CHAPTER 12

Mixed NPs in Spanish-English bilingual speech

Using a corpus-based approach to inform models of sentence processing

Jorge R. Valdés Kroff
University of Florida

Bilinguals speaking with other bilinguals engage in codeswitching (CS). CS is not a priori predictable, yet bilinguals suffer no appreciable costs to communication. One hypothesis explaining this ease is an exposure-driven account whereby speakers converge upon conventional production patterns, which may help guide comprehension. In this study, I quantify and investigate the use of grammatical gender in Spanish-English mixed noun phrases using a bilingual spoken language corpus. Results reveal a robust gender asymmetry where masculine gender is the default gender when switching into an English noun (e.g. *un* juice 'the$_{masc}$ juice,' *un* cookie 'the$_{masc}$ cookie'). In contrast, feminine-marked switches are infrequent and used with feminine translation equivalents (e.g. *una* cookie, 'the$_{fem}$ cookie'). This asymmetry forms testable predictions for how bilinguals use grammatical gender in CS comprehension.

Keywords: grammatical gender, mixed noun phrases, gender asymmetry, default gender, corpus study

1. Introduction

Current psycholinguistic research on bilingualism converges on the finding that a bilingual's two languages are simultaneously active to varying degrees (e.g., Kroll, Sumutka, & Schwartz, 2005). Although a bilingual speaker may intend to produce or comprehend solely in one language, lexical information from the non-target language is also accessible. Researchers have been greatly informed by the evidence suggestive of this parallel co-activation, yet the overwhelming focus is on how bilinguals are able to produce or comprehend solely in one language. Nevertheless,

DOI 10.1075/ihll.11.12val

bilinguals engage in a specialized linguistic skill known as *codeswitching* (CS), defined as the fluid alternation between languages in discourse (Poplack, 1980). By its very act, CS requires the heightened co-activation of a bilingual's languages in order for a speaker to successfully integrate the phonological, morpho-syntactic, semantic, and discourse properties of both languages.

In addition, CS presents an informative scenario into how bilinguals negotiate cross-linguistic structures that may not fully be equivalent across the two languages. Codeswitches often occur within major syntactic clause boundaries (i.e., intrasentential CS), yet the grammatical features of the bilingual's two languages need not be exactly compatible at the switch juncture. For example, Spanish nouns encode for grammatical gender (masculine and feminine, e.g., *carro* 'car_{masc},' *casa* '$house_{fem}$') whereas English does not (e.g., $the_{ø}$ $car_{ø}$). Nevertheless, switching between determiners and nouns is common for Spanish-English bilinguals (Pfaff, 1979; Poplack, 1980). In light of the growing evidence in favor of non-selectivity, bilingual codeswitchers must confront these cross-linguistic differences while seamlessly integrating them in production and comprehension. Therefore, the study of the production of codeswitched speech presents a useful tool to investigate cross-linguistic differences and the subsequent impact to the comprehension system. In this manner, CS provides a fruitful avenue for testing experience-based accounts of sentence processing (Gennari & MacDonald, 2009; Trueswell, Tanenhaus, & Garnsey, 1994).

In the domain of sentence processing, some models promote the view that the production and comprehension systems are tightly linked with the bulk of primary evidence coming from monolingual data. MacDonald and colleagues have proposed one such model, the Production-Distribution-Comprehension (PDC) framework (Gennari & MacDonald, 2009; MacDonald, 1999; MacDonald & Thornton, 2009). This framework adopts an *emergentist* view of language use, following from the hypothesis that language use leads to broad distributional patterns over time. The PDC framework explicitly states that these accumulated distributional patterns will have an impact on comprehension such that in alternating structures, the more frequently used alternative will consequently facilitate comprehension. This effect becomes apparent in optionally equivalent structural choices that speakers regularly encounter in any language. For example, some verbs optionally take either a direct object (DO) or a sentential complement (SC) as an argument, e.g., in English, the verbs *admit* and *believe* as illustrated below.

(1) a. The boy admits [the truth]$_{DO}$
    b. The boy admits [the truth was not discovered]$_{SC}$

(2) a. The boy believes [the truth]$_{DO}$
b. The boy believes [the truth was not discovered]$_{SC}$

Despite the surface equivalence between options a and b in examples (1) and (2), several studies have demonstrated that speakers prefer associating specific verbs with specific argument structures, and this information may be language-specific, a phenomenon known as verb bias or subcategorization (Dussias & Cramer Scaltz, 2008; Garnsey, Pearlmutter, Myers, & Lotocky, 1997). In English, *admit* occurs more frequently with DO complements (DO verb bias, example (1a)) whereas *believe* appears more frequently with SC (SC verb bias, example (2b); Garnsey et al., 1997). Consequently, the PDC framework predicts that speakers will have more difficulty parsing DO biased verbs with SC arguments (1b) and SC verbs with DO arguments (2a).

CS provides another means of testing the production-comprehension link. For production, CS can be characterized as a choice between languages, thereby drawing an analogy to structural alternations in unilingual contexts. In this paper, grammatical gender is the morpho-syntactic feature of focus due to its role in Mixed NP constructions, i.e., codeswitched noun phrases (NP) in which a Spanish determiner (DET, e.g., *el* or *la*) is paired with an English noun, e.g., *el* cookie 'the$_{masc}$ cookie'.[1] The goal of this study is to quantify the production of these Mixed NPs with a particular focus on gender assignment using a bilingual spoken language corpus collected in Miami, Florida. Beginning with a broad approach to quantification, two types of Mixed NPs were extracted: Spanish determiners switching into English nouns (e.g., *el* juice 'the$_{masc}$ juice') and English determiners switching into Spanish nouns (e.g., the *jugo* 'the juice$_{masc}$' Both types of Mixed NPs are included to replicate previous findings that Spanish determiner Mixed NPs more frequently surface in Spanish-English CS (Herring, Deuchar, Parafita Couto, & Moro Quintanilla, 2010).[2] The subsequent analysis focuses on Spanish determiner Mixed NPs by conducting cross-tabulations of the gender assignment of the determiner with the concurrent gender of the Spanish translation equivalent of the English noun.

1. For the remainder of the paper, codeswitches will be demarcated with Spanish elements in Italics. Additionally, following Herring et al. (2010), I use Mixed NP as a theory-neutral term, i.e., some syntactic theories would classify the construction under focus as Mixed DPs.

2. Herring et al. (2010) observe a distributional asymmetry in their corpora but do not attribute a theoretical claim that this asymmetry should always occur. Rather, it could be an epiphenomenon of these Mixed NPs occurring in a higher number of Spanish matrix language clauses (see p. 570).

Foreshadowing the results, feminine marked Mixed NPs are highly infrequent yet occur in bilingual speech. In terms of a psycholinguistic perspective on feminine marked Mixed NPs, rather than treating these infrequent constructions as performance-driven exceptions, their use lead to intriguing questions for the bilingual production and comprehension systems. My ultimate suggestion here is that CS is by and large a planned mode of bilingual speech (e.g., Green, 2011; Soares & Grosjean, 1984). Bilingual codeswitchers are driven to follow community-established production patterns, i.e., in this case, the use of masculine-marked determiners in Mixed NPs, which in turn become useful cues for the comprehension of codeswitched speech.

The remainder of this chapter is organized as follows: first, Spanish-English Mixed NPs are briefly described. The following section summarizes the corpus used for the study and the methods for extraction of Mixed NPs. This is followed by the results of quantification and further analyses concerning the Spanish determiner Mixed NPs. Finally, the paper ends with a discussion on how the results fit with traditional accounts. This discussion is extended with a psycholinguistically-informed hypothesis in which CS is treated as an emergent linguistic system built from the bilingual's constituent languages. I conclude with an overview of how the results inform predictions for experimental research investigating the comprehension of Mixed NPs based on the framework of the PDC model.

2. Spanish-English Mixed NPs

The Mixed NP consists of two main structural elements, a determiner (e.g., demonstrative, article) and a noun phrase, with each element in a different language. Spanish and English determiners differ in that Spanish obligatorily encodes for grammatical gender on some determiners, whereas English only sometimes encodes for number on determiners. As a consequence, Mixed NPs are manifested in several ways[3] as illustrated in Table 1.

As the CS examples in Table 1 highlight, when excluding gender-less determiners in Spanish, e.g. *su* house 'his/her/your/their house,' there are three possible Mixed NP constructions: English determiner + Spanish NP, Spanish feminine determiner + English NP, and Spanish masculine determiner + English NP. These observations naturally lead to the question of gender assignment in Spanish determiner Mixed NPs. As the masculine-marked Mixed NP in Table 1 shows, (i.e.,

3. Remaining examples are taken directly from the Bangor Miami Corpus (Deuchar et al., 2014) and are followed by an anonymized filename and speaker label.

Table 1. Examples of possible NPs across Spanish, English, and CS

Language mode	Language of determiner	Example
Spanish	Spanish	*En alguna parte tiene que ser las cinco de* [*la tarde*] (herring11.GRA) 'Somewhere it has to be 5 o'clock in [the afternoon]'
English	English	And you went to work with [those shoes]? (herring08.ROB)
Codeswitching	English	She got [the *manguera*] (sastre4.fem1) 'She got [the hose]'
	Spanish-feminine	I'm looking for something *con* [*las tres b's*]: *bueno, bonito y barato* (zeledon5.fem1) 'I'm looking for something with [the three bee's]: good, beautiful, and cheap'
	Spanish-masculine	You need to tell him, "Look! *Te voy a poner* [*un* restraining order] on you." (sastre4.fem1) 'You need to tell him, "Look! I'm going to put [a restraining order] on you.'"

un restraining order, Sp. *una orden de restricción/protección*), the gender of the article and that of the Spanish translation equivalent of the noun do not obligatorily match as required in Spanish.

The literature on Mixed NPs in Spanish-English CS has most prevalently focused on (1) the language of the determiner (Herring et al., 2010; Jake et al., 2002) or (2) whether Spanish determiner Mixed NPs follow unilingual Spanish gender assignment constraints (e.g., phonological, semantic, syntactic constraints; Clegg, 2006; Poplack, Pousada, & Sankoff, 1982). On the language of the determiner, the most prominent current frameworks applied to CS (Minimalist approach [MP]), MacSwan, 1999, 2000; Matrix Language Framework [MLF], Myers-Scotton, 1993, 2000; Myers-Scotton & Jake, forthcoming) agree that the determiner is most likely to surface in Spanish (this is categorical under MP and highly likely under MLF), at the most basic level because Spanish has grammatical gender[4] (cf. Herring et al., 2010). On the question of gender assignment in Mixed NPs, previous accounts have included broader contact phenomena, such as phonologically adapted borrowings, e.g., *la breca* ('the brake', Sp. *el freno*; Clegg, 2006) and

4. How grammatical gender influences language of determiner centers on feature checking operations (MP) or Myers-Scotton and Jake's (2015) more recent notions of speech planning "cost" where grammatical gender results in "earlier" and hence, less costly activation (see pp. 417–418; 434, 436). Specific details of both accounts fall outside the scope of this chapter; see references included.

subsequently, give prominent weight to phonological constraints as explained in Section 4. In the study reported here, I limit Mixed NPs to phonologically unadapted mixed elements and focus on gender assignment in Spanish determiner Mixed NPs.

Given the possible permutations for Spanish-English Mixed NPs, the research question under investigation is whether speakers show a clear preference for one combinatorial form. Due to the cross-linguistic difference between Spanish and English in the use of grammatical gender, speakers may adopt an English-like pattern and neutralize the grammatical gender of Spanish nouns. Thus, speakers would show a preference for the use of masculine determiners regardless of the gender of the Spanish translation equivalent, e.g., *el* cookie 'the$_{masc}$ cookie'. Speakers may choose instead to adhere to the Spanish grammatical gender system, favoring a constraint hierarchy that follows Spanish. For example, the gender of human referents may constrain gender assignment in CS, e.g., *la* mother 'the$_{fem}$ mother' because *mother* has an unambiguous female referent. Finally, speakers may adopt a hybrid strategy that is neither fully English- or Spanish-like. I explore this question using the Bangor Miami Corpus, explained in more detail in the following section.

3. Current study[5]

3.1 Materials and participants

The Bangor Miami Corpus was obtained in collaboration with Margaret Deuchar and colleagues. They have made the corpus publicly available online at <http://www.talkbank.org/data/BilingBank/Bangor/>. Detailed information concerning the bilingual corpus is found in Deuchar, Davies, Herring, Parafita Couto, and Carter (2014). Here, I summarize the most relevant details. The corpus was collected over a period of two months, April-June, in 2008 in Miami, Florida. Two on-site assistants helped a member of the research team in recruitment of Spanish-English bilingual participants, which involved employing the "friend of a friend" strategy advocated by Milroy (1987).

5. The research reported in this paper was supported in part by an NSF Minority Postdoctoral Research Fellowship (SMA-120364) to Jorge Valdés Kroff and NSF Dissertation Improvement Grant (BCS-1124218) to Paola Dussias, Chip Gerfen, and Jorge Valdés Kroff. The author would like to thank Paola Dussias and Rena Torres Cacoullos for their continuing guidance and feedback on this work.

Potential recruits were administered a Language History questionnaire. In order to address inherent problems concerning the "Observer's Paradox",[6] recruited individuals were asked to choose their own conversation partners and to select their preferred place for recordings. Individuals were briefed before recordings began that the primary objective of the study was to investigate how bilinguals speak with each other without any specific mention of CS. Most recordings were made in pairs, although some of the recordings include more individuals. All recordings lasted at least 35 minutes.

The completed corpus includes 27 separate sound files composed of 85 speakers (62% female). Of the total group, 73% of participants rated their proficiency as high in both languages. Ages ranged from 9 to 66 years old (M = 32 years old). There were 43 different responses for Occupation with the top three responses including Student (n = 23), Teacher (n = 6), and Office Manager (n = 4). Participants gave 19 different responses for Nationality, with the top three responses including American (n = 23), Cuban or Cuban-American (n = 28), and Colombian (n = 7).[7]

This procedure resulted in spontaneous and natural conversation that at times reached very intimate levels, indicating that conversation partners did not feel constrained by the presence of the recording equipment. Topics ranged widely, including discussions on food, social life, jobs, school, travels, etc. Once the recordings were completed, researchers acquired consent from any individual who was recorded, including unannounced visitors, with individuals given the opportunity to indicate if there were sections that they did not want to include in the final recording. Deuchar et al. report that participants did not elect to omit any significant portion of their recordings. As an additional step, the corpus omits the first five minutes of each recording to remove participants' initial phases of discomfort or novelty to being recorded.

3.2 Methods

Mixed NPs were extracted from 25 of the total 27 sound files. At the time of extraction, only a subset of the files (n = 16) had been completely transcribed (they are currently all transcribed). For sound files with completed transcriptions, the

6. The Observer's Paradox, simply put, is the observation that speakers will change their speech habits in the presence of others who are not members of their speech community (Labov, 1972).

7. Cuban and Cuban-American were counted separately in Deuchar et al. (2014) but are collapsed here.

Table 2. Sample of spreadsheet data entry

File	Line/ time	Sample	Token	Comments	Spanish translation
sastre11.mal1	1:45	and put all [the muebles]	the muebles	female speaker asks about los muebles in previous Spanish turn	los muebles
herring10.SAR	256–267	entonces todos [esos restaurants] that are participating will have booths	esos restaurants		los restaurantes
zeledon8.fem1	6:52	ahí está Sunset a Lakes, que es la escuela donde yo estaba que es [un neighborhood very upscale]	un neighborhood very upscale	very fluid CS	el barrio

CLAN transcription program (MacWhinney, 2000) was used to listen to sound files while concurrently reading the transcription. For non-transcribed files (nine sound files), the program Praat (Boersma & Weenink, 2012) was used instead.

Every instance of a Mixed NP was recorded as a unique token in a spreadsheet file (see Table 2). The following information was recorded for each token in the spreadsheet.

- File: The filename of the sound file.
- Line/Time: The line number of the corresponding transcription or the timestamp of the sound recording of the extracted sentence.
- Sample: The full sentence context.
- Token: The extracted Mixed NP.
- Comments: Extraneous comments indicating notes such as whether the token had been mentioned previously in unilingual speech, whether the token refers to a human referent, what the token refers to if the meaning was ambiguous, etc.
- Spanish Translation: Spanish translation listed with definite article.

Mixed NPs were of the form DET NP where the noun phrase could include single words, e.g., *el* [dress]$_{NP}$, or multi-word constituents, e.g., *el* [red dress]$_{NP}$. The following criteria further constrained token extraction. Mixed NPs that begin with

a Spanish determiner that did not mark for grammatical gender were excluded (e.g., *su* house 'his/her/your/their house'). Bare nouns that surfaced in the other language were also excluded. Finally, Mixed NPs that carried phonological adaption into the other language were not included (e.g., *la breca* 'the brake' cf. in non-contact varieties of Spanish, *el freno*; Clegg, 2006).

3.3 Results

Following the criteria outlined above, a total of 316 Mixed NP tokens were extracted from the corpus. The distribution of Mixed NPs is presented in Table 3. The compiled corpus is available as an online supplemental at http://ufdc.ufl.edu/IR00006198/00001.

Overwhelmingly, Mixed NPs were comprised of Spanish determiners with a following English noun (total of 96%). This distribution pattern replicates previous findings observing that Mixed NPs are more likely to include a Spanish determiner and an English noun (e.g., Herring et al., 2010; Jake et al., 2002; Pfaff, 1979). As noted by Herring et al., these infrequent English determiner Mixed NPs occur in otherwise English clauses (i.e., where English is the matrix language) and include the use of culturally-specific items such as food (examples (3) and (4)).

(3) Maybe I can take [some *agüitas*$_{fem}$] to you, and you can put it in your bag (sastre2.LUI)
'Maybe I can take [some juices/sodas] to you, and you can put it [sic] in your bag'

(4) Just hop in the car and go get [some *pastelitos*$_{masc}$] (herring6.NIC)
'Just hop in the car and go get [some cakes]'

English determiner Mixed NPs were roughly split between feminine and masculine Spanish nouns (58% feminine, n = 7; $\chi^2 = 0.08$, df = 1, p = 0.77). Of the feminine noun tokens, three were repetitions of an inanimate object (the *manguera* 'the hose'), one was a human referent (a *vieja* 'an old woman'), two were cultural borrowings related to food (some *agüitas* 'some juices/sodas', the best *harina* 'the best flour'), and one referred to a store (a *botánica* 'a natural health store'). For the

Table 3. Total distribution of Mixed NPs in Bangor Miami Corpus

Determiner	Total	Percentage
English	12	3.8%
Spanish-masculine	296	93.7%
Spanish-feminine	8	2.5%
Total	316	100%

masculine noun tokens, one token was an inanimate referent (the *muebles* 'the furniture'), one a cultural borrowing related to food (some *pastelitos* 'the cakes'), and three were human referents (a *guardia* 'a guard', a *gringo*, the *cucaracha* guy 'the cockroach guy'). Of these human referent tokens, one is an established borrowing, although the speaker had just mentioned *un americano* 'an American' in the same turn (but I saw *un americano* 'an American', I saw a *gringo* driving it [sastre12.fem1]); the other represents a creative neologism as the speaker was talking about an exterminator.[8]

Despite the frequency of Spanish determiner Mixed NPs, feminine-marked Mixed NPs were exceedingly infrequent in the corpus (n = 8). These Mixed NPs were numerically the least frequent form, even compared to English determiner Mixed NPs (2.5% vs. 3.8%); however, these proportions are not statistically different ($\chi^2 = 0.45$, df = 1, p = 0.5). Focusing the analysis on the Spanish determiner Mixed NPs, the dominant pattern is for masculine-marked Spanish determiners followed by English nouns regardless of the gender of the Spanish translation equivalent, e.g., *el* cookie 'the$_{masc}$ cookie'. Masculine-marked Mixed NPs include a higher proportion of masculine translation equivalents (n = 185, 62.5%) as compared to feminine translation equivalents (n = 103, 34.8%, $\chi^2 = 20.31$, df = 1, $p < 0.001$), with the remaining eight tokens either representing ambiguous human referents (n = 4) or nouns that include translation equivalents that can either be feminine or masculine and were not specified in prior context (n = 4). For full results of masculine-marked Mixed NPs, refer to the online supplemental.

In contrast, the feminine-marked Mixed NPs are almost unambiguously feminine Spanish translation equivalents (seven out of eight tokens), which included two female human referents (*la* assistant 'the$_{fem}$ assistant,' Sp. *la asistente/ayudante, la* cheerleader *pesada* 'the$_{fem}$ annoying cheerleader,' Sp. *la animadora pesada*), one cultural borrowing (*la nueva* Miss USA 'the$_{fem}$ new Miss USA,' Sp. *la nueva Miss EEUU*), one proper noun (*la* Notre Dame church 'the$_{fem}$ Notre Dame church,' Sp. *la catedral de Notre Dame*), one mass count noun with human referent (*la* command staff 'the$_{fem}$ command staff,' Sp. *la comandancia/el personal*), and three inanimate objects (*las tres* bee's 'the$_{fem}$ three bee's,' Sp. *las tres bes*, *las* sheets 'the$_{fem}$ sheets,' Sp. *las sábanas*, *una* cookie 'a$_{fem}$ cookie,' Sp. *una galleta*). The overwhelming preference for masculine-marked Mixed NPs further replicates previous findings, although at a much higher rate than previously reported (e.g., Jake et al., 2002; Poplack et al., 1982). The quantitative results provide support for the hypothesis that masculine is the default gender in Spanish-English bilingual

8. This token also represents a switch from English determiner to either a mixed compound noun (*cucaracha-guy*) or to a Spanish adjective and then to an English noun. In either case, if the token is removed, the results do not appreciably change.

speech. However, given that this preference is not categorical, I further explore the use of feminine-marked Mixed NPs in the Bangor Miami corpus.

3.4 Female referent Mixed NPs

Despite the low number of tokens of feminine Mixed NPs, of particular interest are the usage patterns underlying their gender assignment. One constraint that the literature has cited as strongly favoring feminine gender assignment is (semantic) animacy (Clegg, 2006; Otheguy & Lapidus, 2003; Poplack et al., 1982). However, despite the perceived strength of this constraint, Otheguy and Lapidus observe that either this constraint has shifted or never has been as strong as claimed. Yet, the innovative use of masculine Spanish determiners with female human referents is also not categorical.

The Bangor Miami corpus reveals several instances of female human referents that surface with masculine determiners.[9]

(5) *Ella es* [*un* renaissance woman]$_{NP}$ (sastre5.fem1)
'She is [a$_{masc}$ renaissance woman]'

(6) A. – she was in Platinum before @al *y ahora es* [*el* manager]$_{NP}$ *aquí* (zeledon8.fem1)
'A. – she was in Platinum [Gym] before. Now she is [the$_{masc}$ manager] here'

In (5), the referent is transparently a biological female, as evidenced by mention of the word *woman* as well as use of the pronoun, *ella* 'she.' In contrast, the Spanish translation equivalent would unambiguously surface with the feminine determiner *una* 'a$_{fem}$.' Similarly, in (6) the pronoun *she* indicates that the referent is female. Nevertheless, *manager* appears with a Spanish masculine determiner, *el* 'the$_{masc}$.'

The corpus further reveals examples that refer to humans but do not immediately indicate biological gender. For some of these cases, sex can be determined by context – either by previous mention or discourse context – and continue to indicate the use of female human referents with masculine-marked Spanish determiners.

(7) *tú eres el, tú eres* [*el* case manager]$_{NP}$ *y quiere que* [*el* case manager]$_{NP}$ *lo revise* (zeledon6.fem2)
'You're the, you're the$_{masc}$ case manager and he/she wants [that person] to go over it'

9. Proper names of individuals are obscured and represented with an arbitrary letter initial. Phonetic utterances that are not identifiable are represented with the symbol @.

Table 4. Percentage of human referents in Mixed NPs by assigned gender of determiner

Determiner	Males	Females	Total
Masculine	8 (100%)	8 (73%)	16 (84%)
Feminine	0 (0%)	3 (27%)	3 (16%)
Total	8 (42%)	11 (58%)	19 (100%)

The context leading to the utterance in (7) establishes that the referent is female. The speaker is an administrative assistant at a health services management office. She is retelling a workplace conflict that she encountered with a co-worker. She recounts passing on a message from a boss to another co-worker who was the case manager in charge of several patients' records. The speaker repeatedly refers to the co-worker by name thereby establishing unambiguous female reference.

To further explore the role of sex as a possible constraint on gender assignment in Spanish determiner Mixed NPs, the subset with identifiable human referents was analyzed. Unambiguous human reference was established either by direct semantic reference (use of a name or pronoun) or by previous discourse context. These criteria resulted in a total of 19 tokens. The simple cross-tabulation between Spanish determiner and sex of referent is presented in Table 4.

Table 4 reveals two important findings. First, human male referents are all categorically assigned masculine determiners. On the other hand, only 27% of human female referents are assigned feminine determiners. This distribution indicates that sex and animacy asymmetrically constrain the Spanish grammatical gender system in Spanish-English CS as noted by Otheguy and Lapidus (2003). Additionally, although there are only a limited number of feminine-marked Mixed NP tokens (n = 8), human referents represent approximately 1/3 of the subset. Alternatively, male human referents only represent 3% of all masculine marked tokens. The surfacing of feminine determiners in Mixed NPs are more associated with human referents even though they do not represent the majority of feminine-marked Mixed NPs.

4. Discussion

The goal of this study is to examine the production patterns underlying Spanish-English CS in order to make informed predictions for psycholinguistic models of comprehension, such as the Production-Distribution-Comprehension model (e.g., Gennari & MacDonald, 2009). The syntactic structure of focus is the Mixed NP construction due to its status as a thoroughly investigated form in Spanish-English CS (e.g., Jake et al., 2002; Otheguy & Lapidus, 2003; Pfaff, 1979; Poplack,

1980). The PDC model hypothesizes that production and comprehension are tightly linked such that frequent production patterns should facilitate comprehension. One method to quantify distribution patterns is to extract targeted structures from natural language corpora. To that end, this study makes use of a bilingual spoken language corpus collected in Miami, FL.

The results unambiguously point towards a strong preference for the use of masculine gender in the production of Mixed NPs in Spanish-English bilingual speech. In particular, the usage of masculine Spanish articles with English nouns is not constrained by Spanish gender assignment unlike in Spanish unilingual contexts. In contrast, Mixed NPs with feminine marked determiners do not follow the same usage pattern. Although the number of feminine tokens was small, around 3% of the entire corpus, these infrequent Mixed NPs were overwhelmingly restricted to English nouns with feminine Spanish translation equivalents, e.g., *la* cookie 'the$_{fem}$ cookie,' Sp. *la*$_{fem}$ *galleta*$_{fem}$. This immensely asymmetric distribution favors a default gender assignment strategy in bilingual speech production (Jake et al., 2002).

In terms of exposure-based processing models (e.g., PDC), the predictions for the comprehension of Mixed NPs are now clear. If bilinguals generate expectations for when a CS may occur, then masculine-marked Spanish determiners should serve as a useful cue for signaling an impending codeswitch. This ability to signal potential codeswitches, which may or may not follow the grammatical gender constraints of Spanish, has broader implications for the use of grammatical gender in online sentence processing. Previous studies on Spanish monolingual speakers have consistently shown a grammatical gender effect such that a match between gender-marking functional items, e.g., determiners, and nouns show facilitation in processing in informative contexts (e.g., when presented different gender items). In a somewhat counterintuitive fashion, then, the prediction for CS bilinguals is that they are less likely to use masculine gender as marked on determiners to facilitate noun identification, at least when processing codeswitched speech. That is, bilinguals should not show faster processing towards target masculine nouns because in essence, a masculine determiner could be followed by any noun regardless of gender in CS, i.e., both *el* car 'the$_{masc}$ car' and *el* cookie 'the$_{masc}$ cookie' are acceptable. Alternatively, bilinguals should not expect a codeswitch after a feminine determiner as these constructions are exceedingly rare. Even in the case of a feminine-marked codeswitch, they should have strong biases to expect nouns that are feminine in Spanish.

In light of the overwhelming gender asymmetry in Mixed NPs, the question arises as to why feminine-marked Mixed NPs should appear in production. Although other researchers have noted this asymmetry (Jake et al., 2002; Otheguy & Lapidus, 2003, Pfaff, 1979), there is considerable debate as to how gender is

assigned in Mixed NP constructions. Researchers generally favor phonological and/or semantic constraints as driving gender assignment (see Clegg, 2006; Poplack et al., 1982). As the results in Table 4 show, biological gender only applies to a subset of the data reported here and does not strongly constrain gender assignment for female human referents (see also Otheguy & Lapidus, 2003; Poplack et al., 1982, cf. Clegg, 2006).

The phonological accounts stem from the correlation between Spanish nouns endings and gender. Bull (1965) observed that the most prevalent word-final phonemes that correlated with masculine are /e/, /n/, /o/, /r/, and /s/. For feminine gender, Bull listed /a/, /d/, /ión/, and /is/ as being the most reliable word-final endings. If a phonological constraint is operant in gender assignment in Mixed NPs, then English words that fall into these phonological categories should also apply this phonologically-driven gender assignment rule, e.g. English words ending in *-a* should be assigned feminine gender.

Clegg (2006), following a variationist framework, concluded that the results of his study provided strong support for operant phonological constraints; however, only 453 (around 50%) of the 899 English words examined by Clegg matched Bull's phonological class endings for Spanish. Because half of the data set did not fall under Bull's classification, Clegg claimed that these atypical phoneme endings should be assigned masculine gender, following Spanish phonology. This assumption may have unnecessarily inflated the success of the phonological account by applying a circular logic to final-word ending classification (see also Poplack et al., 1982 for a similar logic). The patterns are more likely a reflection that the Spanish and English phonological systems do not overlap. Additionally, Clegg's study includes a broader scope of English-origin nouns, such as phonologically adapted nouns, e.g. *troca* from Eng. truck, *breca* from Eng. brake.

The results reported here do not lend strong support for phonologically-driven gender assignment in Mixed NPs. Of the feminine-marked tokens, none fall under Bull's classification for feminine gender. For masculine-marked tokens, 36 tokens (12%) follow Bull's classification, with the most frequent word-final phoneme being /r/. In addition, several tokens (n = 6) occur in which the cognate morpheme *-ion* does not automatically result in a correspondent feminine-marked Mixed NP, e.g. *estos* conversat[ion]s, Sp. *estas*$_{\text{fem}}$ *conversac[ion]es*$_{\text{fem}}$ 'these conversations'.

Consequently, the results reported here support the hypothesis that masculine is the default gender in Spanish-English Mixed NPs in tandem with Jake et al. (2002). This hypothesis, however, is an unsatisfactory account for why and how a small number of feminine-marked Mixed NPs arise in the corpus. Instead, they lead to further questions: (1) Under what circumstances do feminine-marked Mixed NPs surface? (2) How does a theory that specifies the use of a default

gender in Mixed NPs also account for the production of feminine-marked Mixed NPs? The limited number of feminine-marked Mixed NPs highlights that these forms are a highly restricted set. To resolve the apparent discrepancy between the overwhelming use of a default gender strategy with the appearance of feminine-marked Mixed NPs, I suggest that CS is an emergent linguistic system built upon the bilingual's constituent languages. Under this hypothesis, CS bilingual have learned a set of distributional patterns for CS that are different from their unilingual modes and form the basis for planning upcoming codeswitches in discourse. For Mixed NPs, a planned codeswitched utterance involves the adoption of a default gender strategy and hence forms that are marked with masculine determiners. Consequently, feminine-marked Mixed NPs are not planned codeswitched utterances. Instead, they are exceptional switches that occur on-the-fly in speech planning.

4.1 Feminine-marked Mixed NPs

The highly infrequent use of the feminine determiner in Mixed NPs suggests that these tokens may be the exception to regular production patterns in CS. The number of feminine marked tokens in the corpus is too few to allow for a deeper investigation of this hypothesis here. However, two pieces of evidence offer tentative support to the idea that feminine-marked Mixed NPs are less planned than masculine-marked Mixed NPs and may guide future research. First, feminine-marked Mixed NPs as exceptions should singly be embedded in otherwise unilingual discourse. Under this hypothesis feminine-marked forms are more likely to appear as singleton switches in larger stretches of Spanish as illustrated in (8)–(10).

(8) *sí pero fíjate que a ti todavía no te han puesto* [*la* assistant]$_{NP}$ *ahí a trabajar* (sastre3.fem1)
'Yeah, but look, they still haven't put in place [the$_{fem}$ assistant] to work for you'

(9) *tú sabes que nosotros no # no vamos a ir a tu casa a recoger* [*las* sheets]$_{NP}$ *y hay que lavar todo eso* (herring16.MAL)
'You know that we are not going to go to your house to pick up [the$_{fem}$ sheets] and all of that [stuff] has to be washed'

(10) *¿Qué es lo que ella quiere, qué es* [*una* cookie]$_{NP}$? (sastre4.fem1)
'What does she want, what is [a$_{fem}$ cookie]?'

Conversely, masculine-marked determiners are more likely to appear in alternational type switches or in longer stretches of discourse in which elements in both languages are apparent.

(11) *pero no tenían* [*el* flag]$_{NP}$ out there? (sastre9.fem2)
'but didn't they have [the$_{masc}$ flag] out there?'

(12) *entonces* [*todos esos* restaurants]$_{NP}$ that are participating will have booths (herring10.SAR)
'Then all of [those$_{masc}$ restaurants] that are participating will have booths'

In this study, only one feminine-marked token (12.5%) is not a singleton switch (Table 1). In contrast, 67 masculine-marked Mixed NPs (22.9%) appeared in non-singleton switches. Clearly, singleton switches still dominate bilingual speech as reported elsewhere (e.g. Poplack, 1980); however, the hypothesis here is that feminine-marked mixed NPs should be highly restricted to singleton switches.

Second, if feminine-marked Mixed NPs are indeed less planned switches, speakers should produce more disfluencies, repetitions, and pauses leading up to CS. Similarly, disfluencies indicating apparent failure to retrieve feminine referents, which result in recasts with masculine-marked Mixed NPs should also surface. Several examples from the corpus offer support for these predictions. The first example (13) demonstrates a speaker with a high number of disfluencies (as represented by frequent repetitions) and a reformulation of the target NP, *la pesada* 'the annoying [one],' with a feminine-marked Mixed NP, *la* cheerleader *pesada* 'the annoying cheerleader'.[10] The remaining examples (14)–(17) highlight failure to retrieve the intended Spanish feminine noun, with a subsequent switch to a masculine-marked Mixed NP, (e.g., *una, un* hammock in (14)).[11]

(13) *no no no no hay un hay una parte que que la la la pesada* [*la* cheerleader *pesada*] *está tomando* (herring7.SEB)
'No, no, no, no … there's a, there's a part [of a performance] that that the$_{fem}$ annoying, [the$_{fem}$ annoying cheerleader] is taking'

(14) *y eso que no la has puesto en una # cómo se llama esto una* [*un* hammock] *de esos* (sastre1.SOF)
'And you didn't even put it on a… what are those things called, one of those, a$_{fem}$, [a$_{masc}$ hammock]'

10. An anonymous reviewer points out that the use of *la pesada* directly before ***la*** cheerleader ***pesada*** may represent an anticipation of a word order conflict instead of a disfluency. This alternative is possible, although the recording reveals several disfluencies leading up to the codeswitch.

11. These examples were not counted as feminine-marked Mixed NPs in the corpus; rather, they are used here to illustrate and support the hypothesis that feminine-marked Mixed NPs are a result of unplanned codeswitches.

(15) *Diciendo de que el* sales office *# le había dicho que cruzara la calle que nosotros íbamos a hacer la ##* [*el*$_{c}$ orientation]$_{NP}$ (herring3.ASH)
'… saying that the sales office had told him/her to cross the street, that we were going to do the$_{fem}$, [the$_{masc}$ orientation]'

(16) *Cambiar las paredes, quitar las paredes, poner las* [*los* dry walls]$_{NP}$, *esos nuevos* (sastre1.KEV)
'To change the walls, remove the walls, place the$_{fem}$, [the$_{masc}$ dry walls]. Those new ones'

(17) *La la la la* [*el* strawberry] *echa una rama* (sastre2.AVA)
'The$_{fem}$, the$_{fem}$, the$_{fem}$, the$_{fem}$, the$_{masc}$ strawberry puts out a stem…'

These examples underscore the exceptional status of the feminine-marked Mixed NP. Example (14) is particularly striking as the speaker explicitly indicates difficulty retrieving the Spanish word that nevertheless elicits a congruent feminine determiner. Failure to retrieve the intended Spanish noun results in a subsequent switch to English with a concurrent switch to using default gender. This same speaker later makes reference to *una hamaca* 'a$_{fem}$ hammock$_{fem}$,' having remembered the prior Spanish word that she was trying to access. This account does not exclusively suggest that masculine-marked Mixed NPs are void of disfluencies; rather, as in the case of singleton switching, disfluencies and frequent repetitions or recasts should occur more frequently with feminine-marked Mixed NPs.

5. Conclusions

Despite the overwhelming distributional asymmetry in gender-marked Mixed NPs, bilinguals must learn this gender asymmetry in order to successfully comprehend bilingual speech. Bilinguals should anticipate an increased likelihood for CS to happen after a masculine determiner in appropriate contexts, but they must also be prepared for the more infrequent occasions in which CS follows a feminine determiner. Alternatively, CS bilinguals must learn that feminine determiners will not be followed by masculine translation equivalents whereas masculine determiners may be followed by either masculine or feminine translation equivalents. Thus, these bilinguals ultimately must learn a hybridized pattern for gender assignment of Mixed NPs in order to successfully comprehend codeswitched speech. This pattern for gender assignment is noticeably different from gender assignment in Spanish. If CS is an emergent and learned system, then this gender asymmetry, as well as any other asymmetric distributions specific to CS, must be learned amongst a community of codeswitchers. Consequently, whether a bilingual has immersed herself in such a community, i.e., the bilingual profile in

terms of usage and exposure to CS, should result in observable group differences in the production and comprehension of CS. Similarly, community differences in the preferred pattern of use in CS can and should arise because the specific structure that a community adopts may be influenced by a host of linguistic and extra-linguistic variables. Indeed, recent research indicates that Spanish-Basque CS has settled on the use of feminine-marked Mixed NPs as the dominant pattern (Parafita Couto, Munarriz, Epelde, Deuchar, & Oyharçabal, 2015). Subsequently, the predictions for online processing of gender should be different between Spanish-Basque and Spanish-English bilinguals.

Although CS as an emergent system receives scant support in the literature (cf. Gardner-Chloros, 2009), emergent approaches offer an alternative as to how to account for asymmetrical structural distributions. In this chapter, I have highlighted one such asymmetric distribution with grammatical gender in Mixed NPs. Additionally, other structures prevalent in CS such as the language of the complementizer (that, Sp. *que*) should also result in asymmetric production distributions and thus asymmetric comprehension patterns. Currently, most theoretical approaches attempt to fit CS phenomena into parsimonious accounts of permissible syntactic switch sites without addressing asymmetric distributions, e.g., on the status of whether codeswitches can happen between a determiner and a noun phrase (e.g., Belazi, Rubin, & Toribio, 1994; Di Sciullo, Muysken, & Singh, 1986), or address asymmetries in the overall contribution of each language to codeswitched speech, (Myers-Scotton, 1993, 2000). Here, I suggest that an emergentist approach further anticipates asymmetric differences in usage as not only possible but highly likely.

In terms of psycholinguistic models of sentence processing, CS is a powerful tool that can be used to elucidate the role of experience in guiding comprehension processes (e.g., PDC framework, Gennari & MacDonald, 2009). In particular, focusing on codeswitches that bring together cross-linguistic differences, such as the presence or absence of grammatical gender in Spanish-English bilingual speech, can reveal how distribution patterns in production impact comprehension, ultimately revealing the dynamic nature of language learning and use.

References

Belazi, H. M., Rubin, E. J., & Toribio, A. J. (1994). Code switching and X-bar theory: The functional head constraint. *Linguistic Inquiry*, 25(2), 221–237.

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer. Retrieved from <http://www.fon.hum.uva.nl/praat/>

Bull, W. E. (1965). *Spanish for teachers: Applied linguistics*. New York, NY: Ronald Press.

Clegg, J. H. (2006). Lone English-origin nouns in the Spanish of New Mexico: A variationist analysis of phonological and morphological adaptations. (Unpublished PhD dissertation). University of New Mexico, Albuquerque, NM.

Deuchar, M., Davies, P., Herring, J. R., Parafita Couto, M. C., & Carter, D. (2014). Building bilingual corpora: Welsh-English, Spanish-English and Spanish-Welsh. In E. M. Thomas & I. Mennen (Eds.), *Advances in the study of bilingualism* (pp. 93–110). Bristol: Multilingual Matters.

Di Sciullo, A., Muysken, P., & Singh, R. (1986). Government and code-mixing. *Journal of Linguistics*, 22, 1–24. doi:10.1017/S0022226700010537

Dussias, P. E., & Cramer Scaltz, T. R. (2008). Spanish-English L2 speaker's use of subcategorization bias information in the resolution of temporary ambiguity during second language reading. *Acta Psychologica*, 128(3), 50–113. doi:10.1016/j.actpsy.2007.09.004

Gardner-Chloros, P. (2009). *Code-switching*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511609787

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, 93(37), 58–93. doi:10.1006/jmla.1997.2512

Gennari, S. P., & MacDonald, M. C. (2009). Linking production and comprehension processes: The case of relative clauses. *Cognition*, 111(1), 1–23. doi:10.1016/j.cognition.2008.12.006.

Green, D. W. (2011). Language control in different contexts: The behavioral ecology of bilingual speakers. *Frontiers in Psychology*, 2, 1–4. doi:10.3389/fpsyg.2011.00103.

Herring, J. R., Deuchar, M., Parafita Couto, M. C., & Moro Quintanilla, M. (2010). 'I saw the *madre*': Evaluating predictions about codeswitched determiner-noun sequences using Spanish-English and Welsh-English data. *International Journal of Bilingual Education and Bilingualism*, 13(5), 553–573. doi:10.1080/13670050.2010.488286

Jake, J. L., Myers-Scotton, C., & Gross, S. (2002). Making a minimalist approach to code-switching work: Adding the Matrix Language. *Bilingualism: Language and Cognition*, 5(1), 69–91. doi:10.1017/S1366728902000147

Kroll, J. F., Sumutka, B. M., & Schwartz, A. I. (2005). A cognitive view of the bilingual lexicon: Reading and speaking words in two languages. *International Journal of Bilingualism*, 9(1), 27–48. doi:10.1177/13670069050090010301.

Labov, W. (1972). *Sociolinguistic patterns*. Oxford: Blackwell.

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In B. MacWhinney (Ed.), *The emergence of language*. Mahweh, NJ: Lawrence Erlbaum Associates.

MacDonald, M. C., & Thornton, R. (2009). When language comprehension reflects production constraints: Resolving ambiguities with the help of past experience. *Memory & Cognition*, 37(8), 1177–1186. doi:10.3758/MC.37.8.1177.

MacSwan, J. (1999). *A minimalist approach to intrasentential code switching*. New York, NY: Garland.

MacSwan, J. (2000). The architecture of the bilingual language faculty: Evidence from intrasentential code switching. *Bilingualism: Language and Cognition*, 3(1), 37–54. doi:10.1017/S1366728900000122

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed). Mahwah, NJ: Lawrence Erlbaum Associates.

Milroy, L. (1987). *Observing and analysing natural language: A critical account of sociolinguistic method*. New York, NY: Blackwell.

Myers-Scotton, C. M. (1993). *Duelling languages: Grammatical structure in codeswitching*. Oxford: Oxford University Press.

Myers-Scotton, C. M. (2000). The Matrix Language Frame model: Development and responses. In R. Jacobson (Ed.), *Codeswitching worldwide* (Vol. 2, pp. 23–58). Berlin: Mouton de Gruyter.

Myers-Scotton, C. M., & Jake, J. (2015). Cross-language asymmetries in code-switching patterns: Implications for bilingual language production. In J. Schwieter (Ed.), *The Cambridge handbook of bilingual processing* (pp. 416–458). Cambridge: Cambridge University Press. doi: 10.1017/CBO9781107447257.019

Otheguy, R., & Lapidus, N. (2003). An adaptive approach to noun gender in New York Contact Spanish. In R. Cameron, L. López & R. Núñez Cedeño (Eds.), *A romance perspective on language knowledge and use*. Amsterdam: John Benjamins. doi: 10.1075/cilt.238.17oth.

Parafita Couto, M. C., Munarriz, A., Epelde, I., Deuchar, M., & Oyharçabal, B. (2015). Gender conflict resolution in Spanish-Basque Mixed DPs. *Bilingualism: Language and Cognition*, 18(2), 304–323. doi: 10.1017/S136672891400011X.

Pfaff, C. (1979). Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English. *Language*, 55(2), 291–318. doi: 10.2307/412586

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: Towards a typology of code-switching. *Linguistics*, 18(7/8), 581–618. doi: 10.1515/ling.1980.18.7-8.581

Poplack, S., Pousada, A., & Sankoff, D. (1982). Competing influences on gender assignment: Variable process, stable outcome. *Lingua*, 57, 1–28. doi: 10.1016/0024-3841(82)90068-7

Soares, C., & Grosjean, F. (1984). Bilinguals in a monolingual and a bilingual speech mode: The effect on lexical access. *Memory and Cognition*, 12(4), 380–386. doi: 10.3758/BF03198298

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33, 285–318. doi: 10.1006/jmla.1994.1014