Testing alternative theoretical accounts of code-switching: Insights from comparative judgments of adjective-noun order

(In press, *International Journal of Bilingualism*)

Hans Stadthagen-González*
(Corresponding Author)
Department of Psychology
University of Southern Mississippi
Hardy Hall 314
730 East Beach Blvd.
Long Beach, MS 39560
USA
h.stadthagen@usm.edu

M. Carmen Parafita Couto
Leiden University Center for Linguistics
Witte Singel complex
Van Wijkplaats 3, 005B
2311 BX Leiden
Netherlands
m.parafita.couto@hum.leidenuniv.nl

C. Alejandro Párraga
Computer Vision Center,
Computer Science Department,
Universitat Autónoma de Barcelona,
Edifici 'O', Carrer de Les Sitges
Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallès) · Barcelona · Spain
Alejandro.Parraga@cvc.uab.es

Markus F. Damian
School of Experimental Psychology
University of Bristol
The Priory Road Complex,
Priory Road, Clifton BS8 1TU
UK
M.Damian@bristol.ac.uk

Abstract

Objectives: Spanish and English contrast in adjective-noun word order: e.g. brown dress (English) vs. *vestido marrón* ('dress brown', Spanish). According to the Matrix Language model (Myers-Scotton, 2002; MLF) word order in code-switched sentences must be compatible with the word order of the matrix language, but working within the Minimalist program, Cantone and MacSwan (2009; MP) arrived at the descriptive generalization that the position of the Noun Phrase (NP) relative to the adjective is determined by the adjective's language. Our aim is to evaluate the predictions derived from these two models regarding adjective-noun order in Spanish/English code-switched sentences.

Methodology: We contrasted the predictions from both models regarding the acceptability of code-switched sentences with different adjective-noun orders that were compatible with the MP, the MLF, both, or none. Acceptability was assessed in Experiment 1 with a 5-point Likert and in Experiment 2 with a two-alternative forced choice (2AFC) task.

Data and analysis: Data from both experiments were subjected to Linear Mixed Model analyses. Results from the 2AFC task were also analyzed using Thurstone's (1927) Law of Comparative Judgment.

Conclusions: We found an additive effect in which both the language of the verb and the language of the adjective determine word order.

Originality: Both experiments examine noun-adjective word order in English/Spanish code-switched sentences. Experiment 2 represents a novel application of Thurstone's Law of Comparative Judgements to the study of linguistic acceptability which yielded clearer results than Likert scales. We found convincing evidence that neither the MLF nor the MP can fully account for the acceptability of noun-adjective switches.

Implications: We suggest that advances in our understanding of grammaticality in code-switching will be achieved by combining the insights of the two frameworks instead of considering them in isolation (Eppler et al. 2016), or by espousing a probabilistic model of code switching (e.g. Bresnan, 2007).

KEYWORDS: code-switching, adjective-noun order, matrix-language frame, minimalist program, two-alternative forced choice, Thurstone's Law.

Testing alternative theoretical accounts of code-switching: Insights from comparative judgments of adjective-noun order

Code-switching is the phenomenon by which bilinguals go back and forth between their two languages in the same conversation (for an overview see Deuchar, 2012). It is generally accepted that switches between languages do not occur at random, but follow specific patterns and rules (Bullock & Toribio, 2009; Guzzardo Tamargo, Mazak & Parafita Couto, 2016, i.a.), and the typifying of those rules represents a rich and active field of research in the bilingualism literature. However, scholars do not agree on the best theoretical account of these regularities. Recently, an interest has emerged in evaluating the predictions of theoretical models to try to disentangle between theoretical predictions (see Blokzijl, Deuchar, & Parafita Couto, submitted; Eppler, Luescher & Deuchar, 2016; Fairchild & Van Hell, 2015; Herring, Deuchar, Parafita Couto & Moro Quintanilla, 2010; Parafita Couto, Boutonnet, Hoshino, Davies, Deuchar & Thierry, 2017; Parafita Couto, Deuchar & Fusser, 2015; Parafita Couto & Gullberg, submitted; vanden Wyngaerd, 2016). These researchers have shown that different theories account for some aspects of the observed data, but there is no overarching theory that can explain all the code-switching patterns.

In what follows we will experimentally examine the accuracy of predictions regarding adjective-noun order in code-switching derived from two particular models of code-switching: Cantone & MacSwan (2009) within the Minimalist Program and the Matrix Language Framework (Myers-Scotton, 2002). Of particular interest are the so-called conflict sites (Poplack & Meechan, 1998), instances where the grammars of the two languages differ. For example, the default order for noun-adjective constructions in English is adjective-noun as in ‘red book’, whereas the default order in Spanish is the opposite (*libro rojo* – literally ‘book red’). In Spanish, some adjectives may appear before or after the noun with a change in meaning. Post-nominal adjectives have a basic, attributive meaning (1a), while prenominal adjectives have a more restricted, intrinsic meaning (1b).

(1) a. *un hombre grande* b. *un gran hombre*

'a tall man' 'a great man'

However, in this paper, we will focus on the regular cases with post-nominal adjectives.

Conflict sites offer the possibility of directly contrasting the predictions of each of the two models under study. In the context of Spanish-English code-switching, we may consider whether ***libro*** *red*, *red* **libro**, ***rojo*** *book*, and *book* ***rojo*** are equally probable. It is unclear how a conflict is resolved (i.e., what constrains word order) when a Spanish noun is combined with an English adjective, or *vice versa*. Some comparative analyses addressing the question of adjective-noun order in code-switched noun phrases have been done couched within the two theoretical models (MP and MLF) but with different outcomes (cf.; Parafita Couto et al. 2015, 2017; Parafita Couto & Gullberg, submitted; vanden Wyngaerd, 2016).

Of the two models in question, Myers-Scotton's (2002) MLF differentiates the languages involved: one language is known as the matrix language (ML), the other as the embedded language (EL). In a code-switched clause, the ML is assumed to provide the morphosyntactic frame and allows predictions regarding acceptable constructions. On the other hand, generativist MacSwan (2005a, 2005b) argues that all instances of code-switching may be accounted for based on general mechanisms of grammar (for further details see MacSwan, 2005a, 2005b; Jake, Myers-Scotton & Gross, 2005).

The MLF predicts that (i) 'late outsider morphemes' such as finite verb morphology marking subject-verb agreement and (ii) word order within a clause that contains code-switching will be sourced from the ML. Previous studies of Welsh-English code-switching show that Welsh-English bilinguals tend to produce clauses with code-switching in which the language of the finite verb morphology matches clause word order (Davies, 2010; Davies & Deuchar, 2010; Deuchar & Davies, 2009; Deuchar, 2006; Herring et al. 2010; Parafita Couto et al., 2015). Myers Scotton's model would predict that if the bound morphology of the finite verb is from language A, adjective and noun word order in a code-switched noun phrase should also be dictated by language A.

Cantone and MacSwan (2009) in their analysis of noun-adjective order in Italian-German code-switching follow Cinque's (1994) proposal that a universal base underlies adjective position, with adjectives assumed to universally precede the noun. On this view, differences in word order between English and Spanish follow from overt movement of the noun in Spanish to a position to the left of the adjective, resulting in the contrasting surface word order outlined above, with adjectives preceding the noun in English and following it in Spanish. Cantone and MacSwan (2009) reach the following descriptive generalization based on survey and naturalistic data from Italian-German code-switching, "while the data remain slightly ambiguous, a relatively clear pattern has emerged in both the survey data and the naturalistic data confirming the general view of previous researchers, namely, that the word order requirements of the language of the adjective determine word order in code-switching in DP-internal contexts" (2009:266-267). We can test this descriptive generalization, which in the context of Spanish-English code-switching, would mean that English adjectives should precede Spanish nouns whereas Spanish adjectives should follow English nouns. Of course, other analyses within the Minimalist approach may make different predictions, but since Cantone and MacSwan's descriptive generalization is specific to code-switching within adjectives and nouns we have chosen to test it.

There have been attempts to evaluate these two approaches at different grammatical switching points, but they have been conducted in different language pairs and using different methodologies. Also, some were based on production data while others were based on comprehension data. Hence, it is perhaps not surprising that they have also yielded conflicting results. For example, Herring, et al. (2010) examined the predictions of both theories regarding determiner-noun switches from naturalistic Spanish-English and Welsh-English code-switching data. They do not find statistically significant differences between the accuracy of the predictions of the two theories. Moving on from naturalistic data, Fairchild and Van Hell (2015) experimentally examined (through a series of picture naming tasks) the ability of the MP and the MLF to explain determiner-noun switches in Spanish-English bilinguals. However, they also did not find support for either theory. Focusing specifically on adjective-noun order in code-switching, Parafita Couto et al. (2015) designed a study to evaluate the predictions of these two models within Welsh-English mixed nominal constructions (between the adjective and the noun) by using a multi-task approach comprising (1) naturalistic corpus data, (2) an elicitation task, and (3) an auditory judgment task. They found that Likert-style judgment tasks are not very useful in this community of code-switchers, due to the stigmatized nature of the phenomenon. However, the data from the naturalistic corpus and the elicitation task supported the relative superiority of the MLF model. Nevertheless, it was only a small proportion of their data that could distinguish between the two models, so no definite conclusion could be reached. Parafita Couto et al. (2017) conducted a follow-up study using electrophysiology. Their results showed that it was the MLF that appeared to provide a better account of the linguistic mechanism involved, however, there was some ambiguity revealed in the control contrasts, where no differences were found in the two control conditions. On the other hand, vanden Wyngaerd (2016), basing her analysis on elicited data and judgment tasks, examined word order in French-Dutch mixed nominal constructions, finding support for Cantone and MacSwan's generalization. And yet an analysis of spontaneous Papiamento-Dutch code-switching production (Parafita Couto & Gullberg, submitted) could not distinguish between the predictions of both models. Pablos, et al. (submitted) evaluated these predictions using event-related brain potentials (ERPs) to measure online comprehension of code-switched utterances and following the design used in Parafita Couto et al. 2017. Their results seem to point to code-switching not being restricted at modification sites (DiSciullo, 2015).

As described above, the formulation of these theoretical models that attempt to describe and predict code-switching behavior in bilinguals has mainly relied on naturalistic corpora analysis and Likert scale ratings. Multiple studies (Herring et al. 2010, Parafita Couto, Deuchar & Fusser, 2015; Parafita Couto & Gullberg, submitted; Eppler et al., 2016) have cited corpus data with no definitive conclusion with regards to constraints on word order within code-switched noun phrases, either between the determiner and the noun or between the noun and the adjective. This situation highlights a limitation of corpus data: despite its descriptive richness and ecological validity, it is not probative in nature. Although corpora are very useful in generating hypotheses and can also be used to falsify them if counterexamples are available, data derived from corpora cannot, on their own, be used to prove predictions because no corpus is exhaustive (i.e. there is always the possibility that counterexamples exist beyond the collected sample). Other studies have used acceptability judgments (for a review of this technique, see Schütze, 1996), where informants provide a yes or no answer as to whether they accept or reject a given sentence as "correct" or "acceptable", or rate how well the sentence "sounds" on a given scale (e.g.: “On an ascending scale of 1 to 7, how acceptable do you find the following sentence?”). As pointed out by Cowart (1996), this technique is very similar to introspective judgments that are routinely used in the field of psychophysics and signal detection theory to determine the limits of our sensory system (e.g.: how dim a light our eyes can sense). While acceptability judgments afford more control, and, potentially, more probative value, than corpus data, they are vulnerable to different types of response biases that may obscure their results (see Parafita Couto et al. 2015 for a discussion of how the stigma associated with code-switching may affect grammaticality judgments).

From the signal detection theory literature (e.g.: Gescheider, 1997; Green & Swets, 1988) we can learn about some of the vulnerabilities of yes/no and Likert-style judgments. Yes/no judgments can suffer from criterion effects, where factors external to the variable under study may affect the *willingness* of the informant to *report* a “yes” or a “no” answer (for a more detailed discussion of detection thresholds, see Gescheider, 1997). Scaled responses present a different set of challenges, particularly regarding consistency of use of the scale. In order to position all items along the same scale, informants must not only “calibrate” their ratings along the entire range, but also keep a memory record of the ratings on earlier stimuli in order to avoid possible shifts in the internal rating scale when new items are presented (Parraga, 2015). Both yes/no and Likert-style acceptability judgments have attracted criticism because of lack inter-rater reliability (Labov, 1972, 1975; Ross, 1979; Stokes, 1974; Bader & Häussler, 2010) and stability within the same informants at different times or under different testing conditions (e.g.: Carol, Bever & Pollack, 1981; Nagata, 1988; Snow & Meijer, 1977).

From very early on, the psychophysics literature (e.g.: Fechner, 1876) identified and addressed these problems concerning introspective judgments by using *paired comparison* *scales,* also known as the 2-Alternative Forced Choice (2AFC) method. This is still now preferred for quantifying detecting fine-grained differences between value judgments that can only be made based on subjective criteria (David, 1988; Parraga, 2015).

The Two-Alternative Forced Choice Task and the Law of Comparative Judgment.

In a 2AFC task, participants are presented with pairs of stimuli and must choose which one of the two items is “better” according to a specified criterion. The impossibility of choosing a “tie” is intrinsic to the method of two-alternative forced choice (hence the “forced” in its name) and it is at the base of Thurstone’s theoretical approach to scaling. The probability of two different stimuli having exactly the same value on the judgment scale is considered negligible, and thus no “tie” is allowed when making the pairwise judgment. Also, if indeed there was no difference, the probability of a particular judge to pick one option over the other would be 50%, so any differences caused by the forced choice on a particular data point would even out across several judges and instances of the comparison (David, 1988). In our case, the criterion for judgment is met by deciding which of the sentences in a pair is more similar to the way the informant would speak. Besides greatly reducing the response bias effects mentioned above (Green & Swets, 1966), comparative judgments are a more natural task for participants than rating scales. Nunnally (1976, p. 40) states that “People simply are not accustomed to making absolute judgments in daily life, since most judgments are inherently comparative [...] people are notoriously inaccurate when judging the absolute magnitude of stimuli.” Paired comparisons also greatly reduce the demands on memory required for avoiding shifts in the internal rating scale (Parraga, 2015). One criticism leveled at traditional acceptability judgments such as Likert scales or yes/no judgments (e.g.: Branigan & Pickering, 2016), is that participants may not share the same definition of "grammatical" or "acceptable" among themselves or with the researcher. Branigan and Pickering also point out that traditional grammaticality judgments suffer from "source ambiguity", that is, that their outcomes may be the result of artifacts, hence they don't tap into an internal grammar, and they do not provide direct evidence about structure. These weaknesses stem from the fact that such judgments are based on an absolute, external concept that is used as a criterion of membership to a category (i.e.: grammatical, accepted, etc.) By contrast, in 2AFC tasks participants are asked to make a simple, direct comparison between two options based on their own subjective preference, minimizing the influence of factors external to the two sentences being compared. For example, as mentioned before, in many communities CS carries a stigma that could affect acceptability judgment tasks and lead informants to reject sentences that their linguistic systems would in fact generate (cf. Munarriz & Parafita Couto, 2014; Parafita Couto, Deuchar & Fusser, 2015; Anderson, 2006; Giancaspro, 2013). This would be reflected in depressed scores, and possibly even floor effects, for code-switched sentences. The 2AFC task allows us to circumvent this problem because participants compare one code-switched sentence against another, they are not asked to compare a code-switched sentence against an ideal grammatical value.

Gustav Fechner (1876), the pioneer of psychophysics, proposed that the proportion of times an item is chosen over another in a series of pairwise comparisons provides a measure of the distance between the two items in some pleasantness continuum. For example, if item A is chosen over item B half the time, both objects are equally pleasant. Correspondingly, if object C is consistently chosen over item A, then by the same measure, object C is likely to be the most pleasant item of the three. Thurstone (1927), with his *law of comparative judgment*, generalized this concept into a measurement model which converts simple pairwise comparisons between stimuli into one-dimensional quality scores (for further explanation see Bock and Jones, 1968 and Torgerson,1958). Thurstone proposed that the proportion of times a stimulus is judged as "better" than another is related to the number of psychological scale units separating the two sensations. In this method, which is the standard scaling method in many disciplines today, pairwise judgments are performed many times by a comparatively large number of subjects, resulting in statistically robust results. The outcome of the analysis is a ranking of preference for stimuli along an interval scale that reflects the relative distance between conditions. Moreover, the results from this analysis can be tested for significance using standard, parametric statistical analyses.

Despite the multiple benefits associated with the 2AFC method attested by its widespread use in psychophysics and other areas of research, its use for linguistic acceptability judgments has been conspicuously absent so far. We decided to use a traditional Likert scale task (Experiment 1) as a baseline for comparison to illustrate the benefits of a novel application of the 2AFC task and Thurstone's (1927) Law of Comparative Judgments (Experiment 2), applying both methods to test the predictions on adjective-noun order in code-switching derived from the MLF and the MP. In particular, each of these models (Myers-Scotton, 2002; Cantone & MacSwan, 2009) makes specific, testable predictions regarding the word order of adjectives and nouns within code-switched noun phrases depending on the language of each part of speech. We directly contrasted sentences that reflected (or not) the structures derived from such predictions. This contrast generated four conditions: sentences that followed the CS pattern predicted by either the MP (but not the MLF), or the MLF (but not the MP), both, or neither. Table 1 presents all possible combinations of noun- and adjective order and language (for each direction of switch), as well as the predictions from each model as to whether each combination should be acceptable to bilinguals or not.

[Table 1 about here]

If code-switching is indeed constrained by the grammatical properties proposed by one model or the other, the presence of such property should predict the acceptability of a given sentence. For example, if the matrix language determines word order within the noun phrase (as proposed by the MLF model), sentences 1, 3, 5, and 7 should show higher acceptability than the others; if, on the other hand, word order is defined by the language of the adjective itself (as according to Cantone and MacSwan's generalization), sentences 2, 4, 6, and 8 would have higher scores. Furthermore, if code-switching acceptability is *entirely* predicted by one of the models, there should be no difference between the preferred construction (whichever one may be) and those sentences "acceptable" according to both models (exemplars 1 and 5); conversely, the least favored construction according to that model should not show an advantage over sentences rejected by both models (exemplars 4 and 8). In Experiment 1 we asked participants to express their preferences by means of Likert scale ratings.

Experiment 1

Participants

A total of 40 early English/Spanish bilinguals took part in this experiment. Most were born in the United States (N=34) or moved there before the age of 5 (N=6; mean age when they migrated

=3:6 years). Self-reported age of acquisition for Spanish (Mean: 2.05, S.D.: 1.28) was generally earlier than for English (Mean: 2.53; S.D.: 1.58), but all participants were able to speak both languages by the time they started elementary school. There were 24 men and 16 women and their mean age was 30:3 years (S.D.: 9.6; Range: 18-56). All participants were second-generation bilinguals (at least one of their parents was born in Mexico) and all stated that they spoke the Mexican variety of Spanish. All of them resided in states with large populations of Spanish speakers, namely California (N=22), Texas (N=13), Colorado (N=3), Arizona (N=1), and New Jersey (N=1). The highest level of education for 22 participants was high school, 17 had attended at least some college, and 1 attended graduate school. When asked "When chatting with friends and family that speak both English and Spanish, do you mix Spanish and English in the same sentence?" all but 3 participants responded in the positive[1].

Participants were recruited through Amazon Mechanical Turk, an online crowdsourcing website that has been shown to be a good source of participants for collection of acceptability judgments (Gibson, Piantadosi & Fedorenko, 2011). Participants were paid a small fee for completing the study and only workers with an acceptance rate of 90% or above and at least 100 tasks completed were allowed to take part in the study (following the guidelines proposed by Peer, Vosgerau & Acquisti, 2014).

Language Proficiency and Dominance. In order to take part in the acceptability judgment task, each participant completed an English and a Spanish test to confirm their proficiency in both languages. These tests were an adaptation from the Online Placement Tests used by Oxford University's Language Centre (Oxford University Language Centre, n.d.). The tests were modified to reflect Latin American (rather than Iberian) verb conjugations and vocabulary (e.g.:

"*ustedes*" instead of "*vosotros*" for the second person plural pronoun), and geographical landmarks to reflect U.S. or Latin American locations (e.g.: New York instead of London). Only participants that attained a score of 34 (out of 50) or more were allowed to continue with the study. This range of scores is classified as "Higher proficiency" by the Oxford website. Participant's average scores were about the same in both languages (English: M =44.9; S.D. =2.8; Spanish: M =43.9; S.D. =3.9; t(39) =2.02, *p* = .2). All participants stated that they were able to speak Spanish before age four and English by the time they entered elementary school. They declared to be slightly more proficient in English than Spanish, with ratings of 4.00 (S.D.: 0.00) vs. 3.80 (S.D.: 0.46) respectively, in a scale of 1 to 4, where 4 indicates that they are "Confident in extended conversations".

Materials

Critical trials: We compiled 5 translation-equivalent "base sentences" in Spanish and English (please see the Appendix for a list of base sentences). We created two sets of code-switched sentences, one in which the first switch went from English to Spanish (E.g.: *I like the MARRÓN dress; I like the dress MARRÓN*) and another one in which it went from Spanish to English (E.g.: *ME GUSTA EL VESTIDO brown; ME GUSTA EL brown VESTIDO*). Each base sentence was modified into code-switched forms following four switching patterns going from Spanish to English and four going from English to Spanish as follows:

MP-/MLF+: Sentences that follow the predictions of the Matrix Language Frame but not the Minimalist Program (E.g.: *I like the dress MARRÓN; ME GUSTA EL brown VESTIDO*)

MP+/MLF-: Sentences that follow the predictions of the Minimalist Program but not the Matrix Language Frame (E.g.: *I like the MARRÓN dress; ME GUSTA EL VESTIDO brown*)

MP-/MLF-: Sentences that do not follow the predicted pattern from either theory (e.g.: *I like the VESTIDO brown; ME GUSTA EL MARRÓN dress)*

MP+/MLF+: Sentences that follow the predictions of both models (E.g.: *I like the brown VESTIDO; ME GUSTA EL dress MARRÓN*)

In summary, we had 20 critical trials (4 variations for each base sentence) for each direction of switching (English to Spanish and Spanish to English).

In Spanish, all nouns were masculine or invariable (e.g.: *vestido* – 'dress', *baño* – 'bath', *cuaderno* – 'notebook', *cantante* – 'singer', *abogado* – 'attorney') and all adjectives were invariable with respect to gender (*marrón* – 'brown', *caliente* – 'hot', *verde* – 'green', *Nicaragüense* – 'Nicaraguan', *independiente* – 'independent'). Proper names in the sentences were chosen so that they were commonplace in both Spanish and English (e.g.: Max, Claudia). We avoided using nouns and adjectives whose onset would elicit changes in the preceding indefinite article depending on their position in some conditions (that is, avoid changes such as "Hugo is *a* singer estadounidense" vs. "Hugo is *an* estadounidense singer").

Filler trials: We included 48 non-critical trials for each direction of switch (English to Spanish and Spanish to English). In these filler trials, the focus of the CS was not the adjective but the determiner or the adverb, and just as with the critical items they were all variations of a set of base sentences (12 base sentences with 4 variations each). Some of the results for those trials will be reported elsewhere. By including these filler trials plus the quality control trials described below, critical trials made up only about a third of all pairs seen by participants. This was done to make it harder for raters to engage in strategic choices for their response (Cowart, 1996).

Quality control trials: There were 8 quality-control trials for each session; they consisted of sentences with inter-sentential code-switches. Half of the sentences had an uncontroversial error that could be easily detected if the sentences were read carefully (e.g.: *LA PASÉ MUY BIEN, the music *were excellent*). These errors were equally distributed among the following factors: first vs. second half of the sentence, English vs. Spanish portion, and type of error (verb tense, number agreement, gender agreement, and word order). If a participant failed more than 2 of these trials, they were removed from the sample and substituted with a new participant.

Procedure

The survey was administered online using Qualtrics and testing occurred across 3 separate sessions: one in which participants completed the language tests and the background questionnaire, and two counterbalanced sessions in which participants were presented with sentences containing NPs in which the switch went from English to Spanish and from Spanish to English, respectively. Test sessions were about a week apart from each other. Participants were given the choice of reading the instructions in English or Spanish; all but 4 participants chose to complete the questionnaire in English. Participants were informed that they would see a series of sentences and that they were to indicate on a 5-point scale how "permitted" a sentence was according to the way they would speak to- or hear from another bilingual person. In the scale, a score of 1 stood for “never permitted” while 5 stood for “always permitted”. Participants were then presented with the 76 code-switched sentences as described above (20 critical items, 48 fillers, and 8 quality control items). Each sentence was presented one at a time and the order of presentation was individually randomized for each participant. Participants had to rate each item before progressing to the next one and could not go back to previous sentences.

Results

Table 2 shows the mean acceptability ratings for each condition (on a scale from 1 to 5). We did not find significant correlations between acceptability ratings and test scores in either language (all *r* values < .22, all *p* values > .16). This is probably because language proficiency was constrained by our selection criteria requiring participants to be highly proficient in both languages and most participants scored near the top of the scale on both tests.

[Table 2 about here]

Rating scores were analyzed with a linear mixed effect model approach (Baayen, Davidson & Bates, 2008) using *lme4* (Bates, Mächler, Bolker & Walker, 2015). We included CS pattern and Direction as fixed factors, and participants and sentences as random structures. Statistical results were obtained using the *anova()* function of the package *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2014) with the Satterswaithe approximation for degrees of freedom. As suggested by Barr, Levy, Scheepers and Tily (20113), we analyzed the results with a “maximal” random effect structure in which slopes corresponding to both fixed factors were included for participants. We included a random intercept for participants (a model which included random slopes for both participants and sentences failed to reach convergence)[2]. The results showed a highly significant effect of CS pattern, $F(3, 52.35) = 24.35$, $MSE = 21.12$, $p < .001$, no effect of Direction, $F < 1$, and a significant interaction between CS pattern and Direction, $F(3, 54.90) = 2.93$, $MSE = 2.54$, $p = .042$. Simple effects analysis was conducted for CS pattern, separately for each level of Direction. For sentences where the switch went from English to Spanish, CS pattern was highly significant, $F(3, 48.37) = 20.23$, $MSE = 18.20$, $p < .001$, and a comparison of the levels of CS pattern with the Tukey HSD test showed that all levels differed significantly from each other, *z*s ≥ 3.11, *p*s ≤ .010, except for the MP+/MLF- and the MP-/MLF+ CS pattern which were not significantly different, $z = 1.24$, $p = .585$. For sentences where the switch went from Spanish to English, again CS pattern was highly significant, $F(3, 51.26) = 8.38$, $MSE = 7.18$, $p < .001$. Tukey HSD tests showed that the MP+/MLF+ differed significantly from all other CS patterns, *z*s ≥ 3.29, *p*s < .005, whereas the remaining three CS patterns did not differ significantly from each other, *z*s ≤ 2.43, *p*s ≥ .068.

In summary, MP+/MLF+ switches are the most preferred in both directions, while MP-/MLF- switches are the least preferred in the Spanish to English direction. There were no significant differences between MP+/MLF- and MP-/MLF+ sentences. The interaction between direction and condition reflects that, even though both directions show the same pattern of results, the differences between conditions were smaller in the Spanish to English direction. This may be because Spanish is more flexible than English regarding word order. The results from Experiment 1 show evidence of a hierarchy of preference for certain CS patterns, particularly in the English to Spanish direction, but do not differentiate between the two models of CS under evaluation. The average acceptability scores were all close to each other, slightly on the positive side of the middle of the scale.

In Experiment 2 we repeated the acceptability judgments for the same code-switched sentences with a different group of participants, but this time using a two-alternative forced choice presentation of the stimulus and using Thurstone's Law of Comparative Judgment (Case V) for the analysis of the results to find out if this alternative method provides a clearer picture of this phenomenon. Thurstone's analysis generates an interval scale based on comparisons of pairs of code-switched sentences that indicates a ranking of acceptability but also the relative distance between conditions. Given the suitability of this method for measuring subtle differences in introspective judgments, we expected to obtain a more fine-grained picture of the contrasts under study.

Experiment 2

Participants

A total of 40 early English/Spanish bilinguals living in the United States took part in this experiment, none of which participated in Experiment 1. Thirty-five participants were born in the United States, while the rest immigrated before age 5 (mean age when they migrated =3:7 years). Self-reported age of acquisition for Spanish (Mean: 1.83, S.D.: 0.96) was generally earlier than for English (Mean: 2.58; S.D.: 1.55), but all participants were able to speak both languages by the time they started elementary school. There were 22 men and 18 women and their mean age was 29:5 years (S.D.: 8.8; Range: 18-57). All participants were second-generation bilinguals (at least one of their parents was born in Mexico) and all stated that they spoke the Mexican variety of Spanish. All of them resided in states with large populations of Spanish speakers, namely California (N=15), Texas (N=13), Arizona (N=5), New York (N=3), New Jersey (N=2), Colorado (N=1) and Nevada (N=1). The highest level of education for 21 participants was high school, 17 had attended at least some college, and 2 attended graduate school. When asked "When chatting with friends and family that speak both English and Spanish, do you mix Spanish and English in the same sentence?" all but 2 participants responded in the positive. Participants were also recruited through Amazon Mechanical Turk and completed the same questionnaire and language tests as those in Experiment 1.

Language Proficiency and Dominance. Participant's average scores in the language proficiency tests were slightly but significantly better for English (M =45.2; S.D. =2.8) than for Spanish (M =43.7; S.D. =3.8); t(78) =2.01, $p < .05$. All participants stated that they were able to speak Spanish before age four and English by the time they entered elementary school. They declared to be slightly more proficient in English than Spanish, with ratings of 3.98 (S.D.: 0.16) vs. 3.68 (S.D.: 0.47) respectively.

Materials

Critical trials: Participants were presented with the same critical sentences as Experiment 1, but they were compared pairwise among themselves so that each variation of a base sentence was contrasted with each of the other variations combinatorially (MP+/MLF+ vs. MP-/MLF+; MP+/MLF+ vs. MP+/MLF-; MP+/MLF+ vs. MP-/MLF-; MP-/MLF+ vs. MP+/MLF-; MP+/MLF- vs. MP-/MLF-). The number of combinations of sentence pairs is given by the formula (n*(n-1)/2), where n is the number of conditions being evaluated, in our case we had 4 conditions and thus 6 pairwise comparisons per base sentence. This generated 30 judgments for each direction of the switch (English to Spanish and Spanish to English). Sentences were contrasted only within their own directionality set in two separate sessions, that is, code-switched sentences going from English to Spanish were only contrasted with other sentences also going from English to Spanish and so on. This was done in order to control for possible preferences in the directionality of the switch (Parafita Couto et al. 2015) and insulate the effect of noun-adjective order, which is the focus of the present study. This was also a pragmatic decision because the combinatorial outcome of comparing all possible pairs of the resulting 8 variations for each sentence made the experiment unfeasible (it would have generated 28 combinations per base sentence instead of 6).

Monolingual trials: We presented a separate group of 40 early Spanish/English bilinguals with monolingual versions (either English or Spanish) of the base sentences used in this experiment. English and Spanish sentences were presented separately and the order of presentation of language blocks was counterbalanced across participants. The monolingual sentences were presented in pairs in which the adjective was pre- or post-nominal and asked them to choose which one was more acceptable. In Spanish, 99% of responses indicated a preference for a post-nominal adjective, while in English 98.5% of responses indicated a preference for pre-nominal adjectives.

Filler trials: In each session we included 72 non-critical trials where the focus of contrast between choices in a pair was not the adjective but the determiner or the adverb so that the critical trials made up only about a third of all pairs seen by participants. Some of the results for the filler trials will be reported elsewhere.

Quality control trials: There were 12 quality-control trials that consisted of pairs of sentences with inter-sentential code-switches. One of the members of each pair had an uncontroversial error that could be easily detected if the sentences were read carefully. If a participant failed more than 2 of these trials, they were removed from the sample and substituted with a new participant.

Procedure

The survey was administered online using Qualtrics and testing occurred across 3 separate sessions: one in which participants completed the language tests and the background questionnaire, and two in which they completed half of the 2AFC trials. Participants were presented with sentences containing NPs in which the switch went from English to Spanish in one session and from Spanish to English in the other, counterbalancing across participants. At the beginning of each session participants were given the choice of reading the instructions in English or Spanish; all but 5 participants chose to complete the questionnaire in English. The instructions informed participants that they would see a series of sentence pairs, and asked them to pick the one closer to the way they would speak to another bilingual person. They were asked to make a choice even if both sentences sounded “right” or both sounded “wrong”. Then participants were presented with the 114 pairs of code-switched sentences described above. The pairs of sentences were presented one at a time and the order of presentation of the pairs (critical trials, fillers, and quality control trials), as well as the order of each sentence within each pair, was individually randomized for each participant. Participants had to make a choice for each trial before progressing to the next one and could not go back to previous sentences.

Results

As outlined above, in this experiment on each trial, two versions of one of the five base sentences corresponding to two out of the four CS patterns were presented in pairs, and on each trial the participant chose the preferred sentence. Hence, six possible combinations of CS patterns existed, with each CS pattern occurring in three of them. We tallied preference responses such that per participant, base sentence, Direction, and CS pattern, preference was expressed as a percentage of possible choices.

These responses were analyzed using Thurstone's (1927) Law of Comparative Judgment, Case V, which analyzes participants' pairwise comparison of the stimuli to generate a ranking of preference among conditions as well as a measure for relative comparison between them. The results of Thurstone's analysis must be interpreted within the context of signal detection theory (c.f. Cowart, 1996): every measurement includes a noise component that usually follows a normal distribution. If we repeatedly measure the same construct, we will end up with a normal distribution centered on the likely "ideal" value of the magnitude, with a lesser probability of obtaining large errors of judgment than small errors in either direction. Thurstone's "measure" provides the center or mean of such a normal distribution, which Thurstone called "discriminal dispersions", for each condition. These measures can be interpreted values on an interval scale that represents a psychological continuum (in our case, the acceptability of the sentences). The unit of measurement along that scale is defined as the standard deviation of the distribution (Brown & Peterson, 2009), so the measure itself provides information about its variability.

Table 3 shows the rank order and Thurstone's measure for each condition. The measure values are relative to the pattern with the lowest acceptability (which is by convention set to 0).

The 95% confidence interval for this set of data was 0.071; differences between all conditions fall outside of this interval.

[Table 3 about here]

We also statistically analyzed the preference of each CS pattern, as shown on the right side of Table 3. Once again, we did not find significant correlations between acceptability ratings and test scores in either language (all r values < .24, all p values > .20). As in the first experiment, we analyzed the results with a linear mixed effect model which included CS pattern and Direction as fixed factors, and participants and sentences as random structures, and for which random slopes were included for participants. The results showed a highly significant effect of CS pattern, $F(3, 52.60) = 110.85$, $MSE = 79.60$, $p < .001$, no effect of Direction, $F < 1$, and a significant interaction between CS pattern and Direction, $F(3, 51.71) = 5.96$, $MSE = 4.28$, $p = .001$. Simple effects analysis showed that for sentences where the switch went from English to Spanish, CS pattern was highly significant, $F(3, 39) = 88.54$, $MSE = 60.92$, $p < .001$, and a comparison of the levels of CS pattern with the Tukey HSD test showed that all levels differed significantly from each other, zs ≥ 4.99, ps < .001, except for the MP+/MLF- and the MP-/MLF+ CS pattern which did not significantly, $z = 2.14$, $p = .139$. For sentences where the switch went from Spanish to English, again CS pattern was highly significant, $F(3, 45.45) = 39.12$, $MSE = 30.09$, $p < .001$. Tukey HSD tests showed that all levels differed significantly from each other, zs ≥ 2.71, ps < .032, except for the MP+/MLF- and the MP-/MLF+ CS pattern which did not significantly, $z = 1.64$, $p = .346$.

The results of Experiment 2 show a clear hierarchy of preference for the CS patterns tested. When contrasting the two CS models under study, the results seem to indicate a preference for the predictions of the MP over MLF. It is important to point out, however, that MP-/MLF+ condition is preferred over the MP-/MLF- condition, and that, similar to the results of Experiment 1, the preferred condition was that in which both constraints were satisfied. Similarly to Experiment 1, the interaction between direction and condition reflects that, even though both directions show the same pattern of results, the differences between conditions were somewhat smaller in the Spanish to English direction.

Discussion

In this study, we investigated the contrasting predictions derived from the MLF and the MP (specifically, the generalization by Cantone & MacSwan, 2009) regarding the mechanisms underpinning code-switching between the noun and the adjective. In two critical conditions, the predictions for adjective positioning made by the MLF and the MP differed. The MLF predicts a violation when the adjective position is incompatible with the word order of the sentence's Matrix Language, while the MP predicts a violation when the adjective position is disallowed by the language of the adjective. While Experiment 1 (with traditional Likert scales) didn't yield a clear differentiation between the predictive power of each model, Experiment 2 indicated an advantage for the MP. However, the difference between the MP+/MLF- and MP+/MLF+ conditions points towards an additive effect in which both the language of the verb and the language of the adjective are used to determine word order in the NP. If decisions were based only on the language of the adjective (as predicted by the MP), there would be no difference between MP+/MLF- and MP+/MLF+ because the language of the verb would have nothing to add above and beyond what can be explained by the language of the adjective in regards to the acceptability of a given sentence. On the other hand, the fact that there is a difference between

MP-/MLF+ and MP-/MLF- confirms that the language of the verb contributes to the acceptability decisions of informants.

Our results do not lend support to the suggestion that it is just the matrix language or the language of the adjective that determine the relative order of adjectives and nouns in code-switched nominal constructions. An earlier proposal by Santorini and Mahootian (1995) and Mahootian and Santorini (1996) is also not supported by our data. They postulated that all combinations of adjectives and nouns are possible because only heads determine the position of its complements, and adjectives are nominal adjuncts. This proposal was disputed in the literature, but recently DiSciullo (2014) argued along similar lines that code-switching is possible in modification sites. She assumes that adverbial and adjectival modifiers occupy the specifier position of a functional (F) category asymmetrically c-commanding the lexical projections it modifies. Code-switching is predicted to be possible in these sites, as code-switching may occur at the juncture of an Adjective and a Noun, since Adjectives are generated by External Merge in the specifier of a functional category. Instead of supporting one theory over another, our results point to a model where insights from different frameworks are combined: while features are important (as in the lexicalist/generative view), we should also consider abandoning a strict version of lexicalism and adopting constructionist approaches (cf. Eppler et al., 2016).

Trying to combine insights from both the MLF and the MP, a potential interpretation of our findings would be to postulate a potential parallel between the MLF and the MP, as already proposed by Radford, Kupisch, Köppe, and Azzaro (2007). They argued that it is possible to equate the MLF's notion of morphosyntactic frame with the MP's notion of Phase (a derivational unit in minimalist syntax). Their rationale is that "the head of a phase is responsible (via a form of selection) for “handing over” functional features to subordinate items within the phase” (Radford et al, 2007, p. 245). In a similar fashion, within the MLF, a mixed utterance has a morphosyntactic frame (from the matrix language), and morphemes from an embedded language can be inserted into this frame. Properties of the matrix frame will determine the nature of the embedded morphemes which can be inserted within it. Our results point to the adjective position being partially dependent on the verb in the higher Complementizer Phrase. Hence, we could speculate that our data suggests that the MLF may dominate the whole Complementizer Phrase (CP) phase, thus making both theoretical models compatible and complementary. Additionally, our results seem to provide evidence against the more general proposal within Minimalism that the DP (Determiner Phrase) is a separate phase (Svenonius, 2004, Hiraiwa, 2005), since information outside the DP needs to be taken into account when building a nominal construction. Hence, beyond supplementing the theoretical debate between proponents of the MLF and the MP, our results have important implications for any possible analysis of noun phrase structure. Crucially, this holds true for mixed nominal constructions as well as for monolingual nominal constructions, but using mixed nominal constructions in our stimuli allowed us to see what otherwise would have remained hidden in monolingual grammars. However, further research on different switching points is needed before we can say that the notion of frame within the MLF and the notion of phase within the MP can indeed be equated.

An alternative way to look at our results (i.e. hierarchical rather than dichotomous acceptability of CS patterns) is by rejecting formal, rule-based theories of code-switching and espousing a “probabilistic” approach (e.g.: Bresnan, 2007; Bresnan, Cueni, Nikitina, & Baayen, 2004; Kootstra, van Hell, & Dijkstra, 2009) in which the acceptability and use of code-switched structures are based on multiple (sometimes competing) probabilistic constraints, usually based on frequency or recency of exposure to particular syntactic structures. As suggested by an anonymous reviewer, a constraint-based approach makes it possible to conceptualize predictions from MLF and the MP in terms of multiple constraints on the grammatical form of code-switched sentences, which determine the form of code-switched in parallel (in a probabilistic manner). For example, Bresnan (2007) on her study on English dative alternation, found that the probability of occurrence of a given sentence affected how acceptable it was judged to be. Within the field of code-switching research, Kootstra et al. (2009) investigated the role of shared word order and alignment with a dialogue partner in the production of code-switched sentences. They found that participants had a clear preference for using the shared word order when they switched languages, but also aligned their word order choices and code-switching patterns with the confederate. It is conceivable that the hierarchy of preference yielded by our experiments could be explained by different levels of exposure to each type of construction, instead of the combination of formal internalized rules. These possibilities must be tested empirically, something that falls outside of the scope of the present study.

Kootstra et al. 2009 also suggest that the equivalence constraint (Poplack, 1980) may well be a probabilistic constraint on code-switching, surfacing as a general tendency amenable to interaction with other forces of code-switching. While it is true that violations of the equivalence constraint have been attested in naturalistic production (see for ex. Bentahila & Davies (1983) for Arabic-French; Berk-Seligson (1986) for Spanish-Hebrew; DiSciullo, Muysken & Singh (1986) for Italian- English, and Parafita Couto & Gullberg, submitted for Welsh-English, Spanish-English, and Papiamento-Dutch), it is also true that they are not as common as other types of switches that do not occur at conflict sites. For the particular case of adjective-noun switches, the low occurrence of this type of switch may also be due in part to the fact that attribute adjectives are not used frequently in naturalistic speech. However, while structural equivalence may in fact explain the higher frequency of switches at structurally equivalent points in production data, in this study we were interested in unveiling whether we could discover any regular patterns in comprehension when the equivalence constraint was violated.

An avenue for future research would be to test Spanish-English bilinguals in a different community (for example in Gibraltar). Valdés Kroff (2016) argues that code-switching is a learned behavior and that different code-switching patterns may be learned in different communities of code-switchers. Consequently, whether bilinguals have immersed themselves in such a community, i.e., the bilingual profile in terms of usage and exposure to code-switching, should result in observable group differences in the production and comprehension of code-switching. Similarly, community differences in the preferred pattern of use in code-switching can and should arise because the specific structure that a community adopts may be influenced by a host of linguistic and extra-linguistic variables. The only way to further understand code-switching and the nature of the relation between frequency of production in naturalistic speech on the one hand, and processing on the other, would be to continue to collect multiple types of data from a variety of bilingual populations (see Gullberg et al., 2009).

Code-switching patterns may not only be influenced by community practices, but also by the specific language pair under investigation. In this regard, the generalizability of our results to other language pairs is an empirical question; we are currently in the process of collecting comparable data in several bilingual communities with a variety of language typologies. We are greatly encouraged on those efforts by the interesting results we obtained using this methodology in the study presented here.

As evidenced by the results presented here and attested elsewhere (Gigerenzer, Krauss, & Vitouch, 2004; Gigerenzer & Richter, 1990; Sprouse, 2011; Sprouse & Almeida, 2011; Stadthagen-Gonzalez et al., in press), the use of a 2AFC task offers a high degree of granularity in the discrimination of differences between the acceptability of different conditions when compared to Likert scales. On the other hand, one limitation of this method stems from its combinatorial nature: the number of paired comparisons in a survey quickly escalates when more conditions and/or more exemplars for each condition are included in the design. The number of paired comparisons per exemplar is given by the formula (n*(n-1)/2), where n is the number of conditions being evaluated. In our case, we had 4 conditions, so the addition of each exemplar sentence would increase the number of critical items by 6. This is exacerbated by the need to include filler items, as recommended by Cowart (1996); in this study we chose to have 2 filler comparisons for each critical comparison, so, in total, the addition of one more exemplar would have resulted in the inclusion of at least 18 more items in the survey. Our ability to include a larger variety of items to the survey was thus limited by the number of survey items generated by each extra base sentence. One way of dealing with this problem is to use "partial comparison models" developed in the field of psychophysics (e.g.: Bradley & Terry, 1952) in which it is not necessary to compare all pairs exhaustively in order to generate a Thurstonian scale and hierarchy. Because the use of Thurstone's analysis to the study of linguistic acceptability judgments is in its infancy, we chose the original, full-comparison design for the present study, but we are currently working on adapting the application of the above mentioned partial comparison designs to linguistic research.

A potential criticism to the method used in Experiment 2 is that repetition effects could emerge from presenting variations of the same base sentences several times. This is a possibility for ratings provided in an absolute (Likert) scale such as in Experiment 1: familiarity with the items might affect their acceptability and shift their ratings. However, in a two-alternative forced choice task, participants are asked to judge two sentences relative to each other, and not relative to an absolute value. All sentence variations were presented the same number of times, and their order of presentation was individually randomized for each participant. It is thus reasonable to believe that any potential familiarity (or repetition) effect will not systematically favor a particular condition; it would even out. In the end, the *relative* preference for one condition over another, as reflected in the interval scale generated by Thurstone's analysis, would be preserved.

The use of the 2AFC task used in this study has provided robust and unprecedented insights into the validity of theories of code-switching. As discussed above (and more extensively in the psychophysics literature), the 2AFC task represents an improvement over other methods of measuring introspective judgments and provides clear, interpretable data that can be used for directly contrasting the predictions of linguistic theoretical models. This approach opens the way for systematic testing of predictions from linguistic theory in general, especially in cases where corpus data appear insufficient to test all possible combinations of theoretical predictions. From a theoretical standpoint, the results of this experiment cannot be accounted for by neither the MP nor the MLF model on their own, and thus provide motivation for a re-formulation of our current understanding of the mechanisms underlying the acceptability of code-switched sentences that goes beyond those two popular models. This reformulation may stem from combining the insights of the two frameworks instead of considering them in isolation (see Eppler et al. 2016), or by espousing a probabilistic model of code switching (e.g. Bresnan, 2007; Bresnan, et al., 2004; Kootstra, et al., 2009).

REFERENCES

Anderson, T. (2006). *Spanish-English bilinguals' attitudes toward code-switching: proficiency, grammaticality, and familiarity*. Doctoral dissertation, The Pennsylvania State University, State College, Pennsylvania.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412. doi:10.1016/j.jml.2007.12.005.

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics, 46*, 273-330.

Bentahila, A. & Davies, E.E. (1983) The syntax of Arabic-French code-switching. *Lingua,59,* 301-330. doi: 10.1016/0024-3841(83)90007-4

Berk-Seligson, S. (1986) Linguistic constraints on intrasentential code-switching: a study of Spanish/Hebrew bilingualism. *Language in Society, 15*, 313-348.

Bock, R. D., & Jones, L.V. (1968). *The measurement and prediction of judgment and choice*. San Francisco, CA: Holden-Day.

Blokzijl, J., Deuchar, M., & Parafita Couto, M. (Submitted). Determiner asymmetry in mixed nominal constructions: the role of grammatical factors in data from Miami and Nicaragua. Manuscript submitted to Languages.

Branigan, H.P., & Pickering, M.J. (2016) An experimental approach to linguistic representation, *Behavioral and Brain Sciences*. Advance online publication. doi:10.1017/S0140525X16002028

Bradley, R.A., & Terry M.E. (1952) Rank analysis of incomplete block designs. I. The method of paired comparisons, *Biometrika*, *39*, 324–345

Bresnan, J. (2007) Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston and W. Sternefeld (eds) *Roots: Linguistics in Search of its Evidential Base* (pp. 77–96). Berlin: Mouton de Gruyter.

Bresnan, J., Cueni, A., Nikitina, T. & Baayen, R.H. (2007) Predicting the Dative Alternation. In G. Boume, I. Kraemer and J. Zwarts (eds) *Cognitive Foundations of Interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Science.

Brown, T.C., & Peterson, G.L. (2009). *An enquiry into the method of paired comparison: reliability, scaling, and Thurstone's Law of Comparative Judgment*. Fort Collins, CO: U.S. Department of Agriculture, Forest Service.

Bullock, B.E. & Toribio, A. J. (Eds.) (2009). *The Cambridge Handbook of Linguistic Code-switching*. Cambridge: Cambridge University Press.

Cantone, K. F., & MacSwan, J. (2009). The syntax of DP-internal codeswitching. In L. Isurin, D. Winford & K. de Bot (eds.) *Multidisciplinary Approaches to Codeswitching* (pp. 243-78). Amsterdam: John Benjamins Publishing.

Carol, J.M., Bever, T.G., & Pollack, C.R. (1981). The non-uniqueness of linguistic intuitions. *Language*, *57*, 368-383.

Cinque, G. (1994). On the evidence for partial N-movement in the Romance DP. In G. Cinque, J. Koster, J-Y Pollock, L. Rizzi & R. Zanuttini (Eds.), *Paths towards Universal Grammar. Studies in honor of Richard S. Kayne* (pp. 85-110). Washington, DC: Georgetown University Press.

Cowart, W. (1996). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. California: Sage Publications Inc.

David, H.A. (1988). *The method of paired comparisons* (2nd ed.). New York, NY: Oxford

University Press.

Davies, P. (2010). Identifying word-order convergence in the speech of Welsh-English bilinguals. PhD dissertation, Bangor University.

Davies, P., & Deuchar, M. (2010). Using the Matrix Language Frame model to measure the extent of word order convergence in Welsh-English bilingual speech. In A. Breitbarth, C. Lucas, S. Watts & D. Willis (eds.) *Continuity and change in grammar* (pp. 77–96). Philadelphia: John Benjamins.

Deuchar, M. (2012). Code Switching. In Chapelle, C.A. (ed.) *The Encyclopedia of Applied Linguistics* (pp. 664-675). Chichester, UK: Wiley-Blackwell.

Deuchar, M. (2006). Welsh-English code-switching and the Matrix Language frame model. *Lingua*, *116*, 1986-2011.

Deuchar, M. and Davies, P. (2009). Code-switching and the future of Welsh. *International Journal of the Sociology of Language*, *195,* 15–38.

DiSciullo, A. M. (2014). On the Asymmetric Nature of the Operations of Grammar: Evidence from Codeswitching. In J. MacSwan (Ed.) *Grammatical Theory and Bilingual Codeswitching* (pp. 63-87). Cambridge: MIT Press

DiSciullo, A.M., Muysken, P. & Singh, R (1986). Government and Code-Switching, *Journal of Linguistics, 22*, 1-24.

Eppler, E.D., Luescher, A. & Deuchar, M. (2016) Evaluating the predictions of three syntactic frameworks for mixed determiner–noun constructions. *Corpus Linguistics and Linguistic Theory*, Advance online publication. DOI: 10.1515/cllt-2015-0006

Fairchild, S., & Van Hell, J. (2015). Determiner-noun code-switching in Spanish heritage speakers. *Bilingualism: Language and Cognition.* Advance online publication. DOI:

10.1017/S1366728915000619.

Fechner, G. T. 1876. *Vorschule der Aesthetik, Vol. 2.* Leipzig: Breitkopf & Härtel.

Gescheider, G. (1997). *Psychophysics: the fundamentals (3rd ed.).* New Jersey: Lawrence Erlbaum Associates.

Giancaspro, D. (2013) L2 learners' and Heritage speakers' judgments of code-switching at the auxiliary-VP boundary. In J. Cabrelli-Amaro, G. Lord, A. de Prada, & J. Aaron (Eds.), *Selected Proceedings of the 16th Hispanic Linguistics Symposium* (pp. 56–69). Somerville, MA: Cascadilla Proceedings Project.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to Obtain and Analyze English Acceptability Judgments. *Language and Linguistics Compass, 5*, 509-524.

Gigerenzer, G., Krauss, S. & Vitouch, O. (2004) The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (ed.) *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage.

Gigerenzer, G., & Richter, H. (1990) Context effects and their interaction with development: Area judgments. *Cognitive Development*, 5, 235–264.

Green, D. M., & Swets, J. A. (1988). *Signal Detection Theory and Psychophysics*. Los Altos, CA: Penninsula.

Guzzardo Tamargo, R.E., Mazak, C., & Parafita Couto, M.C. (Eds.) (2016) *Spanish-English CodeSwitching in the Caribbean and the U.S.* Amsterdam: John Benjamins.

Gullberg, M., Indefrey, P., & Muysken, P. (2009). Research techniques for the study of code-switching. In Bullock, B. E. & Toribio, J. A. (Eds.) The Cambridge Handbook of linguistic code-switching (pp. 21-39). Cambridge: Cambridge University Press.

Herring, J., Deuchar, M., Parafita Couto, M.C., & Moro Quintanilla, M. (2010). "When I went to Canada, I saw the madre": evaluating two theories' predictions about codeswitching between determiners and nouns using Spanish-English and Welsh-English bilingual corpora, *International journal of bilingual education and bilingualism, 13,* 553-573.

Hiraiwa, K. (2005) Dimensions of Symmetry in Syntax: Agreement and Clausal Architecture. PhD dissertation, Massachusetts Institute of Technology.

Jake, J. L., Myers-Scotton, C. M., & Gross, S. (2005). A response to MacSwan (2005): Keeping the Matrix Language. *Bilingualism: Language and Cognition, 8,* 271–276.

Kootstra, G. J., Van Hell, J. G., & Dijkstra, T. (2012). Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency. *Bilingualism: Language and Cognition*, *15,* 797-819.

Labov, W. (1972). Some principles of linguistic methodology. *Language in Society, 1,* 97-120.

Labov, W. (1975). *What is a linguistic fact?* Lisse: Peter de Ridder.

MacSwan, J. (2005a) Codeswitching and generative grammar: A critique of the MLF model and some remarks on "modified minimalism." *Bilingualism: Language and Cognition, 8,*1-22.

MacSwan, J. (2005b) Comments on Jake, Myers-Scotton and Gross's response: There is no "matrix language." *Bilingualism: Language and Cognition, 8,* 277-284.

Mahootian, S., & Santorini, B. (1996). Code-switching and the complement/adjunct distinction. *Linguistic Inquiry, 27*, 464–479.

Munarriz A. & Parafita Couto M.C. (2014) ¿Cómo estudiar el cambio de código? Incorporación de diferentes metodologías en el caso de varias comunidades bilingües. *Lapurdum. Basque Studies Review, 18*, 43-73.

Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford, UK: Oxford University Press.

Nagata, H. (1988). The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *Journal of Psycholinguistic Research, 171*, 1-17.

Nunnally, J.C. (1976). *Psychometric theory*. New York: McGraw Hill.

Oxford University Language Centre (n.d.). “Placement Tests.” lang.ox.ac.uk. http://www.lang.ox.ac.uk/tests/index.html (accessed July 1st, 2015)

Pablos, L. Parafita Couto, M.C., Boutonnet, B., de Jong, A. Perquin, M., de Haan A., & Schiller, N.O. (submitted). Adjective-Noun order in Papiamento-Dutch code-switching. Manuscript submitted to *Linguistic Approaches to Bilingualism.*

Parafita Couto, M.C., Boutonnet, B., Hoshino, N., Davies, P., Deuchar, M. & Thierry, G. (2017). Testing Alternative Theoretical Accounts of Code-Switching using Event-related Brain Potentials: A Pilot Study on Welsh-English. In: Lauchlan F., Parafita Couto M.C. (Eds.) *Bilingualism and minority languages in Europe: Current trends and developments*. Newcaslte: Cambridge Scholars. 242-256.

Parafita Couto, M. C., Deuchar, M., & Fusser, M. (2015). How do Welsh-English bilinguals deal with conflict? Adjective-noun order resolution. In G. Stell, & K. Yakpo (eds.) *Code-Switching Between Structural and Sociolinguistic Perspectives* (pp. 65-84). Berlin: De Gruyter.

Parafita Couto, M.C. & Gullberg, M. (submitted). Code-switching within the Noun Phrase. Evidence from three corpora. Manuscript submitted to International Journal of Bilingualism.

Parraga, C.A. (2015). Perceptual Psychophysics. In G. Cristobal, M. Keil & L. Perrinet (eds.)

Biologically-Inspired Computer Vision: Fundamentals and Applications (pp. 81-108). New York, NY: Wiley.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46,* 1023-1031.

Poplack, S., & Meechan, M. (1998). How languages fit together in codemixing. *International Journal of Bilingualism, 2,* 127-38.

Radford, A., Kupisch, T., Köppe, R., & G. Azzaro, G. (2007). Concord, Convergence and Crash-Avoidance in bilingual children. *Bilingualism: Language and Cognition, 10,* 239-256.

Ross, J.R. (1979). Where's English? In C.J. Fillmore, D. Kemper & W.S. Wang (eds.) *Individual differences in language ability and language behavior* (pp. 127-163). New York: Academic Press.

Santorini, B., & Mahootian S. (1995) Code-switching and the syntactic status of adnominal adjectives. *Lingua, 96,* 1-27.

Schütze, C.T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, Il: University of Chicago Press.

Snow, C., & Meijer, G. (1977). On the secondary nature of syntactic intuitions. In S. Greenbaum (ed.) *Acceptability in language* (pp. 163-177). The Hague, the Netherlands: Mouton.

Sprouse, J. (2011) A Test of the Cognitive Assumptions of Magnitude Estimation: Commutativity does not Hold for Acceptability Judgments. *Language, 87,* 274-288.

Sprouse, J. & Almeida, D. (2011). Power in acceptability judgment experiments and the reliability of data in syntax. Ms., University of California, Irvine & Michigan State University.

Stadthagen-Gonzalez, H., López, L., Parafita Couto, M.C., & Parraga, C.A. (in press) Using two-alternative forced choice tasks and Thurstone's Law of Comparative Judgment for Acceptability Judgments in Code Switching. *Linguistic Approaches to Bilingualism*.

Stokes, W. (1974). All of the work on quantifier-negation isn't convincing. In M.W. La Galy, R.A. Fox & A. Bruck (eds.) *Papers from the tenth regional meeting, Chicago Linguistic Society* (pp. 692-700). Chicago: Chicago Linguistic Society.

Svenonius, P. (2004). On the edge. In D. Adger, C. de Cat & G. Tsoulas (eds.) *Peripheries: Syntactic edges and their effects* (pp. 261-287). Dordrecht: Kluwer.

Thurstone, L. (1927). A Law of Comparative Judgment. *Psychological Review, 34*, 273-286.

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York, NY: John Wiley & Sons.

Valdés Kroff, J. R. (2016). Mixed NPs in Spanish-English bilingual speech: Using a corpus based approach to inform models of sentence processing. In R. E. Guzzardo Tamargo, C. M. Mazak, & M. C. Parafita Couto (Eds.), *Spanish-English code-switching in the Caribbean and the US* (pp. 281–300). Amsterdam, The Netherlands: John Benjamins.

vanden Wyngaerd, E.V. (2016) The adjective in Dutch-French codeswitching: Word order and agreement. *International Journal of Bilingualism*. Advance online publication. DOI: 10.1177/1367006916632302

Footnotes

[1] For both experiments the results of our analyses with and without the participants that stated that they didn't code-switch were virtually identical, so we decided to keep them in the sample.

[2] The R syntax for this analysis was: `anova(lmer(Score ~ CS_pattern * Direction + (1 + CS_pattern * Direction | Participant) + (1 | Sentence)))`

Table 1. *Prediction of acceptability for code-switched constructions according to the MP and MLF models.*

English→Spanish			Spanish→English		
	Exemplar	Predicted Acceptability		Exemplar	Predicted Acceptability
(1)	I like the *marrón* dress	MP-/MLF+	(5)	*Me gusta el vestido* brown	MP-/MLF+
(2)	I like the dress *marrón*	MP+/MLF-	(6)	*Me gusta el* brown *vestido*	MP+/MLF-
(3)	I like the *vestido* brown	MP-/MLF-	(7)	*Me gusta el marrón* dress	MP-/MLF-
(4)	I like the brown *vestido*	MP+/MLF+	(8)	*Me gusta el* dress *marrón*	MP+/MLF+

MP+ =Acceptable according to the Minimalist Program; MP- =Not acceptable according to the Minimalist Program; MLF+ =Acceptable according to the Matrix Language Frame model; MLF- =Not acceptable according to the Matrix Language Frame model.

Table 2. *Mean acceptability rating for each condition*

Condition	English→Spanish	Spanish→English
MP+/MLF+	3.71 (S.D. 0.93)	3.71 (S.D. 0.86)
MP+/MLF-	3.08 (S.D. 0.90)	3.25 (S.D. 0.96)
MP-/MLF+	3.24 (S.D. 0.94)	3.12 (S.D. 1.04)
MP-/MLF-	2.60 (S.D. 0.88)	2.89 (S.D. 1.04)

Table 3. *Ranking and measure for each condition*

		Thurstone's Measure		Mean % Preference (S.D.)	
Rank	Condition	English→Spanish	Spanish→English	English→Spanish	Spanish→English
1	MP+/MLF+	2.26	1.32	39.0 (7.9)	34.0 (8.5)
2	MP+/MLF-	1.29	0.82	27.3 (8.3)	27.0 (7.8)
3	MP-/MLF+	0.90	0.50	22.2 (9.1)	22.8 (10.5)
4	MP-/MLF-	0	0	11.5 (6.7)	16.2 (6.9)

APPENDIX

Monolingual base sentences

English	**Spanish**
I like the brown dress	Me gusta el vestido marrón
I need a hot bath	Necesito un baño caliente
Samuel wrote it all in his green notebook	Samuel lo escribió todo en su cuaderno verde
Hugo is a Nicaraguan singer	Hugo es un cantante Nicaragüense
They had to hire an independent attorney	Tuvieron que contratar un abogado independiente

Code-Switched Critical Sentences

MP-/MLF+

English → Spanish	Spanish → English
I like the marrón dress	Me gusta el vestido brown
I need a caliente bath	Necesito un baño hot
Samuel wrote it all in his verde notebook	Samuel lo escribió todo en su cuaderno green
Hugo is a Nicaragüense singer	Hugo es un cantante Nicaraguan
They had to hire an independiente attorney	Tuvieron que contratar un abogado independent

MP+/MLF-

English → Spanish	Spanish → English
I like the dress marrón	Me gusta el brown vestido
I need a bath caliente	Necesito un hot baño
Samuel wrote it all in his notebook verde	Samuel lo escribió todo en su green cuaderno
Hugo is a singer Nicaragüense	Hugo es un Nicaraguan cantante

They had to hire an attorney independiente	Tuvieron que contratar un independent abogado

MP-/MLF-

English → Spanish	Spanish → English
I like the vestido brown	Me gusta el marrón dress
I need a baño hot	Necesito un caliente bath
Samuel wrote it all in his cuaderno green	Samuel lo escribió todo en su verde notebook
Hugo is a cantante Nicaraguan	Hugo es un Nicaragüense singer
They had to hire an abogado independent	Tuvieron que contratar un independiente attorney

MP+/MP+

English → Spanish	Spanish → English
I like the brown vestido	Me gusta el dress marrón
I need a hot baño	Necesito un bath caliente
Samuel wrote it all in his green cuaderno	Samuel lo escribió todo en su notebook verde
Hugo is a Nicaraguan cantante	Hugo es un singer Nicaragüense
They had to hire an independent abogado	Tuvieron que contratar un attorney independiente