphemes differently from those classifications based on the lexical vs. functional distinction, or the open vs. closed class distinction. Specifically, system morphemes are not the same as the functional elements or closed-class items in other linguistic models. They are defined in opposition to content morphemes. Content morphemes are defined as assigning or receiving thematic roles; system morphemes do not. Prototypical system morphemes are affixes and some function words that are free forms but do not occur alone, such as determiners and clitics.

The MLF model contains two principles that can be interpreted as hypotheses about the differing roles of the participating languages. These were first presented in Myers-Scotton (1993a, 1997:82) as the Morpheme Order Principle (MOP) and the System Morpheme Principle (SMP). They specify the elements in a bilingual constituent that must come from only one participating language; in effect, support of these principles identifies this language as the ML.

19.2.1 Exemplifying the MLF model

Example (1) comes from a corpus of Turkish-Dutch CS, where elements from the EL appear in italics. Turkish is verb-final, and in this example the inflected (main) verb yap “do” occurs after its predicate (the Dutch infinitive), not before it as it would in Dutch. This configuration supports the MOP, which states that only one of the participating languages supplies morpheme order in such constituents. Note as well that all instances of subject-verb agreement come from Turkish. This supports the SMP, which states that only one of the participating languages supplies a certain type of system morpheme (SM), now called an outsider late SM under the 4-M model. Subject-verb agreement is such a morpheme. Based on the example’s support of the MOP and the SMP, Turkish is identified as the ML.

(1) Turkish–Dutch

O diyor ben utmak-en yap-ti-m diyordu kiz-nam
he.say.PROG.3SG 1SG finish-INF do-PRET-1SG say.IMP.3SG girl-with
“He says ‘I broke up with a girl.’”

(Backus 1992:107)

Example (2) supports both principles as well. Note the order of certificate and its modifiers; they follow Swahili word order, not that of English. In addition, although the main verb is from English (depend), subject-agreement, an outsider morpheme, (i.e., class 9) comes from Swahili, agreeing with a subject mentioned before (saa hio is an introductory phrase, not the subject). These data support Swahili as the ML in this example.
19.2.2 Three premises summarizing the MLF model

Three basic premises have always structured the MLF model (Myers-Scotton 1993a, 1997):

1. Participating languages do not play equal roles in the bilingual clause.
2. In bilingual constituents within this clause, not all morpheme types can come equally from the ML and EL.
3. The SMP limits the occurrence of system morphemes that build clausal structure of the ML.

19.2.3 Relating the MLF model to other data

The implicit domain of the MLF model always has been participating varieties that are not mutually intelligible. It may well apply to other varieties, but that would be an unintended bonus. As already noted, the model applies only to what is defined above as Classic CS. This type of CS contrasts with Composite CS in which the abstract grammatical structure underlying surface configurations still comes largely from one language, but also partially from another. The Abstract Level model (see Myers-Scotton and Jake 2001; Myers-Scotton 2002a) is especially relevant to Composite CS. More research may show Composite CS to be more common than Classic CS. The USP applies to both types of CS, of course, to the extent there is an ML, and so does the 4-M model; both are universal.

19.3 The language production model

A general language production model of four levels, the conceptual level, the mental lexicon, the formulator, and the surface level (see Levelt 1989) accommodates CS and other contact phenomena. The conceptual level is pre-linguistic, and includes speaker intentions, as well as other cognitive components, such as memory. The critical factor in resolving competition at the conceptual level is which lemma entry (either from the ML or the EL) best conveys the speaker's semantic and pragmatic intentions (La Heij 2005). Intentions activate semantic and pragmatic features that are bundled together, pointing to language-specific lemmas in the mental lexicon.
If CS becomes part of the cognitive plan, the ML is selected at the conceptual level. Speakers must be able to produce well-formed utterances for the language selected as the ML because it provides the grammatical frame of the bilingual clause; they may be less proficient in the EL, but are not necessarily so. Often the ML is the speakers’ L1, although not necessarily. Selecting a language as the ML is largely unconscious, although the process draws on various resources, especially the participant’s cognitive system (i.e. memory about social aspects of contexts compared with the nature of current contexts).

But satisfying semantic and pragmatic intentions at the conceptual level is not the only issue. CS data imply a matching process – checking – between the abstract requirements regarding the structural well-formedness of the ML and a potential EL element in a bilingual clause (Myers-Scotton and Jake 1995). This is referred to as congruence checking. There must be some degree of semantic match, but more critical is a grammatical match.

Lemmas in the mental lexicon include directions that map semantic information to grammatical structure, directions needed at the next level, the formulator. Thus, lemmas contain information beyond word meaning about thematic roles and selectional restrictions that have syntactic consequences, such as argument structure. For example, the verb hit assigns the thematic roles of Agent and Patient to a subject and object, respectively. Other lemmas in the mental lexicon underlie late SMs that become salient at the level of the formulator and will build syntactic structure (see §19.4.1).

Lemmas point to language-specific morpho-syntactic constraints located in the formulator. The formulator assembles larger constituents. The mental lexicon also contains language-specific Generalized Lexical Knowledge (GLK) that reflects the grammatical competence of speakers in their languages (Myers-Scotton and Jake 1995; Myers-Scotton 2002a). GLK plays an important role in congruence checking between languages and explains how EL lemmas without close ML counterparts occur in CS because their features can be checked against ML Generalized Lexical Knowledge.

Incomplete congruence can have repercussions for CS. Significant incongruence may mean that optimal CS mixed constituents with EL elements entirely framed by the ML do not occur. Instead, compromise strategies such as entire, well-formed phrases in the EL (EL islands), may occur. Sometimes EL content morphemes occur in ML frames as bare forms, without the SMs that would make the phrase well-formed in the ML. Note that the occurrence of bare forms implies that the ML is an abstract construct and not necessarily identical with the morpho-syntax of the language that is its source.

19.4 The 4-M model

The 4-M model does not replace the MLF model; rather, it offers a more precise description of morpheme types by viewing them in terms of their
syntactic roles and how they are activated in language production. For convenience, the model employs the term “morpheme” to refer both to the abstract entries in the mental lexicon that underlie surface realizations and to the surface realizations themselves. The model separates out three types of system morpheme: early SMs, and two types of late SMs, bridges and outsiders. The MLF model’s SMP is often misunderstood as applying to all SMs. However, it was always intended to constrain only one type of SM, now called outsiders (Myers-Scotton 1993a, 1997:82). The 4-M model keeps the division between content and system morphemes, but explicitly recognizes significant divisions between morpheme types.

The primary division is between morphemes that are conceptually-activated (e.g. nouns and verbs) and those that are structurally-assigned (e.g. AGR elements). Content morphemes are conceptually-activated. They are based on the speaker’s pre-linguistic intentions; recall the semantic/pragmatic feature bundles that speakers’ intentions activate. But early SMs (e.g. plural affixes) are also conceptually-activated; they flesh out the meaning of their content morpheme heads that “indirectly elect” them (see Bock and Levelt 1994). Because they are structurally-assigned, late SMs contrast with both content morphemes and early SMs in an important way with many ramifications for both monolingual and language contact data.

19.4.1 The Differential Access Hypothesis

In CS, and in line with what the USP would predict, the distribution of morpheme types across the ML and EL is quite different. Not only are there distribution differences between content and system morphemes, but also within the category of SMs itself. Recognizing this motivates new ways of classifying morphemes and leads to the 4-M model. In turn, how the 4-M model classifies morpheme types leads to a hypothesis that abstract differences at the production level account for surface level differences in morpheme types. The Differential Access Hypothesis (DAH) offers an explanation for the observed differences. The DAH is the following:

The different types of morpheme under the 4-M model are differentially accessed in the abstract levels of the production process. Specifically, content morphemes and early SMs are accessed at the level of the mental lexicon, but late SMs do not become salient until the level of the formulator.

(cf. Myers-Scotton 2002a:78, 2005a)

The hypothesis suggests the following scenario. As already noted, lemmas underlying content and early SMs send language-specific directions to the formulator to build larger linguistic units. To build these units, these directions contain information about assigning late SMs. These late SMs become salient only when they are structurally-assigned at the formulator. Separating the activation of abstract elements underlying surface
morpheme types echoes Garrett’s view that “major and minor grammatical category words behave quite differently” (1993:81). However, he and others, such as Ullman (2001), who posits that the grammar and lexicon are two separate systems, do not differentiate the distribution of different types of SMs.

This theory differs from contemporary linguistic theories that project “functional” elements as the heads of maximal projections. The following sections exemplify EL morphemes in CS in terms of SM types, showing how their distribution follows the USP and implies the DAH.

19.4.2 Early SMs

Early SMs are so-designated because they, along with their content morpheme heads, become salient in the mental lexicon as the basic building blocks of constituent structure, such as NP, VP, AP. Yet, they are still SMs because only content morphemes receive and assign thematic roles. Early SMs typically occur with the content morpheme heads that select them. Early SMs may be free or bound. For example, definite articles are early SMs but always occur with nouns in English.

Plural and derivational affixes are examples of early SMs. Unfortunately, to date, few studies include quantitative evidence on the distribution of either type of early SM. However, in one quantitative study considering determiners in bilingual NPs in a Spanish–English corpus, 151/161 (94%) of English nouns in well-formed mixed NPs occur with Spanish determiners, such as el garage (Jake et al. 2002). Because Spanish can be identified as the ML, the overwhelming number of these mixed NPs supports the USP because ML structure is maintained in these NPs even though the noun is from English. However, as early SMs, definite articles can come from the EL without violating the SMP, and occasionally do, as in Palestinian Arabic–English CS el pharmacy is very boring [ . . . ] (Okasha 1999:110).

Verb satellites (also called particles) that occur with what are often called phrasal verbs are also early SMs because they depend on their heads for their appearance and they add meaning to their heads. Under the MLF model, these and other derivational morphemes may come from the EL because they are not the type of morpheme that the SMP restricts. EL phrasal verbs often appear with their EL verb satellites. An example from Swahili–English CS shows this, u-na-chase after (“you are chasing after”), as does another example from Arabic–English: [an engine is] locked up. In an Ewe–English example, an Ewe object suffix -e “him” can attach to the verb, as required in Ewe, but the particle remains in the EL, English: keep-e away from Eun (“keep him away from Eun”) (Amuzu 1998:53). That Ewe supplies the third person singular object suffix shows how the USP is supported as grammatical structure from the ML is maintained.
19.4.2.1 Plural markers as early SMs
Perhaps the most common early SM studied in Classic CS corpora is the plural marker. The language of origin of the plural marker in CS varies in four ways: the possible combinations are these: (1.) EL plural marking only, (2.) no plural marking at all, (3.) ML plural marking only, (4.) plural marked from both the EL and ML.

(1.) Most often, plural is marked on an EL noun by its EL plural affix, but no overt ML affix, as in Welsh–English CS in the phrase: y motorway-s na'r dual carriageway-s ("on the motorways nor the dual carriageways") (Deuchar 2006). In a Moroccan Arabic–Dutch example, duk artikel-en ("those articles"), the Dutch (EL) noun occurs with a Dutch plural suffix, but its plural determiner from Arabic (ML) shows agreement, thereby maintaining ML structure and the USP’s dictates (Boumans 1998:37). In example (3) workers has no ML plural marking, but its agreements indicate that the EL noun is operating as a class 2 Swahili noun. The demonstrative hawa ("these") and the associative wa ("of") show plural agreement (class 2).

(3) Swahili–English
Mbona ha-va worker-s wa East Africa Power and Lighting
wa-ka-end-a strike [. . .]
Why dem-c.2 worker-pl c.2-assoc East Africa Power and Lighting
c.2-consec-go strike
"Why did workers of East Africa Power and Lighting go [on] strike
[. . .]"
(Myers-Scotton 1993a, 1997:96)

In some language pairs with morphologically rich MLs, this is a frequent pattern. Even though an EL noun occurs without the ML plural marker, there is evidence that the ML assigns plurality features to the EL noun (see Myers-Scotton 2002a:127–31).

(2.) In some language pairs, an EL noun appears with no plural markings. In na date zingine ("with other dates") from Swahili–English, even though date has no plural at all, its modifier (zi-ningine) has a prefix from noun class 10 (zi-), indicating date is intended as a class 10 plural.

(3.) Perhaps less frequently, but still often, an ML affix marks plurality and there is no EL plural affix. For example, a Turkish (ML) plural suffix (-lar) occurs in the otherwise Dutch (EL) phrase klant-lar weglat-en (customer-pl plur away-chase-INF "chase away customers") (Boeschoten 1991:90).

(4.) Finally, sometimes both EL and ML early SMs occur with an EL content morpheme head; they are usually affixes. Both convey plurality, although they may contain other information as well. This “double morphology” can occur with other early SMs, but occurs most often
with plural affixes. For example, in the Acholi-English example lu-
civilian-s ("the civilians"). Acholi lu encodes both definiteness and
plural (Myers-Scotton 2005d) and English -s also encodes plural.

19.4.2.2 Early SMs and internal EL islands
When an EL early SM, particularly a plural marker, occurs with its EL
content morpheme head, it often occurs in a construction of EL elements
framed by the ML, as in Spanish–English tant-a-s thing-s ("so.many.FEM-PL
things-s"). In such instances, early SMs together with their content mor-
pheme heads are small EL islands (internal EL islands). These islands are
well-formed in the EL, but are part of a larger mixed constituent framed by
the ML. They are like other EL islands, but are smaller than phrasal level
constituents, full EL islands (e.g. the PP, on the weekend). Many internal EL
islands contain the crucial "chunks" of collocations that are then framed
by the ML (see Backus 2003, on multi-morphemic “chunks”). For example,
French–English la real thing (King 2000:100), Cajun–French le highest class
(Brown 1986:404), and sa little salary (see (4) below) do not occur as
maximal EL constituents. Instead, they occur framed by an ML element.
This is evidence that the USP is observed in bilingual speech whenever
possible.

(4) Wolof–English
Sa little salary rek la [. . .]
2SG.poss little salary only 3SG.cop [. . .]
"It is only your little salary, [. . .]"

(Haut 1995:52)

19.4.3 Late system morphemes
In contrast to early SMs, two types of late SMs are structurally-assigned.
The term “late” suggests that they are not activated until a later production
level. While early SMs largely build semantic structure, late SMs build
syntactic structure. These late SMs are labeled “bridges” and “outsiders.”
The DAH, discussed in §19.4.1, explains observable differences in data
distribution by postulating a fundamental difference in how late SMs are
accessed. It states that not all morphemes become salient at the same level
of language production. Information about content morphemes and early
SMs is available at the level of the mental lexicon; late SMs do not become
salient until the level of the formulator. The role of late SMs is to construct
larger constituents out of conceptually-activated morphemes; they assem-
ble phrases and connect phrases to realize full clauses. Put simply, late SMs
satisfy the requirements of the USP that constituents maintain a consistent
structure. "The late system morphemes [. . .] indicate relationships within
the clauses; they are the cement that holds the clause together" (Myers-
Scotton 2006a:269).
19.4.3.1 Bridge late SMs

Bridge late SMs provide “bridges” between elements that make up larger constituents. There is an important difference between bridge SMs and outsider SMs. For information about their form (and, indeed, their presence), bridges depend on information within their maximal projection, while outsider SMs depend on information outside the maximal projection in which they appear. Also, bridges seem to have an invariant form (they constitute a single allomorph); in contrast, outsiders seem to be part of a paradigm or conjugation (with more than one allomorph). English of is an example of a bridge SM, as in requirements of the college; so is ’s in Lena’s shoe. In French, de is an equivalent bridge, as in le français de Bruxelles (“the French of Brussels”). Example (5) shows a similar bridge from Hindi (ki) with a partitive meaning.

(5) Hindi–English
merii paatnii saaRii kii choice kar-egii
my wife saree of choice do-FUT.3SG.FEM
“My wife will choose a saree.”

(Ritchie and Bhatia 1999:273)

Because language-specific requirements for phrasal well-formedness vary, bridges are required in some languages, but not in others. For example, in many languages, weather expressions require a bridge. In these expressions, the subject pronoun does not receive a thematic role, e.g. French il pleut or English it is raining. In such expressions, the pronoun it is different from referring indefinite/antecedent third person singular it, a content morpheme in English (as in Where is the book? It is on the table.). Similarly, in American English, in certain expressions, determiners are bridges, not early SMs, as in this exchange: Where’s John? He had to go to the hospital. No definite hospital is indicated.

In CS corpora, most bridges come from the ML. Example (6) from an Acholi–English corpus shows a bridge me coming from the ML, Acholi. In the entire corpus, an English bridge occurs in only one formulaic El island (cost of living). Altogether, 42 Acholi associative constructions have at least one NP from English.

(6) Acholi–English
Chances me accident pol ka i-boarding taxi
chances assoc accident many if 2sg-board taxi
“[The] chances of [an] accident [are] many if you board [a] taxi.”

(Myers-Scotton 2005d:12)

EL bridges occur very rarely in mixed constituents, although there is an exception noted in the literature. When Arabic is the EL, sometimes it supplies the bridge djal in a clause framed by French as in French–Moroccan Arabic connaissance dyal la personne “knowledge of the person” (Bentahila and Davies 1998:38). The presence of dyal in
such cases does not violate the SMP. It is clearly a bridge SM, not an outsider.

19.4.3.2 Outsider late SMs
The second type of late SM is the outsider. As noted above, this morpheme type differs from bridges in that the presence and form of an outsider depends on information that is outside of the element with which it occurs. This information can come from another element in another constituent, or from the discourse as a whole. For example, subject-verb agreement is realized by outsider late SMs. However, in pro-drop languages, a late SM may be co-indexed with a null pronoun, whose relevant grammatical features come from the larger discourse. For example, in Spanish, the -en on the verb corr-en (run-3prt) is a late SM when it occurs as los estudiantes corren ("the students run") or simply corren ("[they] run").

There is good evidence from various sources that outsiders behave differently from other morphemes in many linguistic phenomena – see Myers-Scotton and Jake (2000) on Broca's aphasia and second language acquisition; Myers-Scotton (2002a) on speech errors and attrition; Myers-Scotton (2003) on split or mixed languages; and Wei (2000a) on second language acquisition. There is also scattered evidence in the literature about the distinctive distribution of outsiders in various contact phenomena. For example, Johanson (1998:251-3) notes that Turkic languages frequently borrowed conjunctions from other languages, but they "practically never" borrowed what he calls "relators." These relators include case markers, which are outsider SMs.

Certainly, outsiders are the most crucial and unambiguous purveyors of grammatical structures. They provide a more precise indexing of relations that extends beyond word order and basic constituent structure. Outsiders "knit together elements at another level" (Myers-Scotton 2005c:25). The grammatical relations indexed by outsiders reinforce semantic coherence within the clause and within the larger discourse. Furthermore, "[t]hese characteristics are the basis for an argument that outsider morphemes are the main bastion for maintaining uniform structure [the USP] in a clause" (2005c:25). Given that these characteristics define outsiders, it follows that the distribution of outsider late SMs should be the most defining feature of Classic CS - and it is. With few exceptions, outsiders always come from the ML in mixed constituents.

In some CS data sets, ML outsiders as AGR features occur with EL verbs, as in example (7): the third person singular prefix i- on appartemir shows subject–verb agreement, referring to richesse, the subject of the clause. The object prefix -tu- ("us") refers to the speaker and previously mentioned others. In addition, the class 9 prefix y- on y-ote ("of all") is also an outsider, as is the prefix on y-ake ("his").
(7) Shaba Swahili–French

Donc, (h)jii richesse y-ote (h)jii i-na-tu-appartenir shi
So, c.9.DEM riches c.9-ALL c.9.DEM c.9-NON-PAST.OBJ.1PL-belong us
ba-toto y-ake
c.2-child c.9-his
"So, all these riches, it belongs to us, his children."

(De Rooij 1996:186)

When the ML is a language with case assigning verbs (and/or prepositions), case markers are also outsiders. Almost without exception, EL elements receive the expected ML case marker as in (8), in which Dutch terras receives locative case from Turkish. Similarly, in (9), English grass is inflected with prepositional case from Russian.

(8) Turkish–Dutch

èvet, terras-ta oturuyorlar
yes cafe-LOC sit-PROG.3PL
“Yes, they are sitting at the outdoor cafe."

(Backus 1996:140)

(9) Russian–English

Zachem ty na grass-e valjajesh’sja
what-for you.SG on grass-PREP.SG roll-around
“Why are you rolling around on the grass?”

(Schmitt 2006)

19.5 An overview of morpheme types

An important advantage of the 4-M model over other approaches to classifying morphemes is that it eliminates the problem that lexical category membership does not predict morpheme type. That is, members of a category need not be members of the same 4-M morpheme type. In fact, in terms of the 4-M model’s defining features for morpheme types, some morphemes in the lexical types we consider here (prepositions, pronouns, and complementizers) are content morphemes, but others are early SMs, and still others are either bridge or outsider late SMs. The Chomskyan lexical-functional element dichotomy does not account for these differences. The premises of the 4-M model that allow for such a flexible classification are supported by findings in CS data, other contact phenomena, and evidence from various types of language acquisition and loss. Simply put, not all prepositions, pronouns, or complementizers have the same distribution.

19.5.1 Prepositions

Linguistic theory has long recognized that prepositions do not behave as a uniform class (see e.g. Abney 1987). Under the 4-M model, prepositions
can be content morphemes or any of the three types of SMs. Sometimes the same phonological form fits into more than one category. For example, in *He walked across the street*, *across* assigns a thematic role and is a content morpheme. In CS, content morpheme prepositions can come from the EL, as in (10). There are not many examples of such EL prepositions in mixed constituents; more frequently, they occur in PPs that are EL islands (e.g. *before tomorrow evening*).

(10) **Swahili–English**
Labda, [.] bring it at my home. U-let-e before kesho jioni.
perhaps, bring it to my home. 2sg-bring:subjunct before tomorrow evening
"Perhaps you should bring it to my house. You should bring it before tomorrow evening."

(Myers-Scotton 1993a, 1997:124)

Some prepositions can also be indirectly selected at the conceptual level, and are then early system morphemes: in *he comes across as ill-prepared*, *across* occurs with *come*, its content morpheme head. The discussion in §19.4.2 above includes examples in which prepositions are satellites of phrasal verbs, and suggests that the satellite comes from the same language as the verb, either EL or ML verb. However, sometimes such early SMs occur in the EL even when the verb is in the ML. In (11), the expression "change [something] around" is realized in both Spanish and English, with English supplying the preposition *around*.

(11) **Spanish–English**
Sabes los cambian *around*  
know.pres.2sg them change.pres.3pl around
"You know they change them around."

(Pfaff 1979:303)

Sometimes prepositions are late SMs that are not activated until the level of the formulator; these primarily contribute structure, and not content. For example, prepositions that are bridge SMs make a phrasal constituent well-formed. Although EL bridge prepositions can occur in mixed constituents, very few actually do. As noted above in some French–Arabic CS, Arabic *djal* (equivalent to "of") occurs in associative constituents in French-framed CPs. Below are discussed some instances of EL bridges that occur with more frequency, namely, Comp bridges.

Some locative prepositions are bridges; they do not encode directionality or motion, but locate a figure with respect to a ground (see Talmy 2000). For example, in *Joe's in school*, *in* adds little conceptual information to the mapping of the theme (*Joe*) to the ground (*school*). Such bridge prepositions show variation (e.g. *Joe's at school*). However, *in* can also be a content morpheme or an early SM. In *He's all done in*, *in* is an early SM. Further, *in* is a content morpheme in *In this example, they illustrate the distinction*; like
other content morphemes, the thematic role assigned by in can be questioned, as in *Where do they illustrate the distinction?*

In some languages, prepositions are also outsiders. Consider Spanish *a*. It can be a content morpheme assigning directionality, as in *va a Hamburg* "he/she goes to Hamburg," or an early SM, as in *miremos al año que viene*, ("we are looking forward to the coming year") (al = a + el "to+the.m.def"). As a bridge, it connects purpose infinitives with matrix CPs, as in *prepare a venir* ("prepare to come"). Finally, *a* occurs as an outsider when it assigns case to animate direct objects, as in *veo a Eva* "I see Eva." The prediction is that as ML, Spanish will supply personal *a* to objects in mixed constituents, as in *refieres a tus coworkers* ("you are talking about your coworkers") (Jake et al. 2002), but that as EL, Spanish NP animate objects do not have to occur with personal *a*, as in *the police officers have seen un ladrón* ("The police officers have seen a thief") (Belazi et al. 1994:230).

The 4-M model articulates how morphemes are classified. Even so, the fact that one prepositional form can be activated at more than one level and is thus subject to different conditions in CS demands careful analysis. The SMP requires that all outsider prepositions come from the ML in mixed constituents. The distribution of bridges in most CS also supports the USP; one language, the ML, provides most of the grammatical frame.

### 19.5.2 Pronouns

Pronouns are another lexical category that is not uniform because they can be members of any of the four morpheme types (see Jake 1994). Some are content morphemes; i.e. they occur in argument position and receive thematic roles. As content morphemes, EL pronouns can occur in clauses framed by the ML. For example Klintborg (1995) reports English pronouns in Swedish-framed clauses in Swedish–American English CS, as in *När vi var hemma sista gången me and min hustru* ("When we were home last time me and my wife") and [*...* he var smed för Tyket* ("[... ] he was a blacksmith by trade").

But even when pronouns are content morphemes, EL pronouns occur very infrequently except in EL islands. Why? First, preference is for ML elements. Also, ML counterparts play a role. When pronouns in the ML are clitics or affixes licensing null pronouns in argument position, they are outsider SMs and must come from the ML (see the Blocking Hypothesis (Myers-Scotton 1993a, 1997:120) which requires congruence with the ML).

However, EL content morpheme pronouns that establish topics or contrast occur widely in bilingual clauses. They convey both conceptual and procedural information, as noted by Wilson and Sperber (1993:21). Example (12) shows an English pronoun within a Malay
grammatical frame, expressing a “dual notion of fusion and contrast” (Jacobson 2000:68).

(12)  Malay–English

Oh! About the recent controversy? I tak bersetujulah kalau women stay at home.

Oh! About the recent controversy? I not agree-emph if . . . “Oh! About the recent controversy? I don’t agree that women should stay at home.”

Other examples from diverse language pairs abound. Haust (1995) includes examples of Mandinka emphatic pronouns occurring in English-framed CPs and Wolof-framed CPs. English contrastive topic pronouns occur in Spanish-framed CPs, as in You estás diciéndole [sic] la pregunta in the wrong person (“You are asking the question in the wrong person”) (Sankoff and Poplack 1981:13).

Some researchers have commented on “pronoun doubling,” as in (13), but this is not the true doubling that occurs with early SMs. Each pronoun is activated independently and occurs in a separate position in the bilingual CP. In (13), for example, the Arabic discourse emphatic pronoun nta (“you”) is adjoined under Comp and the French tu (“you”) is an agreement clitic not in argument position. A null pronoun is assumed to occur in subject position. In (14), Arabic ?ihna (“we”) is adjoined under Comp and English we occurs in subject position.

(13)  Moroccan Arabic–French

nta tu va travaillier

2SG.EMPH 2SG go work-INF

“You, you are going to work.”

(Bentahila and Davies 1983:313)

(14)  Palestinian Arabic–English

?ihna we are supposed to be nudris-ing

1PL.TOP . . . study-PROG

“We, we are supposed to be studying.”

(Myers-Scotton et al. 1996:27)

And in the Spanish–English example cited above, the English emphatic pronoun you is a topic, and a Spanish null pronoun occurs as the subject.

In sum, the distribution of pronouns in CS reflects their classification under the 4-M model. The requirements of the MLF model foreshadow the import of this classification. Both the SMP, which requires ML outsider pronouns in mixed constituents, and the Blocking Hypothesis, which requires cross-linguistic congruence, imply how critical it is to recognize morpheme type at the abstract level of clause construction. Taken together, they maintain the integrity of the frame in line with the USP.
19.5.3 Complementizers and other clause connectors

Complementizers and complementizer-like elements are similar to prepositions and pronouns in not showing a uniform distribution in CS. In current syntactic theory, COMP is the head of any clause identified as CP, projection of Complementizer. Variation among COMP elements themselves and cross-linguistic variation in their patterning in CS complicate their discussion. Also, there is no uniform agreement regarding what elements are rightly classified under COMP. Complementizers include not just elements such as that, but also subordinating conjunctions, relative clause markers, other elements that indicate clause boundaries, and even coordinating conjunctions. These elements are discussed according to how they are elected and how they participate in the construction of a CP.

Like pronouns, most complementizers convey procedural knowledge. Many constrain the truth-conditions of propositions and participate in the discourse-thematic structure of propositions. For example, porque “because” in (15) assigns a discourse-level thematic role of Cause or Reason.

(15) Spanish–English
trabaje menos porque then I didn’t know some of his business
work.PRET.1SG less because . . .
“I worked less because then [i.e. at that time] I didn’t know some of
his business.”

(Pfaff 1979:312)

Most complementizers straddle two CPs. In this way, they are at the intersection between inter-sentential CS and intra-sentential CS. In example (15), porque is the head of the subordinate CP (hereafter CP2), yet it is within the domain of a matrix clause (hereafter CP1), and is in the language of CP1.

19.5.3.1 Overview of complementizer types

Several factors play a role in determining the source of complementizers. These include the type of morpheme, the grammatical requirements of the participating languages, and the overall pattern in the discourse. Complementizers from one of the participating languages are preferred if that language is typically the ML in mixed constituents throughout the corpus.

Under the 4-M model, some complementizers and complementizer-like elements are bridge SMs, especially complementizers such as “that.” Similar complementizers are multi-morphemic elements that include a bridge and an outsider SM. For example, in Arabic, ?in- (“that-”) occurs with a suffix agreeing with the subject of CP2. Finally, many subordinators and coordinators are content morphemes (e.g. French alors “then” or German aber “but”). In many corpora, content morpheme subordinators
and coordinators tend to come from the ML of either CP. In (15) above, 
*porque* comes from the ML of CP1.

19.5.3.2 The language of the complementizer
If CS occurs at the clause boundary, is the complementizer in the language
framing CP1 or CP2? It appears that in some language pairs, the comple-
mentizer can come from either language. In Spanish–English CS, for
example, *que* can introduce an English CP (see 16), and *that* can introduce
a Spanish CP (see 17). However, the complementer can also be in the ML
of CP2, as in (18) and (19).

(16) Tonces salió eso que she wanted to take mechanics
“Then it turned out that she wanted to take mechanics.”
(Lipski 1978:258)

(17) They sell so much of it that *lo están sacando* y many people [...] 
...it be 3pl take-PART and ...
“They sell so much of it that they’re taking it out and many people
[...]
(Pfaff 1979:312)

(18) Sí, but the thing is *que empiezan bien recio* and [...] 
“Yes, but the thing is that [[they]] start[3pl] pretty fast and [...]”
(Pfaff 1979:315)

(19) El profesor dijo *that the student had received an A *
“The professor said that the student had received an A.”
(Belazi et al. 1994:234)

19.5.3.3 *That*-type complementizers as bridges
The distribution of complementizers such as *that* and *que* supports their
analysis as bridge SMs. *That*-like complementizers allow a larger constit-
uent, a multi-clause structure, to be constructed out of an embedded CP and
a matrix CP. And unlike content morphemes and early SMs, bridges convey
little representational meaning. In this way, *that*-like complementizers are
different from other complementizers such as *when and before*, and their
equivalents across languages.

It is not surprising that bridge complementizers can come from
either language with CS at clause boundaries. This is because, although
bridges join two constituents together, they are invariant placeholders
satisfying well-formedness conditions for the larger unit. In some lan-
guage pairs, *that*-like complementizers come from one specific partici-
paring language, regardless of the ML of CP1 or CP2. For example,
in Chichewa–English CS, the bridge complementizer always comes
from Chichewa, as in (20). In Simango’s (1996) corpus, *kuti* introduces
over 20 English-framed clauses, but *that* never introduces any English
clauses.
(20)  Chichewa–English
   [. . .] a-ka-tsimidz-e kuti *this was the end*
   [. . .] 3SG-CONSEC-confirm-SUBJUNCT that . . .
   '[. . .] he confirms that this was the end.'

(Simango 1996)

This suggests that language-specific factors are involved. For example Chichewa requires a complementizer, whereas English does not.

19.5.3.4 Outsider complementizers
As noted above, Arabic *?inn*-agrees with the subject of CP2. In Arabic–English CS, these multi-morphemic complementizers always come from Arabic. For example, Okasha (1999) reports that *?inn*-introduces 10 clauses entirely in English, as in (21), and 15 bilingual clauses framed in English. No English complementizers occur with Arabic clauses.

(21)  Palestinian Arabic–English
   ka’an el-*doctor* yišuq *?innu it is not reliable*
   PERF:3MASC:be the-doctor imperf:3MASC:doubt that:3MASC it is not reliable
   '[he] was, the doctor, doubting that it was not reliable'

(Okasha 1999:71)

In the case of Arabic–English CS, the USP is better satisfied when the complementizer comes from the ML of CP1. This configuration means that the following IP (clause) is in the EL of CP2. That is, *?inn-* does not just bridge two CPs; it coindexes an embedded CP with the ML of CP1, and frames the entire multi-clausal constituent in one language, Arabic. Thus the English clause in (21) is an embedded IP island (and English is not the ML of CP2).

Not only are Arabic bridge complementizers inflected with outsider SMs, but so are subordinators, which are content morphemes in many languages. In (22) *li?anhum* "because" agrees with the third person plural subject *they* of the English IP.

(22)  Palestinian Arabic–English
   [. . .] huma biyi[a]d[a]n9ooli kul haga li?anhum *they can afford it*
   [. . .] they HAB:IMP:3PL:pay:1SG every thing because:3PL . . .
   '[. . .] they pay for everything [for me] because they can afford it.'

(Okasha 1999:123)

19.5.3.5 Content morpheme complementizers
Adverbal-like subordinators are conceptually-activated content morphemes. They introduce discourse-thematic roles. When they are unaffected, the language of the complementizer can be different from the language of the CP it introduces. Thus, *porque* "because" can introduce an English CP, as in (15) above. And in many other language pairs, the language of such subordinators appears free. For example, in Wolof-French
CS, French subordinators introduce clauses that are bilingual or mixed (Swigart 1992a, 1992b, 1992c).

Under the 4-M model, most coordinating conjunctions are also content morphemes. They reflect procedural knowledge and are truth conditional, although some have more procedural content than others. In (23), English conjoints two Xhosa clauses.

(23) Xhosa–English

[…] ba-se-msebenzi-ni and umalume be-ka-khal-a kude ku-na-thi […]

[…] 2.pl-loc-work-loc and my uncle PST of-stay-FV far loc-have-us […]

“[…] they were at work and my uncle lived far from us […]”

(Myers-Scotton 2005b)

19.5.3.6 Summary: The language of the complementizer

In summary, although EL complementizers do not appear in CS as freely as some categories, participation largely depends on their morpheme type. EL subordinators and coordinators, usually content morphemes, are quite robust. However, subordinators or that-like complementizers that include an outsider seem to always come from the ML, even when the outsider depends on a CP2 whose clause is otherwise framed by the EL. That-like complementizers which are bridges are a mixed bag. With some languages, these complementizers must come from the ML, with others they do not.

When the complementizer of CP2 is in the ML of CP1, it reinforces uniformity across clauses. This suggests that only when the complementizers of the participating languages are congruent enough not to violate language specific requirements do complementizers come from the ML of CP2 (e.g. Spanish–English CS). However, no matter what the morpheme type, whatever ML dominates in the discourse seems to preference complementizers from that language, reflecting a more general organizing principle, the USP.

19.6 Conclusion

One goal of this chapter has been to offer implications for the nature of bilingual production and processing that arise from considering naturally occurring CS corpora in terms of the MLF model and the 4-M model. In turn, this leads us to a more universally applicable characterization of linguistic structure, the Uniform Structure Principle (USP). In bilingual utterances, there is no a priori reason to expect uniformity in clause structure; the significance of the principle for contact linguistics is that the USP predicts uniformity. It preferences the structures of only one of the participating languages. The extent to which this principle is supported
implies certain preferences for how the cognitive component supporting bilingual speech is structured.

Early researchers in contact linguistics avoided CS, focusing instead on possible contact-induced change or dialectal variation. For example, the father of modern contact linguistics, Weinreich (1954, 1967) famously dismissed bilingual CS clauses in a way that implied that looking for organizing principles in CS was a theoretical dead end. In contrast, the USP, along with the MLF and 4-M models, predict that a principled account of CS is possible. Thus, the mantra of the USP is "no chaos allowed." Such an account depends on premises about predictable divisions between the roles of participating languages and morpheme types. In turn, these divisions motivate a model of language production and present implications about organization within the cognitive components supporting language.

19.6.1 Predictable patterns
The bulk of this chapter is descriptive, but with the theoretical goal of demonstrating how the asymmetries that one finds in CS show a predictable pattern. The goal here has been to demonstrate that the contributions of morphemes of participating languages depend on the four morpheme types and to relate this observation to production. When one views morphemes in terms of these types and in terms of the Differential Access Hypothesis, a principled explanation for differences in their cross-linguistic distributions in CS is forthcoming. Three examples illustrate insights of this chapter.

First, it is predicted that double plural marking on EL nouns is possible, but that double subject-verb agreement or double case marking is not. This disparity is owed to the difference in how these morphemes are accessed in language production. Plural affixes are early SMs while subject-verb agreement and case markers are outsider SMs. EL early SMs can be accessed with their content heads because they are conceptually activated and available in the mental lexicon, but outsiders become salient only later, at the level of the formulator.

Second, a strong preference for the ML to supply "that-type" complementizers at clause boundaries is predicted. However, subordinating complementizers are less constrained and they come from either language. The reason again is differences in morpheme type and hypotheses about their production history. That-type complementizers are bridge SMs and, although not as critical in building clauses as outsiders, are still part of constituent structure and are salient at the level of the formulator. When they come from the ML, the USP is satisfied. In contrast, many subordinators are content morphemes; they are activated by speakers' intentions (conceptually activated), and are available in the mental lexicon.

Third, prepositions fall into all four types under the 4-M model, but not all types are predicted to come from the EL, or with the same frequency. In
fact, EL prepositions do not occur frequently, and this may be because of their role in structuring constituents and the requirement of uniform structure, i.e. the USP. Among prepositions, early SMs are more frequent, perhaps because they allow for an elaboration of pragmatic and semantic structure without creating syntactic structure. Prepositions that are content morphemes (i.e. that assign thematic roles) are the most unconstrained; even so, they are not frequent.

Of the three types of SMs, early SMs are the least constrained because they are conceptually activated, whereas bridges and outsiders are structurally assigned. Even so, the most frequent early SMs seem to be the most contentful ones, plural affixes and definite articles. Only a few types of EL bridge SMs occur, and no EL outsider SMs occur in mixed constituents (except for fairly rare types of EL islands). More research needs to be done on EL islands, but the overall point about prepositions holds for all lexical categories: morpheme type, as discussed at many points above, makes the difference in their distribution.

19.6.2 Cognitive support systems
Such differences in morpheme distribution across languages are systematic and follow from the MLF and 4-M models, as well as the USP; they are also empirically verifiable. However, these differences also imply some speculations about language production and the cognitive systems supporting language. First, the division of labor between languages in CS, with only one language providing the morpho-syntactic frame of the bilingual clause, seems to imply some sort of divisions within the cognitive component supporting this surface asymmetry. Certainly, any intra-clause CS implies that both languages are active during bilingual production, but that the ML has a higher level of activation.

Second, the USP implies that cognitive energy is conserved by allowing only minor-level switches to EL islands. If we look at the frequency with which EL content morphemes are integrated into an ML-framed structure and juxtapose this frequency with the USP's injunction against changing languages, one conclusion is that accessing words from the EL requires a different type or level of activation than creating morpho-syntactic structure.

Third, the role of the EL is explicitly limited. Most obviously, the EL never structures any constituents that include ML morphemes. This means that the EL has little opportunity to supply any outsider SMs to the bilingual clause except in EL islands, but typical EL islands have few structures that would require outsiders. Further, except for occasional early SMs or even less frequent bridge SMs, the EL supplies only content morphemes within constituents structured by the ML. The dearth of outsider SMs from the EL motivates the conclusion that the cognitive component supporting this morpheme type may be independent from that which coordinates
simpler syntactic constructions. Keep in mind the complex tasks that this mental architecture must accomplish: outsiders are critical in signaling thematic roles and other relationships of the semantic-syntactic interface; without them, there can be no clause.

19.6.3 Testing hypotheses

Finally, as already noted at many points above, CS data support the DAH that is derived from the 4-M model. This hypothesis suggests a language production model in which some of the elements underlying surface level morphemes are salient at one level and others are not salient until another level. Specifically, late SMs are not salient until they are called by the lemmas underlying content morphemes to construct larger constituents in the formulator. Obviously, the extent to which this hypothesis is supported has relevance beyond CS and other types of contact phenomena to both child and second language acquisition, among other topics. The results of psycholinguistic experiments testing this hypothesis will add crucial support to a language processing model that accommodates the notion of saliencies at different levels (Myers-Scotton 2006b). Not only does this matter for production models, but also for comprehension models.

19.6.4 Supporting the Uniform Structure Principle (USP)

In sum, this chapter has shown how the distribution of morpheme types in Classic CS, at least, is compatible with predictions of the MLF and 4-M models, and the USP. As an empirical window on divisions of labor between participating languages, CS implies intriguing hypotheses about some ways in which language is supported in our cognitive systems.