Code-switching in Tunisian Arabic: a multi-factorial random forest analysis

Chadi Ben Youssef[1] and Stefan Th. Gries[1]

Abstract

This paper explores the morphosyntactic and cognitive principles influencing code-switching (CS) from Tunisian Arabic to French. We annotate data from the TuniCo corpus for many variables and run a Random Forest to overcome the methodological challenges typically associated with low-resource languages and imbalanced data. We find CS is not affected by any factor in isolation, but by a constellation of interactions. Our results partially confirm previous findings: (*i*) to maintain the code-integrity at the phrase and discourse levels, speakers tend to switch dependent parts-of-speech when the latter's head is switched; (*ii*) NPS are a prime location for CS; and (*iii*) speakers are attuned to the cognitive load they impose on themselves and/or on listeners.

Keywords

code integrity, code momentum, code-switching, French, part-of-speech, random forest, Tunisian Arabic

1. Introduction

To achieve a given communicative goal, speakers must choose between competing strategies and functional units to form their utterances (Du Bois, 1985). From this competition, grammars emerge and are reshaped constantly. If this is true of monolingual settings, then this must be even more salient in bilingual speech communities and diglossic societies (Ferguson, 1959). In such contexts, speakers have to select not only from the affordances of a single language but, rather, from two or more repertoires in constant competition. This can sometimes lead to the occurrence of code-switching (hereafter, CS). CS is 'the alternating use of two languages in the same stretch of discourse by a bilingual speaker' (Bullock and Toribio, 2009: xii). What motivates

[1] Department of Linguistics, University of California, Santa Barbara, Santa Barbara, CA 93106-3100, USA.
Correspondence to: Chadi Ben Youssef, *e-mail*: chadi.benyoussef@ucsb.edu

Corpora 2023 Vol. 18 (3): 297–327
DOI: 10.3366/cor.2023.0289
© Edinburgh University Press
www.euppublishing.com/cor

CS has been one of the most researched phenomena in language contact since the influential publication of Poplack (1980) on the subject. However, studying CS using multi-factorial quantitative techniques is less common and tends to use either Internet written data (e.g., Gambäck and Das [2016]), bilingual immigrant speech communities' data (e.g., Carter *et al*. [2010]) or conversations occurring between a limited number of speakers in an intimate context (e.g., Myslín and Levy [2015]).

This study is an instance of such a multi-factorial corpus study of CS in naturally occurring conversations and narrative sociolinguistic interviews collected in Tunisia, and addresses the question of what motivates and constraints a multilingual speaker to code-switch in a context characterised by diglossia. The challenge is two-fold: (*i*) Tunisian Arabic is a low-resource language, which imposes certain limitations on the operationalisation of a number of hypotheses, and (*ii*) the inherently imbalanced nature of CS corpora makes the use of 'traditional' statistical techniques such as mixed-effects generalised linear regression modelling rather difficult, given how such models are trying to predict rare events (i.e., code-switched occurrences) within limited datasets characterised by some degree of sample bias (which is often the case with corpora of low-resource languages). For these reasons, relying on parametric models is at best technically difficult (i.e., computationally intensive) and at the worst risky in terms of prediction and interpretation. In this study, we address these two challenges to investigate to what degree morphosyntactic, discourse, cognitive/psycholinguistic, and sociocultural factors jointly affect the choice of a bilingual speaker to code-switch, in a diglossic environment using the predictive modelling technique of Random Forests, which we apply to an annotated dataset from the TuniCo corpus (Moerth *et al*., 2014, 2017) and which is better suited to the otherwise statistically difficult nature of such corpus data. In the next section, we briefly survey previous work on code-switching from different sub-fields and theories of linguistics with an eye to identifying the factors that, ideally at least, multi-factorial studies of CS could include.

2. Factors affecting code-switching

2.1 Morphosyntactic factors

By far the most influential theoretical notions regarding grammatical constraints of CS are (*i*) Congruence and (*ii*) the Matrix Language Frame (MLF). Congruence (Sebba, 1998, 2009; and Deuchar, 2005) is the idea that within a potential CS window, the grammatical categories and the word classes of different languages are equivalent but hierarchically asymmetrical. In other words, the dominant language acts as the matrix language (ML) and the secondary language provides the embedded elements. There are two equivalence paradigms:

- **Paradigmatic similarity** between grammatical categories (i.e., the code-switched elements have to be compatible grammatically with other elements inter-sententially); and,
- **Syntagmatic similarity** between word order (i.e., the ML acts as a morphosyntactic frame into which the switched elements are inserted, and thus the word order of the ML has to be followed).

Hence, Sebba (1998, 2009) found that when both paradigmatic and syntagmatic congruences are met, then CS is facilitated, when neither are present, then CS is blocked, and when only one congruence is present, then CS is possible but restricted. The Matrix Language Frame (Myers-Scotton, 1995; Jake *et al.*, 2002; Myers-Scotton and Jake, 2009; and Deuchar *et al.*, 2017) specifies more constraints about the asymmetry between the ML and the embedded language (EL). The theory posits, as for congruence above, that the two languages are asymmetrical where ML systematically dominates, but it adds two over-arching principles:

- The **System Morpheme Principal**. In mixed constituents, system morphemes (function words) are mainly selected from the ML whereas content morphemes are selected from the EL (unless they belong to an EL island). System morphemes are prototypically quantifiers, specifiers and inflectional morphemes. Content morphemes prototypically assign or receive (discourse) 'theta-roles' (e.g., verbs, prepositions, descriptive adjectives and complementisers).
- The **Morpheme Order Principal**. The ML dictates order in mixed constituents.

Crucially, prominent CS researchers lately argued in favour of viewing the idea of 'constraints' governing code-switching as rather general tendencies (Poplack, 2001). In her recent position paper, Deuchar (2020: 16) suggested that 'future research should help us discover the relative role of external and internal factors as well as community norms in accounting for these patterns'. Although she highlights the importance of focussing on the 'invariant' patterns in CS behaviour, Deuchar hopes for a more 'comprehensive' scope that would include variability.

2.2 Cognitive and discourse features of CS

2.2.1 Cognitive processing

CS has been linked with behavioural and neurological costs (Costa and Santesteban, 2004; Gollan and Ferreira, 2009; Hell *et al.*, 2015, 2018; and Verreyt *et al.*, 2016). However, most researchers assume that, from a cognitive perspective, elements selected from EL are (nearly) equivalent to their potential counterparts from ML, equating this equivalence to synonymy in monolingual settings (Sridhar and Sridhar, 1980; Gollan and Ferreira, 2009; and Kutas *et al.*, 2009). Gollan and Ferreira further argued that a speaker will simply choose the first word that comes to their mind, regardless of the language. Hence, cognitive processing alone would lead to selecting the shorter and/or most frequent word (Heredia and Altarriba, 2001). Others argued that bilingual speakers do not access their lexicon symmetrically. For instance, Marian (2009) claimed that nouns are stored within a shared system across languages while verbs or other words are not. Accordingly, nouns are more likely to be code-switched as they are more 'portable', followed by verbs and then other parts-of-speech.

2.2.2 CS and prosody

Despite the paucity of work connecting CS and prosody, the available literature uncovered the existence of certain phonetic cues signaling forthcoming switches (e.g., reduction in speech rate) (Fricke *et al.*, 2016), a different prosodic contour between CS and unilingual speech (Piccinini and Garellek, 2014; and Shen *et al.*, 2020), and prosodic distancing (Torres Cacoullos, 2020). Furthermore, Shenk (2006) argued that the prosodic and discourse structure are the most important factor in predicting the occurrence of CS. She found (in a one-hour corpus of Spanish–English) that CS elements tend to occur at intonation units (IU) boundaries, which have been theorised to correspond to speakers' cognitive processing boundaries (Chafe, 1994).

2.2.3 Predictability

Myslín and Levy (2015) found that following part-of-speech, unpredictability of meaning was the second most explanatory variable in their model. They were able to measure predictability experimentally by having access to the speakers in their corpus and to the community. They determined that speakers tend to produce less predictable words not in L1, rather than the opposite, presumably in an effort to mark important information and invite the listener to pay special attention to it.

2.2.4 Priming and listener accommodation

As shown by a number of classic studies (e.g., Weiner and Labov [1983] and Bock [1986]) and recent ones (e.g., Gries [2005] and Hartsuiker *et al.* [2016]), having processed a certain syntactic structure (because they comprehended or produced it themselves) speakers are more likely to produce it again. In addition, it has been demonstrated that mimicking others' behaviour acts as a social-affiliation-and-solidarity device (Baaren *et al.*, 2009; and Kavanagh and Winkielman, 2016), and Myslín and Levy (2015) found that speakers tend to code-switch to accommodate other participants.

2.3 Sociocultural factors

Poplack (1980), Treffers-Daller (1992), Haust (1995) and Walters (2011) found variation in the number and/or the type of CS according to the gender of the speaker. Walters focussed specifically on CS in Tunisia and argued that the use of French is 'gendered' and dependent on the education level: women and more educated speakers are more likely to code-switch.

2.4 CS or Lexical Borrowing?

Early on, Poplack and Meechan (1998: 127) pointed out that distinguishing CS from Lexical Borrowing (LB) is 'at the heart of a fundamental disagreement among researchers about data'. And even now, it is arguably difficult to distinguish CS from LB (Deuchar, 2020), especially in a high-contact language situation, as for French and Tunisian Arabic (Manfredi *et al.*, 2015; and Lavender, 2017). Nonetheless, some scholars have argued that we can structurally distinguish CS from LB, with the latter exhibiting (more) morphological and phonological integration (Bullock and Toribio, 2009). Others, like Poplack *et al.* (2020) and Myers-Scotton and Jake (2009) contended that only morphosyntactic integration is a reliable metric to distinguish CS from LB. However, we will not explore this distinction in what follows.

2.5 This paper

As mentioned above, whilst many of the above factors, or predictors, have been studied in smaller datasets or in mono-factorial settings – one factor/predictor at a time – there is a dearth of studies devoted to how multiple predictors co-influence CS, both on their own (i.e., as what, in a regression-modelling context, would be captured by multiple but separate main effects) and jointly (i.e., as what, in a regression-modelling context, would be captured by interactions of predictors). In fact, a mono-factorial perspective on a complex phenomenon runs the risk of reporting findings without taking into account things like Simpson's paradox (Blyth, 1972), where individual factors may appear to influence the outcome in certain directions but the effect can be reversed or even disappear when factors are combined. In the following section, we present the methodology we employed to help address this gap and identify which of the previous findings survive multi-factorial scrutiny. In addition, we will also go beyond much existing CS work by including a variety of more cognitive/psycholinguistic and discourse-functional predictors in our analysis.

3. Methodology

For this study, we used data from TuniCo (Moerth *et al.*, 2014, 2017). In Section 3.1, we provide a brief description of the Tunisian linguistic landscape. In Section 3.2, we describe our corpus and the data extraction and annotation procedures and, in Section 3.3, we present our statistical approach.

3.1 Linguistic, social and historical background

Despite the presence of Berber (Gabsi, 2011) and Judeo-Tunisian Arabic (Bar-Asher, 1996), Tunisia is an ethnically and linguistically homogeneous country, where 98 percent of Tunisians identify as Arabs and speak Tunisian Arabic (Walters, 2011). Although the picture drawn here seems rather simple, Tunisian Arabic co-exists with Modern Standard Arabic (MSA) and French, in a 'triglossic' relationship. After independence from France, many factors contributed to the prominence of French over MSA, including the lack of Arabic textbooks and trained instructors as well as the political choices made by the leadership at the time. Consequently, French remained the official language of instruction until the 1980s (Daoud, 2001). And even with the Arabisation reforms, Tunisian students learnt French early (at around eight years of age) and STEM subjects are still taught in French (beginning in high school).

Furthermore, the strong economic and historical ties with France made the main immigration destination[2] and made French cultural products available to generations of Tunisians. Combined with the importance of the tourism industry, one would expect French to be regarded as a prestigious language. Yet, this is not uniformly the case across the country and different communities. In fact, Walters (2011) reported that using French is rather frowned upon outside Tunis.[3] However, the speakers in our corpus are from Tunis and we should not expect any negative attitudes toward CS.

3.2 Corpus data and annotation

The TuniCo corpus was collected by Ines Dallaji and Ines Gabsi in 2013 and contains transcriptions of thirty hours of conversations and narrative sociolinguistic interviews. The speakers are from various socio-economic backgrounds, maximally thirty-five years of age, and grew up and still live in Tunis, all of which controls for dialectal and generational variation. The corpus is encoded according to the guidelines of the Text Encoding Initiative (TEIP5)[4] and contains 142,317 tokens (with 13,154 items /

---

[2] Of the Tunisian diaspora, 88 percent lives in France which in turn constitutes 10 percent of the population (Leaders, 2016).
[3] We would further argue that the prestige of French would correlate with the socio-economic status of speakers.
[4] See: https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html.

14 percent of the tokens being foreign words). Most of the Tunisian Arabic tokens are part-of-speech (POS) tagged through a combination of manual and automatic annotation. However, this approach generated mis-annotations and portmanteau tags (containing multiple tags for certain ambiguous words). Consequently, we relied on semi-automatic and manual annotations to correct and/or add the missing parts-of-speech.

The conversations in the corpus can be divided into four categories depending on the number of participants and whether the researchers collecting the data are participants. In order to analyse comparable conversations, we only retained the subset of tripartite conversations where the researchers are participants, given that they were the most attested type. The subset consists of eleven files containing 56,310 words produced by thirteen main speakers (including the interviewers) and sixteen secondary speakers taking part in the conversations for a limited time.[5] The subset contains 8,224 French words representing approximately 15 percent of the selected sub-corpus.

The data were retrieved and analysed using *R* (R Core Team, 2021). With regard to the compilation of the dataset, we used the XML structure of the corpus to extract each utterance, which corresponds to a turn in conversation, all the words, their parts-of-speech, their respective language, as well as several metadata elements (e.g., the speaker, the file number and the utterance number). Table 1 provides an overview of the distribution of tokens according

Conversation File	Tunisian Arabic	French	Total
Talking to an artist	5,021	3,471	8,492
Medina salesman	7,854	472	8,326
Rapper	7,137	636	7,773
Woman in cafe	7,582	142	7,724
Souq salesman 2	5,997	302	6,299
Student of architecture	4,598	1,491	6,089
Artist in cafe	4,063	829	4,892
Student of architecture 2	2,179	509	2,688
Artist and photographer	1,763	209	1,972
Souq salesman 1	1,179	44	1,223
Tunisian Canadian	713	119	832
Total	48,086	8,224	56,310

Table 1: Distribution of tokens according to the production language across the corpus.

[5] The conversations occurred in public spaces.

Variable	Variable levels
LANG (the language of the word; dependent variable)	FR, TN
WORDPOS (the position of the word within an utterance)	[0, 1]
LENGTH (the length of the word in phonemes)	[1, 17]
POS (the part-of-speech of the word)	ADJ, ADV, ART, CONJ, DISF, INTJ, INTER, N, NUM/ORD, PART, PREP, PRONs, PRON, V
POSPREV (the POS of the previous word)	NONE, ADJ, ADV, ART, CONJ, DISF, INTJ, INTER, N, NUM/ORD, PART, PREP, PRONs, PRON, V
POSFOLL (the POS of the following word)	NONE, ADJ, ADV, ART, CONJ, DISF, INTJ, INTER, N, NUM/ORD, PART, PREP, PRONs, PRON, V
LANGPREV (the language of the previous word)	None, FR, TN
MOMENTUM (the language momentum at the current word)	[–93, 25]
PRIMING (the number of CS elements in the previous utterance)	[–93, 24]
SURPRISAL (the surprisal of the word based on a trigram model)	[0.05, 19.9]
SPEAKER	The speaker label
FILE	The conversation file name

Table 2: Variables used in the annotation of the data and their levels/ranges.

to the production language across the corpus and Table 2 summarises the variables used to annotate the data as well as their respective levels, followed by a detailed description.

3.2.1 Morphosyntactic variables

To test both the notions of Congruence and MLF against the current corpus, we rely on the POS tagging of each word (POS), that of the previous word (POSPREV) and that of the following word (POSFOLL). However, since our corpus is divided into utterances, we cannot test previous intra-sentential findings. Nonetheless, to determine the dominant language at each word in an utterance, we include as predictors the language of the previous word (LANGPREV) as well as a predictor we call MOMENTUM. This variable represents the difference between French and Tunisian Arabic words from the beginning of the utterance up to the current word:

- A negative value indicates that more Tunisian Arabic than French words have been produced so far in the utterance: for example, if, at a certain point in the utterance, seven words were so far in Tunisian Arabic and two in French, this would be represented with a value of –5; in other words, the utterance at this point has a Tunisian-leaning MOMENTUM;
- If the utterance so far contained equally many Tunisian Arabic and French words, this would be represented with a value of 0; and,
- A positive value indicates that fewer Tunisian Arabic than French words have been produced so far in the utterance: for example, if, at a certain point in the utterance, seven words were so far in French and two in Tunisian Arabic, this would be represented with a value of +5; in other words, the utterance at this point has a French-leaning MOMENTUM.

3.2.2 Cognitive and discourse variables

In order to investigate the effects of cognitive/psycholinguistic as well as discourse-functional predictors, we added the following variables to our statistical analysis:

WORDPOS: we compute the word position as its normalised position within an utterance, given below, where $\boldsymbol{W}_{n|n}$ is the word number within an utterance and $\boldsymbol{N}_{w|u}$ is the total number of words in the utterance; thus, the second word in a four-word utterance would score a value of $\frac{1}{3}$:

$$WORDPOS = \begin{cases} 0, & if\ N_{w|u} = 1 \\ \dfrac{W_{n|u} - 1}{N_{w|u} - 1}, & else \end{cases}$$

LENGTH: the length of the word in phonemes.[6]

PRIMING: specifies the number of French words that occurred in the immediately preceding turn or utterance (regardless of who produced it):

- A negative value indicates that the previous utterance contained more Tunisian Arabic then French words: for example, if the previous utterance contained fifteen words in Tunisian and five words in French, this would correspond to a value of –10;

[6] The corpus compilers adapted a Deutsches Institut für Normung standard for the transliteration of the Arabic alphabet, where every sign corresponds to a sound.

- If the previous utterance contained equally many Tunisian Arabic and French words, this would be represented with a value of 0; and,
- A positive value indicates that the previous utterance contained fewer Tunisian Arabic than French words: for example, if the previous utterance contained fifteen words in French and five words in Tunisian, this would correspond to a value of +10.

SURPRISAL: following Hale (2001), Levy (2008) and Smith and Levy (2013), we operationalised the predictability of a given word using its surprisal. A low SURPRISAL score indicates whether, given the word's previous context, a word has a high probability of occurrence or not. The formula given below is used to compute the surprisal of a word Sw_{k+1} given its previous context:

$$Sw_{k+1} = -log_2 Pr(w_{k+1}|w_{k-1}, w_k)$$

To compute the probability for each word in the sample, we used SRILM toolkit (Stolcke, 2002) to train a trigram model on the held-out portion of the corpus (n = 38,038). We estimated the probability of an unseen n-gram using Chen and Goodman's (1998) modified Kneser-Ney smoothing with interpolation to obtain an estimate using the probability of a lower-order n-grams. We ran the trained model on the selected sub-corpus and obtained the probability for each word to occur, as the last word of a trigram.

3.2.3 Speaker-specific control variables

SPEAKER: the information about speakers provided are the name, the occupation and the gender of each speaker; thus, whatever is unique to this speaker can theoretically be captured in this predictor.

FILE: the file names are included as a variable to account for any possible variation across conversations; thus, whatever is unique to this conversation can theoretically be captured in this predictor.

3.3 Statistical evaluation

We first tried to fit a generalised linear mixed-effects regression model with the language of the word as the response variable. Perhaps unsurprisingly, the model never converged and the computer ran out of memory (64 GB), given that we were trying to model a class-imbalanced dependent variable with nearly 57,000 cases (8,824 in French + 48,086 in Tunisian Arabic). In our search for an alternative, we ultimately opted for the predictive modelling technique of Random Forests: not only did Muchlinski *et al.* (2016: 101) find that they 'offer superior predictive power compared to several forms of logistic regression', but, as per Oommen *et al.* (2011), Random Forests are often also superior when it comes to predicting a class-imbalanced response variable (i.e., one characterised by a very uneven distribution of its levels). Hence, like other corpus-linguistic studies (Tagliamonte and Baayen, 2012; Dilts, 2013; Bernaisch *et al.*, 2014; and Deshors and Gries, 2016, 2020), we, too, ultimately went with Random Forests.

A Random Forest (Breiman, 2001) is a tree-based machine learning algorithm that tries:

> [...] to identify structure in the relation(s) between a response and multiple predictors by determining how the dataset can be split up repeatedly into successively smaller groups (based on the values of the predictors) in such a way that each split leads to the currently best possible improvement in terms of classification accuracy [...] for the response variable.
>
> (Gries, 2021: 453)

A Random Forest extends this by adding two layers of randomness,[7] which de-correlates trees, helps identify the importance of predictors and their interactions to the predictions, avoids collinearity problems, and protects against over-fitting. We followed Gries's (2020, 2021) recommendations and included interactions between predictors. All modelling and extraction of numerical results have been performed using *R* (R Core Team, 2021) with the *randomForestSRC* (Ishwaran and Kogalur, 2007) and the *ggRandomForests* (Ehrlinger, 2016) packages. For our dataset, we fit a Random Forest with *ntree* = 2,000 trees, each tree fit on a randomly sampled with replacement subset of the data and *mtry* = nine randomly sampled predictors for each split; the values of these two hyper-parameters performed optimally in our explorations of the forest during the development stage.

4. Results

Gries (2021: Chapter 7) suggests that, to interpret a Random Forest's results, it is crucial to examine:

- The variable importance scores (VIMP), which reflect the absolute size of the effect of a predictor on the response; thus, in regression modelling, the equivalent of how far regression coefficients of (z-standardised) predictors are from 0 (in whatever direction); and,
- The partial dependence plots (PDP), which reflect the direction of the effect of each level of the predictor on the response; thus, in regression modelling, the equivalent of the signed coefficients of

[7] This is achieved by running different trees on bootstrapped samples and by using a randomly selected subset of predictors at every split in every tree.

predictor levels and what they imply about the level's effect on the response.

Due to the large class-imbalance problem, the baseline/no-information rate accuracy of our classification is already a high 85.4 percent, but our model performs significantly better with a 96 percent true prediction/out-of-bag accuracy ($p_{binomial\ test}$ = 0), and 98 percent as the out-of-bag Area Under Curve (AUC, the equivalent of the *C*-score in regression modelling). Figure 1 shows a plot of VIMP-values computed by randomly permuting each variable's values and comparing the prediction error to that of the observed values. A large VIMP-value indicates that the variable is important to obtain accurate predictions, and a value closer to 0 indicates that the variable contributes almost nothing.

Already, Figure 1 shows that LANGPREV, MOMENTUM, POS, LENGTH, WORDPOS and POSPREV have a relatively big effect on the forest's predictions (in that order), whereas SURPRISAL, FILE, POSFOLL, SPEAKER and PRIMING have much smaller VIMP-values and, therefore, hardly contribute to the accuracy of predictions; they could be considered as the equivalent of non-significant for the model. Regarding SURPRISAL, this was expected given the small dataset used to train the *n*-gram language model, and estimated probabilities correspond mostly to the unigrams probabilities. On the other hand, the low VIMP-values for SPEAKER and FILE show that there is little variation across files and between speakers. Accordingly, for the sample in hand, the sociocultural variables and the context of the conversation itself seem to have little influence on predicting the occurrence of French words.

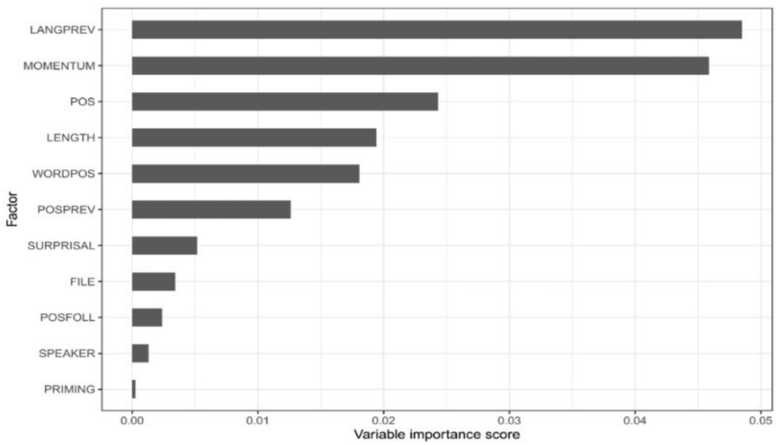

Figure 1: Variable importance scores in the model.

As for the more important variables, the higher VIMP-values indicate that these variables contribute to prediction accuracy, but, as per Gries (2020, 2021), it is important to keep in mind that Random Forests do capture

Interaction	VIMP-1	VIMP-2	Paired	Additive	Association
MOMENTUM:WORDPOS	0.046	0.018	0.046	0.064	0.018
POS:LENGTH	0.024	0.019	0.029	0.044	0.014
LANGPREV:POSPREV	0.048	0.013	0.052	0.061	0.010
MOMENTUM:LENGTH	0.046	0.019	0.073	0.065	0.008
MOMENTUM:POS	0.046	0.024	0.074	0.070	0.004
MOMENTUM:POSPREV	0.046	0.013	0.062	0.059	0.004
LANGPREV:LENGTH	0.048	0.019	0.071	0.068	0.003
LANGPREV:POS	0.048	0.024	0.076	0.073	0.003

Table 3: Overview of the highest association values.

interactions and we should avoid interpreting VIMP-values mono-factorially without investigating possible interactions. To do so, we employ a joint-variable importance approach (Ishwaran, 2007), where the paired importance of each pair of variables is calculated, then subtracted from the sum of the variables' respective VIMP-values. Table 3 gives an overview of the paired association values for the (important) variables in the model where a large association between two variables reflects an interaction that is worth exploring if the univariate VIMP-value for each of the paired-variables is relatively large. We state 'worth exploring' because a high association value between two variables is not equivalent to 'the interaction is significant': it signals, rather, that the interaction should be investigated.

Eventually, it is the analyst's prerogative and responsibility to determine: (*i*) where to draw the line between 'high' and 'low' values (much like the choice of a significance threshold would be); and, (*ii*) if the interaction is of theoretical significance for the research questions asked. Thus, `POS:LENGTH` scored the second largest association value but the two univariate VIMP-values are low relative to the two largest VIMP-values. Accordingly, when investigating the interaction's PDP (see Figure 2) we notice that the association results are driven by certain data points that seem to be of little theoretical interest. Figure 2 shows the mean predicted probability of a word being produced in French on the *x*-axis for the combination of each part-of-speech and each of three word lengths (when attested for the POS in question) on the *y*-axis. In fact, the longer a word is, the more likely it is to be produced in French (regardless of part-of-speech). When the word is longer than eight phonemes, adjectives, adverbs, disfluencies and numerals/ordinals are predicted to slightly prefer French. But this is mainly a by-product of the fact that a word longer than eight phonemes is more likely to be French. This particular behaviour will be more salient and more interpretable in light of other interactions (see Section 4.1).

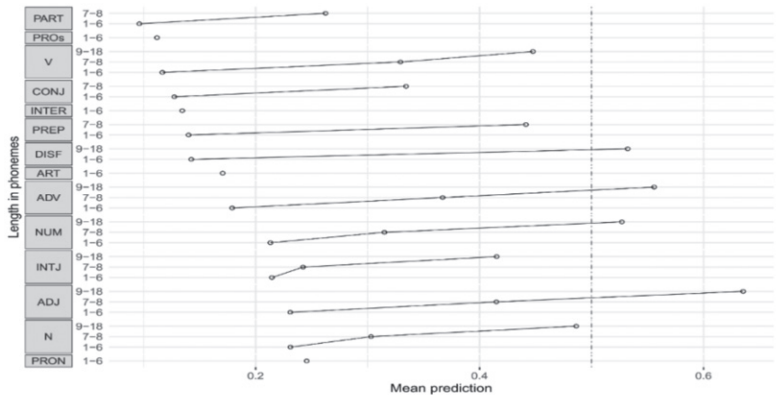

Figure 2: Conditional PDP of POS and LENGTH.

In what follows, we present seven interactions involving the two variables that scored the highest VIMP-values (LANGPREV and MOMENTUM) and the following four variables that have relatively large importance (POS, LENGTH, WORDPOS, and POSPREV):

- Four interactions involving MOMENTUM: MOMENTUM:WORDPOS in Section 4.1.1, MOMENTUM:LENGTH in Section 4.1.2, MOMENTUM:POS in Section 4.1.3, and MOMENTUM:POSPREV in Section 4.1.4;
- Three interactions involving LANGPREV: LANGPREV:POSPREV in Section 4.2.1, LANGPREV:POS in Section 4.2.2, and LANGPREV:LENGTH in Section 4.2.3.

4.1 Interactions with MOMENTUM

4.1.1 Interaction 1: MOMENTUM and WORDPOS

Figure 3 is a conditional PDP of the variable MOMENTUM and its interaction with WORDPOS. Both variables have been factorised.[8] To determine the bins (for these and other variables as needed), we struck an expositorily useful balance between (*i*) the results of classification trees (Hothorn *et al*., 2006), using the *R* package *partykit* (Hothorn and Zeileis, 2015) with the language of the word as the response and the variable of interest as the only predictor and (*ii*) intuitively understandable groupings/bins within each plot. Predictions can take values between 0 and 1 with values closer to 0 and 1 predicting

[8] The practice of exploring interactions of two numeric predictors by factorising at least one of them is widely used in regression modelling: see, for example, the package effects (Fox and Weisberg's [2018] regression textbook).

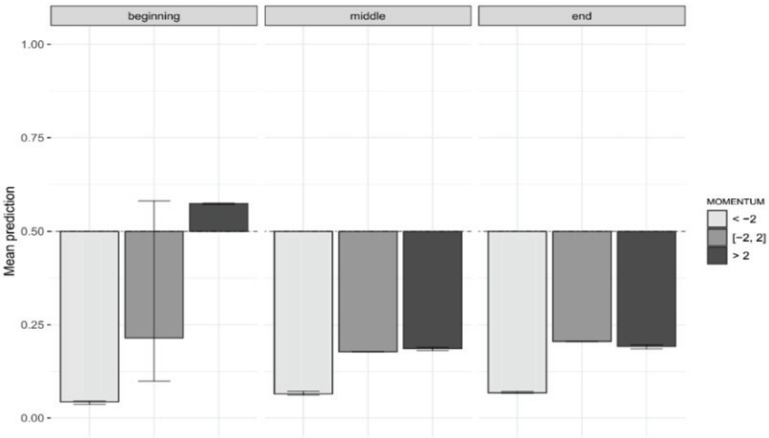

Figure 3: Conditional PDP of MOMENTUM and WORDPOS.

the occurrence of a word in Tunisian Arabic and French, respectively. Given the high value of the AUC, we should be confident that 0.5 is a good cut-off point for converting predicted probabilities into predicted languages. As a reminder, negative MOMENTUM values correspond to points in the utterance dominated so far by Tunisian Arabic and *vice versa*. Each bar in Figure 3 corresponds to the mean prediction for each class of the response, with the error bars representing the range of predictions giving rise to that mean.

Figure 3 shows that, in the middle and the end of an utterance (the two right panels), regardless of MOMENTUM, there is a higher chance of a Tunisian Arabic word to occur (especially in a Tunisian-leaning MOMENTUM) – that is, if a sizable portion of the utterance already was in Tunisian Arabic, speakers are less likely to switch. Fittingly, at the beginning of an utterance (the left panel), the model predicts that speakers tend to stick with the language they started with – that is, when MOMENTUM leans towards Tunisian in the beginning of an utterance, there is a significant chance that words produced within that stretch are Tunisian Arabic; and *vice versa*, if MOMENTUM leans towards French, there is a relatively high likelihood of seeing French words at the beginning of an utterance. Where none of the two languages seem to dominate (i.e., MOMENTUM values close to 0), the predictions tend to favour Tunisian Arabic but with a high degree of uncertainty.

4.1.2 Interaction 2: MOMENTUM and LENGTH

As mentioned above, predictions concerning LENGTH are driven by the fact that longer words generally tend to be French. Nonetheless, Figure 4 shows the interaction of LENGTH and MOMENTUM and despite the large predictions range, the plot is worth some attention. For short words (the left panel), the

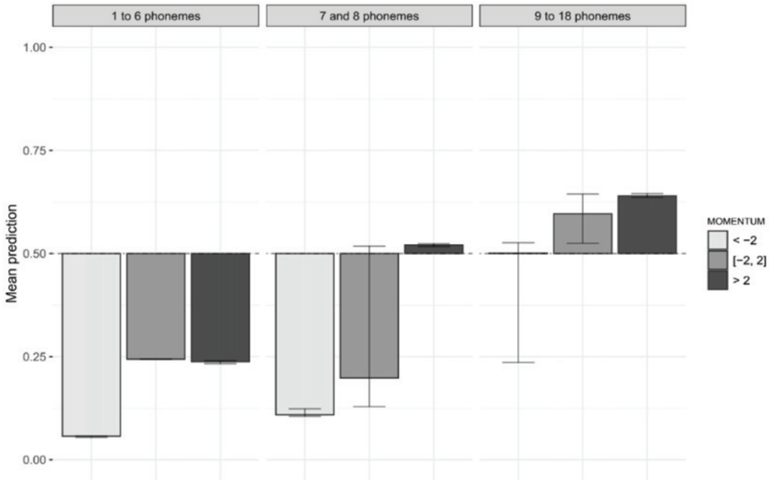

Figure 4: Conditional PDP of MOMENTUM and LENGTH.

model prefers Tunisian Arabic regardless of MOMENTUM (with a slightly lower probability if MOMENTUM is French-leaning or 'neutral'). For long words (the right panel), the reverse tendency can be observed. The predicted language is French regardless of MOMENTUM, but with a higher degree of uncertainty (except for French-leaning MOMENTUM). Last but not least, for intermediately long words (the middle panel) we find that the model predicts them to be non-switched (i.e., produced in Tunisian Arabic) in a Tunisian-leaning or neutral MOMENTUM (although notice the span of predictions' range) and in French in a French-leaning MOMENTUM.

4.1.3 Interaction 3: MOMENTUM and POS

Moving to the interaction between MOMENTUM and POS represented in Figures 5 and 6, which divide the results up by grouping together similarly behaving POS into POS with invariable behaviour (i.e., corresponding predictions do not change as a function of MOMENTUM) in Figure 5 and POS whose behaviour exhibits some variation in Figure 6. Both figures show conditional PDPs, where the predicted probabilities of a word being in French are on the x-axis, and every shade of grey represents a MOMENTUM interval. Examining both graphs, we can see that, generally, when the dominant language is Tunisian Arabic (i.e., negative MOMENTUM in light grey), the probability of a French word occurring is low regardless of the word's POS.

However, when contrasting Figures 5 and 6, we can see that when MOMENTUM is either neutral or French-leaning, prepositions, interjections, conjunctions, verbs and disfluencies are resistant to the change in MOMENTUM and still produced in Tunisian Arabic in neutral or French-dominated stretches

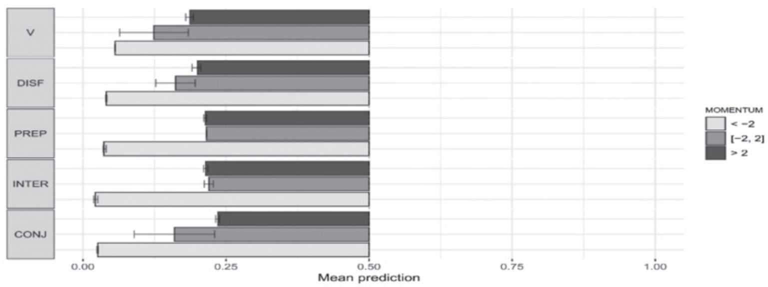

Figure 5: Conditional PDP of MOMENTUM and POS. Invariable parts-of-speech.

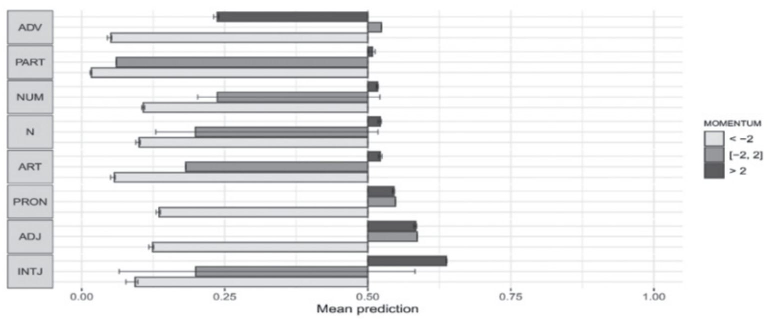

Figure 6: Conditional PDP of MOMENTUM and POS. Variable parts-of-speech.

of talk, whereas a number of other parts-of-speech do follow the French-leaning MOMENTUM and the probability of producing them in French becomes higher in French dominated points in the utterance:

- *N*: Nouns are predicted to occur in French when MOMENTUM is positive. They are, however, predicted to be Tunisian Arabic in a neutral MOMENTUM but with a high degree of uncertainty. Hence, the occurrence of French nouns is very likely in a stretch of talk dominated by French.
- *PART*: Particles have a high probability of being produced in French, in a French-leaning MOMENTUM. Looking more closely at those specific particles, we notice that 63 percent of them were response particles (e.g., *oui* ['yes'] or *non* ['no']) and negation particles (e.g., *ne*, *pas*, *jamais*…).
- *NUM*: Numerals/ordinals are predicted to occur in French if MOMENTUM is positive, and in Tunisian Arabic (with less confidence) if MOMENTUM is negative. It is relevant to note here that Tunisians tend to use French numerals, which may be an artifact of the linguistic history of Tunisia (see Section 3.1).

- *ART*: Articles are predicted to be produced in French, in a French-leaning MOMENTUM. Given that nouns behave similarly, and that they head NPs, this result is, thus, not surprising.
- *ADJ*: Adjectives are likely to be French in a French-leaning and neutral MOMENTUM. Similar to articles, adjectives mostly occur in NPs and it should not come as a surprise that they mirror the behaviour of their phrase's head.
- *PRON*: Pronouns are predicted to occur in French in both a neutral and a French-leaning MOMENTUM. Similarly, although the relationship of pronouns to nouns is not necessarily syntactic (as in occurring in the same phrase), but rather that of reference, it exhibits the same hierarchical structure, where pronouns are dependent on nouns. Thus, pronouns are likely to be switched when nouns tend to be switched.
- *ADV*: Adverbs exhibit a unique behaviour. They are likely to be switched only in neutral MOMENTUM. Inspecting the training data, we noticed that these occurrences mainly correspond to adverbs occurring at the beginning of an utterance (which should remind us of the results in Section 4.1.1). Accordingly, the speakers in the sample seem to start their turns with French adverbs. This correlates with the first author's intuition that Tunisians tend to use certain French adverbs as discourse connectors or sentence modifiers (e.g., *bien-sûr* ['of course'], *déjà* ['already'] or *normalement* ['usually']).
- *INTJ*: interjections are predicted to occur in the language of the MOMENTUM they are produced in. However, whilst interjections are traditionally seen as independent syntactically, they still hold a relationship with their discursive and interactional context (Dingemanse, 2017) and, thus, should reasonably be expected to occur in French in a French-leaning MOMENTUM.

4.1.4 Interaction 4: MOMENTUM and POSPREV

Figures 7 and 8 are conditional PDPs of the interaction MOMENTUM:POSPREV. They respectively group together the invariable POS and the variable POS (from the MOMENTUM point of view). All in all, this is just a confirmation of the results presented in the previous section. First, we see variation in predictions only when the previous part-of-speech is either an article, a numeral/ordinal, an interjection or a pronoun. All other POS precede a Tunisian Arabic word regardless of MOMENTUM. Second, in Figure 8, we see that articles, numerals and pronouns are likely to precede a French word in French-leaning, and neutral MOMENTUM (although notice the range of predictions for the latter). This is yet another indication of the 'supremacy' of nouns over their dependents when it comes to the code integrity of the NP. In other words, when a given stretch of talk is dominated by French, noun modifiers are likely to

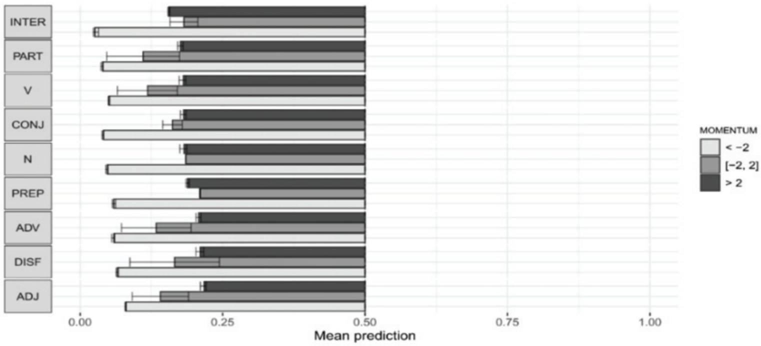

Figure 7: Conditional PDP of `MOMENTUM` and `POSPREV`. Invariable parts-of-speech.

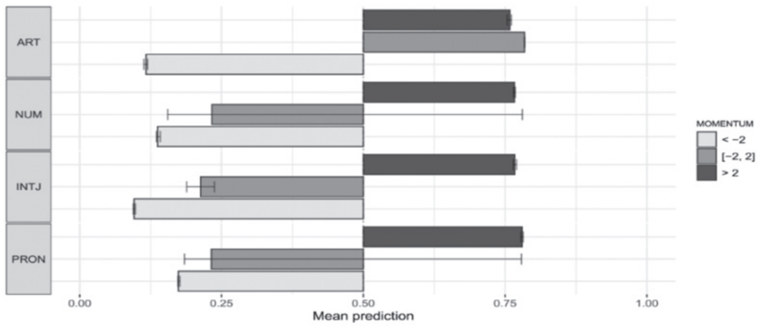

Figure 8: Conditional PDP of `MOMENTUM` and `POSPREV`. Variable parts-of-speech.

be produced in the same language as the noun they modify (which in turn is likely to be produced in French, as outlined previously). The same logic applies to pronouns: although they are not syntactically dependent on nouns, speakers are likely to produce them in French in a French-leaning `MOMENTUM`, perhaps in an effort to reduce the 'cognitive distance' between a reference and an antecedent. Finally, interjections display the same behaviour as previously, where they tend to be produced in French when the immediate context is leaning towards a French `MOMENTUM`.

4.2 Interactions with `LANGPREV`

4.2.1 Interaction 1: `LANGPREV` and `POSPREV`

Figure 9 and 10 are similar to the previous figures, where the predicted probabilities of `LANG`: 'French' are on the *x*-axis, the different shades of grey

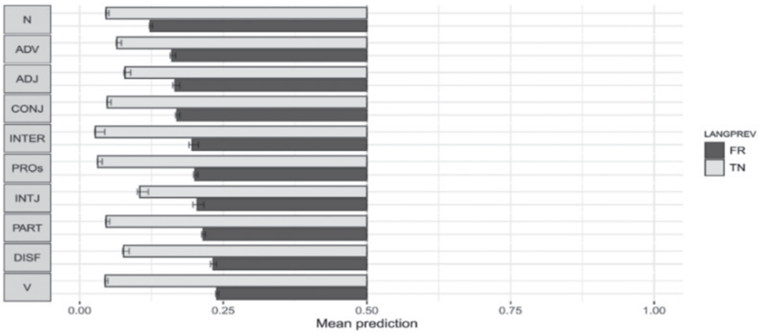

Figure 9: Conditional PDP of LANGPREV and POSPREV. Invariable parts-of-speech.

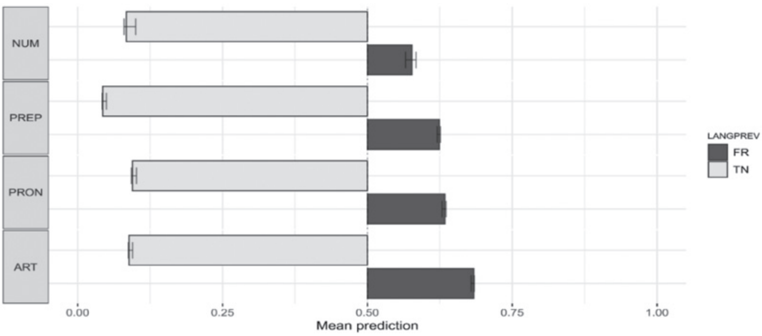

Figure 10: Conditional PDP of LANGPREV and POSPREV. Variable parts-of-speech.

bars indicate the previous language, and the parts-of-speech in each panel represent the POS of the previous word. Both figures show that, when the previous language is Tunisian Arabic, the likelihood of seeing a switched element occurring is negligible. In addition, Figure 9 shows that most POSs are likely to occur in Tunisian Arabic regardless of the previous language – and this is not surprising, given the dataset's imbalance with regard to the variable LANG. However, Figure 10 presents more variation. In fact, words preceded by French numerals/ordinals, prepositions, pronouns and articles tend to be French themselves. This result partially confirms the previous results (see Section 4.1.3) in that dependent POS – especially those dependent on nouns – are more likely to be produced in French within stretches of talk dominated by French. Hence, speakers seem attuned to maintaining (*i*) the phrase code integrity (e.g., in the cases of articles and numerals), (*ii*) the discourse code continuity (e.g., in the case of interjections) or (*iii*) reducing the 'code distance' and the cognitive distance between a referent and a reference (e.g., in the case of pronouns).

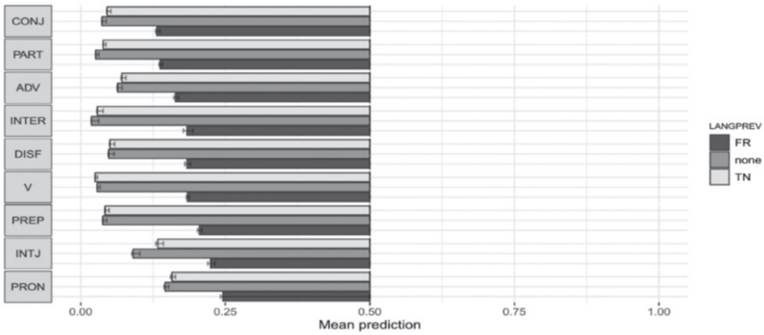

Figure 11: Conditional PDP of LANGPREV and POS. Invariable parts-of-speech.

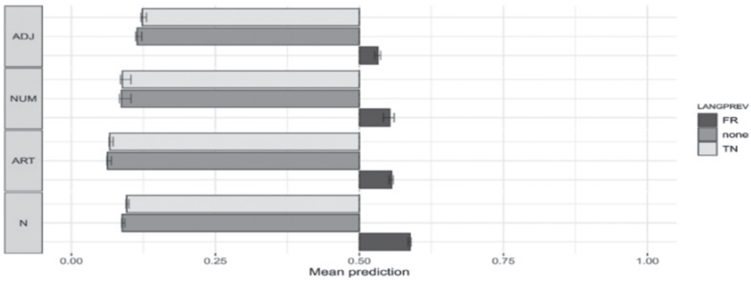

Figure 12: Conditional PDP of LANGPREV and POS. Variable parts-of-speech.

4.2.2 Interaction 2: LANGPREV and POS

The picture here is very similar to the previous interaction. First, Figure 11 (below) re-confirms that most POSs are more likely to occur in Tunisian Arabic regardless of the previous language. Figure 12, on the other hand, shows that nouns, numerals/ordinals, articles and adjectives are predicted to occur in French, if the language of the previous word is French. Hence, this implies that noun phrases are likely to be switched as a unit, with nouns being the most likely to be switched. The latter are, after all, the head of their phrases and seeing their dependents being switched – when they are themselves switched – contributes to the phrase code integrity.

4.2.3 Interaction 3: LANGPREV and LENGTH

Finally, the last interaction of interest in the model concerns the language of the previous word (LANGPREV) and the length of the current word (LENGTH). Figure 13 is divided into three panels according to the previous language;

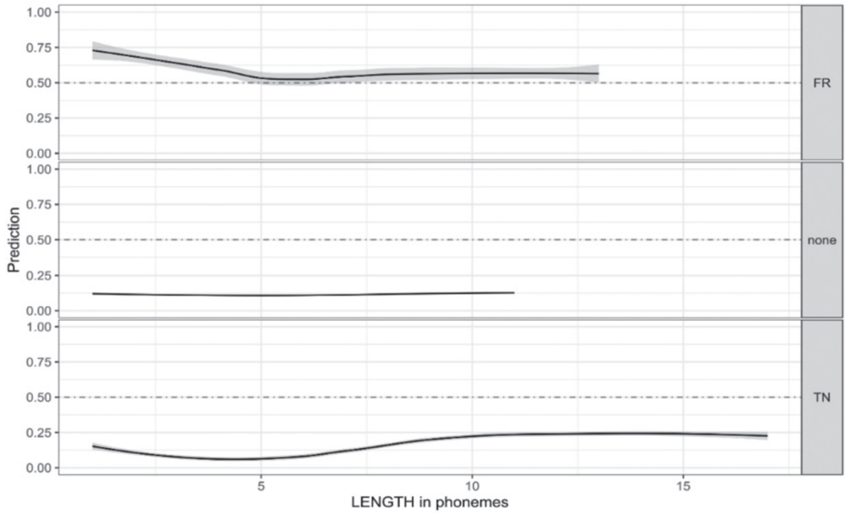

Figure 13: Conditional PDP of LANGPREV and LENGTH.

predicted probabilities are on the *y*-axis and the lengths of words in phonemes on the *x*-axis.[9] When LANGPREV is either 'none' (i.e., the word is located at the beginning of an utterance) or 'Tunisian', the predicted probability of producing a CS element is low. It is worth noting that the probability gets even lower for words of two to six phonemes in length when LANGPREV is 'Tunisian'. These words are often grammatical in nature and, in the light of the results presented so far, the dip in the graph is consistent with the idea that function words tend to keep the same code as their immediately preceding context. More interestingly, the upper panel is concerned with words occurring after a French word. On the one hand, the former words tend to be produced in French themselves. On the other hand, the predictions' line gets closer to the cut-off value of 0.5 as the word gets longer. Hence, speakers tend to produce French words immediately after another French word, but if the current word is projected to be relatively long (> 5 phonemes), its likelihood of being in French is lower; and in fact there is almost an equal chance for it to be produced in French or in Tunisian Arabic.

5. Discussion and conclusions

As discussed above, with this study, we hope to have achieved several goals: we wanted to take a widely studied phenomenon – code-switching – and offer

[9] Contrary to previous figures that included LENGTH, here, it is not factorised. The reason is that, for this interaction, we already had a categorical variable and a continuous variable, and we did not feel the need to simplify the data for expository reasons.

a range of perspectives to it that are so far very much under-represented in such work. More specifically, we wanted to offer a study that:

- Is corpus-linguistic in nature (using a diglossic corpus) and is, despite the low-resource nature of L1, based on a much larger amount of data than most previous CS work;
- Is multi-factorial in nature and, thus, able to study the effect of multiple predictors both separately and simultaneously (as in main effects) and jointly and interactively (as in interactions);
- Covers a wider range of predictors than some previous work by including structural but also cognitive/psycholinguistic and discourse-functional predictors;
- Employs not only powerful predictive modelling methods (Random Forests), which are useful for data that make more 'traditional' modelling methods difficult to apply (e.g., scarcity of data, absence of reference corpora and rare-event modelling), but also goes beyond the usual application of such methods to the study of interactions (which, based on [Gries, 2020], is a rather new development in corpus-linguistic circles); and,
- Owing to all of the above (and with all due humility):
  - Offers the field a range of methodological proposals and examples of how to push CS research towards new boundaries. In addition, we also make a plea for a more general integration of (more) machine learning techniques and (more) computational and Natural Language Processing (NLP) tools into corpus linguistics; and,
  - Allows us to uncover patterns in speakers' usually unconscious CS behaviour that have not been discovered before.

Whilst the methodological innovations, in a sense, 'speak for themselves' in how they provided new results and perspectives on the data, we now turn to the linguistic/conceptual findings. Most theories relating to the morphosyntactic features of CS rely on determining the intra-sentential location and the syntactic hierarchical structure in which the CS elements occur. Although our sample consists of utterances, which in turn comprise different number of sentences, the two most important monofactorial predictors in our model, namely `LANGPREV` and `MOMENTUM` and especially their interactions with other predictors, allow us to have two different perspectives on the syntactic and code context in which a word occurs. `LANGPREV` provides a localised window comprising a word and its immediately preceding context, whereas `MOMENTUM` allows for a larger but fuzzier window of the code momentum in which a word occurs. Thus, our results show that the morphosyntactic factors constraining CS are in constant interaction with the code choices speakers made in their previous stretch of talk. Specifically, nouns are by far the most switched lexical POS when the adjacent context is at least partially in

French; this seems to confirm previous findings (e.g., Marian [2009]). But our results also show that nouns occurring in stretches of talk relatively dominated by French lexemes tend not only to be in French, but also to affect the lexemes whose POS are governed syntactically or semantically by nouns (i.e., articles, adjectives, pronouns and prepositions). In other words, when the code momentum of the utterance favours code-switching (i.e., a French-leaning momentum), not only nouns but the NP (and to a certain degree the PP), as a whole, seems to be a prime location for code-switching to occur. Hence, within these stretches of talk, the competition between the two languages is constrained by the need to maintain the code integrity of the phrase, but not of all types of phrases equally, rather or especially NPs and PPs (which can be argued to be syntactically and/or semantically dependent on the noun). Verbs, on the other hand, and despite being considered amenable to CS (Myers-Scotton, 1995; Jake *et al.*, 2002; Marian, 2009), are rather resilient even when the context is dominated by French. Accordingly, our results lend the existing literature some weight but add some layers of nuance in the context of CS in Tunisian Arabic by showing that confining the focus within the sentence boundaries can lead to overlooking the behaviour of what traditionally has been considered to be at the fringe of the sentence (i.e., interjections) (Dingemanse, 2017). This study shows that preserving code integrity goes beyond the phrase and encompasses the discourse level. Interjections are a case in point as they tend to follow the code momentum in which they occur (i.e., interjections are produced in the language of their immediate context in the conversation). That being said, annotating for sentence boundaries and dependencies would add more granularity to our model and will be included in the further development of the study.[10]

Moreover, when speakers in our sample code-switch they seem to be not only attentive to the discourse-level code integrity, but also to the cognitive load they impose on themselves and/or their interlocutors. First, speakers are more likely to code-switch at the beginning of an utterance and consistently continue to do so (at least for the first third of their utterance), but are less likely to do so at the middle and the end. Thus, the two constraints of (*i*) preserving discourse code integrity as much as possible and (*ii*) minimising cognitive processing load are in competition here. As for (*i*), a speaker could be expected to continue code-switching if they started to do so at a point in their utterance; but as they go further into the stretch of talk, the likelihood of code-switching decreases. This tendency in our data correlates with Verreyt *et al.*'s (2016) findings. Their study revealed that for bilinguals who frequently code-switch, 'the frequent simultaneous activation between strong lexical representations of different languages causes competition and necessitates the bilinguals to engage their executive control mechanism to select representations in the target language, and inhibit the non-target

[10] This includes fine-tuning a pre-trained transformer-based multilingual machine learning model on the current dataset in the hope of achieving a better accuracy in POS and dependencies tagging, and sentence boundary annotation.

language' (Verreyt *et al.*, 2016: 188), thus leading to (*ii*). The competition between (*i*) and (*ii*) makes speakers less likely to code-switch at the middle or end of their turn, given the executive control required. This is also apparent when we look at the previous context of a word in conjunction with its length in phonemes: when the immediate previous context is French, the likelihood of continuing in French is higher when the planned word is shorter. In other words, in our data, if the previous lexeme is in French and the planned lexeme is longer than four phonemes, the chance of the planned lexeme to be French or Tunisian Arabic is about equal; this seems to correlate with the fact that speakers are attuned to the cognitive control required for code-switching in conversation.

Finally, our model revealed that priming, the predictability of a word and the controls of speaker and conversation (which, at a very coarse level, include sociocultural aspects of the speakers) have little effect on predicting CS. However, we have to introduce a number of caveats and consider how to address them in the future. First, the units between which we measured `PRIMING` are utterances, which are often relatively large and have no or little structural/psycholinguistic relevance (compared to sentences, IUs or clauses). We expect to see a bigger effect size if utterances are to be segmented at sentence boundaries. Regarding surprisal/predictability, the absence of comparable reference corpora, especially for the ML/L1, limited our surprisal measure to a relatively small (and imbalanced) dataset and should be interpreted with extreme caution. Therefore, we plan to take advantage of the advances made in synthetic data generation to overcome the class imbalance by generating synthetic data samples for the minority class: for example, the ADASYN (He *et al.*, 2008) and SMOTE algorithm (Fernandez *et al.*, 2018). Furthermore, the scarcity of the data confined the analysis to a limited number of speakers about whom minimal information has been provided. Hence, the apparent non-importance of the variable `SPEAKER` in the model should also be taken with a pinch of salt; `SPEAKER` might just be too indirect a proxy for more granular social and/or sociolinguistic/-cultural variables. Recall that the dataset used for the analysis is a subset of the TuniCo corpus and we hope to include the entire corpus in a future analysis.

Last but not least, the high contact of Tunisian Arabic with European languages, requires distinguishing code-switching from lexical-borrowing. This is manifest in the seemingly odd behaviour of adverbs occurring in neutral momentum. A closer inspection revealed that these adverbs (e.g., *bien-sûr* ['of course'], *bon* ['well'], *déjà* ['already'] and *normalement* ['usually']) can be argued to be, rather, loans. One strategy to address this shortcoming would be trying to differentiate LB from CS by determining their degree of morphological and/or phonological integration (Bullock and Toribio, 2009). This can be accomplished by training a language model to generate phonotactic statistics calculated across the corpus; it might then be possible to set/determine a threshold value that allows us to differentiate CS from LB.

To summarise, this study emphasises the importance of investigating complex linguistic phenomena, such as CS in conversation, through a multi-factorial/predictive modelling lens. Such phenomena are often affected by a constellation of competing as well as interacting factors that can easily be missed when one tackles CS from a mono-factorial perspective. Despite the apparent hurdles that CS and low-resource languages corpora present, we hope that our analysis showcased that extending the toolbox of corpus linguistics to machine learning techniques, whilst not offering the pure and formal hypothesis-testing power many corpus linguists associate with regression models, can still be a more than adequate tool to overcome the inherent challenges posed by limited, biased and noisy observational data.

References

Baaren, R. Van, L. Janssen, T.L. Chartrand and A. Dijksterhuis. 2009. 'Where is the love? The social aspects of mimicry', Philosophical Transactions of the Royal Society B: Biological Sciences 364 (1528), pp. 2381–9.

Bar-Asher, M. 1996. 'La recherche sur les parlers judéo-arabes modernes du Maghreb: État de la question' ('Research on the modern Judeo–Arabic dialects of the Maghreb: a review'), Histoire Épistémologie Langage 18 (1), pp. 167–77.

Bernaisch, T., St.Th. Gries and J. Mukherjee. 2014. 'The dative alternation in South Asian English(es): modelling predictors and predicting prototypes', English World-Wide: A Journal of Varieties of English 35 (1), pp. 7–31.

Blyth, C.R. 1972. 'On Simpson's Paradox and the Sure-Thing Principle', Journal of the American Statistical Association 67 (338), pp. 364–6.

Bock, J.K. 1986. 'Syntactic persistence in language production', Cognitive Psychology 18 (3), pp. 355–87.

Breiman, L. 2001. 'Random Forests', Machine Learning 45 (1), pp. 5–32.

Bullock, B.E. and A.J. Toribio. 2009. 'Themes in the study of code-switching' in B.E. Bullock and A.J. Toribio (eds) The Cambridge Handbook of Linguistic Code-switching, pp. 1–18. Cambridge: Cambridge University Press.

Carter, D., P. Davies, M.D.C.P. Couto and M. Deuchar. 2010. 'A corpus-based analysis of codeswitching patterns in bilingual communities', Revista Española de Lingüística.

Chafe, W.L. 1994. Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing. Chicago: University of Chicago Press.

Chen, S.F. and J. Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling.

Costa, A. and M. Santesteban. 2004. 'Lexical access in bilingual speech production: evidence from language switching in highly proficient bilinguals and L2 learners', Journal of Memory and Language 50 (4), pp. 491–511.

Dallaji, I., I. Gabsi, K. Mörth, S. Procházka and O. Siam. 2017. Linguistic Dynamics in the Greater Tunis Area: A Corpus-based Approach (TUNICO). Available online at: https://tunico.acdh.oeaw.ac.at.

Daoud, M. 2001. 'The language situation in Tunisia', Current Issues in Language Planning 2 (1), pp. 1–52.

Deshors, S.C. and St.Th. Gries. 2016. 'Profiling verb complementation constructions across New Englishes: a two-step random forests analysis of *ing* vs. *to* complements', International Journal of Corpus Linguistics 21 (2), pp. 192–218.

Deshors, S.C. and St.Th. Gries. 2020. 'Mandative subjunctive versus *should* in world Englishes: a new take on an old alternation', Corpora 15 (2), pp. 213–41.

Deuchar, M. 2005. 'Congruence and Welsh–English code-switching', Bilingualism: Language and Cognition 8 (3), pp. 255–69.

Deuchar, M. 2020. 'Code-switching in linguistics: a position paper', Languages 5 (2), p. 22.

Deuchar, M., P. Davies and K. Donnelly. 2017. Building and Using the Siarad Corpus: Bilingual Conversations in Welsh and English. Amsterdam: John Benjamins.

Dilts, P.C. 2013. Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed-Effects Regression. (PhD thesis.) Edmonton, Alberta: Department of Linguistics.

Dingemanse, M. 2017. 'On the margins of language: ideophones, interjections and dependencies in linguistic theory' in N.J. Enfield (ed.) Dependencies in Language, 195–203. Berlin: Language Science Press.

Du Bois, J.W. 1985. 'Competing motivations' in J. Haiman (ed.) Typological Studies in Language, pp. 343–65. (Volume 6.) Amsterdam: John Benjamins.

Ehrlinger, J. 2016. ggRandomForests: Visually Exploring Random Forests. Available online at: https://doi.org/10.48550/arXiv.1501.07196.

Ferguson, C.A. 1959. 'Diglossia', WORD 15 (2), pp. 325–40.

Fernandez, A., S. Garcia, F. Herrera and N.V. Chawla. 2018. 'SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary', Journal of Artificial Intelligence Research 61 (2018), pp. 863–905.

Fox, J. and S. Weisberg. 2018. An R Companion to Applied Regression. (Third edition.) Los Angeles: Sage.

Fricke, M., J.F. Kroll and P.E. Dussias. 2016. 'Phonetic variation in bilingual speech: a lens for studying the production–comprehension link', Journal of Memory and Language 89 (2016), pp. 110–37.

Gabsi, Z. 2011. 'Attrition and maintenance of the Berber language in Tunisia', International Journal of the Sociology of Language 2011 (211), pp. 135–64.

Gambäck, B. and A. Das. 2016. 'Comparing the level of code-switching in corpora' in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 1850–5. 23–28 May 2016. Portorož, Slovenia. European Language Resources Association (ELRA).

Gollan, T.H. and V.S. Ferreira. 2009. ‘Should I stay or should I switch? A cost–benefit analysis of voluntary language switching in young and aging bilinguals’, Journal of Experimental Psychology: Learning, Memory, and Cognition 35 (3), pp. 640–65.

Gries, St.Th. 2005. ‘Syntactic priming: a corpus-based approach’, Journal of Psycholinguistic Research 34 (4), pp. 365–99.

Gries, St.Th. 2020. ‘On classification trees and random forests in corpus linguistics: some words of caution and suggestions for improvement’, Corpus Linguistics and Linguistic Theory 16 (3), pp. 617–47.

Gries, St.Th. 2021. Statistics for Linguistics with R: A Practical Introduction. (Third revised edition.) Berlin: de Gruyter Mouton.

Hale, J. 2001. ‘A Probabilistic Earley Parser as a Psycholinguistic Model’, Second Meeting of the North American Chapter of the Association for Computational Linguistics. 2–7 June 2001. Pittsburgh, Pennsylvania. Association for Computational Linguistics.

Hartsuiker, R.J., S. Beerts, M. Loncke, T. Desmet and S. Bernolet. 2016. ‘Cross-linguistic structural priming in multilinguals: further evidence for shared syntax’, Journal of Memory and Language 90, pp. 14–30.

Haust, D. 1995. Codeswitching in Gambia: Eine soziolinguistische Untersuchung von Mandinka, Wolof und Englisch in Kontakt; with an English summary. Köln: Köppe.

He, H., Y. Bai, E.A. Garcia and S. Li. 2008. ‘ADASYN: adaptive synthetic sampling approach for imbalanced learning’. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–8. Hong Kong, China: IEEE.

van Hell, J.G., C.B. Fernandez, G.J. Kootstra, K.A. Litcofsky and C.Y. Ting. 2018. ‘Electrophysiological and experimental-behavioral approaches to the study of intra-sentential code-switching’, Linguistic Approaches to Bilingualism 8 (1), pp. 134–61.

van Hell, J.G., K.A. Litcofsky and C.Y. Ting. 2015. ‘Sentential code-switching: cognitive and neural approaches’ in J.W. Schwieter (ed.) The Cambridge Handbook of Bilingual Processing, pp. 459–82. Cambridge: Cambridge University Press.

Heredia, R.R. and J. Altarriba. 2001. ‘Bilingual language mixing: why do bilinguals code-switch?’ Current Directions in Psychological Science 10 (5), pp. 164–68.

Hothorn, T., K. Hornik and A. Zeileis. 2006. ‘Unbiased recursive partitioning: a conditional inference framework’, Journal of Computational and Graphical Statistics 15 (3), pp. 651–74.

Hothorn, T. and A. Zeileis. 2015. ‘Partykit: a modular toolkit for recursive partytioning in R’, Journal of Machine Learning Research 16 (118), pp. 3, 905–9.

Ishwaran, H. 2007. ‘Variable importance in binary regression trees and forests’, Electronic Journal of Statistics 1, pp. 519–37.

Ishwaran, H. and U.B. Kogalur. 2007. 'Random survival forests for R', R News 7 (2), pp. 25–31.

Jake, J.L., C. Myers-Scotton and S. Gross. 2002. 'Making a minimalist approach to codeswitching work: adding the Matrix Language', Bilingualism: Language and Cognition 5 (1), pp. 69–91.

Kavanagh, L.C. and P. Winkielman. 2016. 'The functionality of spontaneous mimicry and its influences on affiliation: an implicit socialization account', Frontiers in Psychology 7, 458.

Kutas, M., E. Moreno and N. Wicha. 2009. 'Code-switching and the brain' in B.E. Bullock and A.J. Toribio (eds) The Cambridge Handbook of Linguistic Code-switching, pp. 289–306. Cambridge: Cambridge University Press.

Lavender, J. 2017. 'Comparing the pragmatic function of code switching in oral conversation and in Twitter in bilingual speech from Valencia, Spain', Catalan Review 31, pp. 15–39.

Leaders. 2016. 'Ces Tunisiens dans le monde: Qui sont-ils? Où résident-ils?' Leaders. Retrieved from https://archive.wikiwix.com/cache/index2.php?url=https%3A%2F%2Fwww.leaders.com.tn%2Farticle%2F19702-ces-tunisiens-dans-le-monde-qui-sont-ils-ou-resident-ils#federation=archive.wikiwix.com

Levy, R. 2008. 'Expectation-based syntactic comprehension', Cognition 106 (3), pp. 1126–77.

Manfredi, S., M.-C. Simeone-Senelle and M. Tosco. 2015. 'Language contact, borrowing and codeswitching' in A. Mettouchi, M. Vanhove and D. Caubet (eds) Studies in Corpus Linguistics, pp. 283–308. (Volume 68.) Amsterdam: John Benjamins.

Marian, V. 2009. 7. 'Language interaction as a window into bilingual cognitive architecture' in L. Isurin, D. Winford and K. de Bot (eds) Studies in Bilingualism, pp. 161–85. (Volume 41.) Amsterdam: John Benjamins.

Moerth, K., S. Procházka and I. Dallaji. 2014. 'Laying the foundations for a diachronic dictionary of tunis arabic: a first glance at an evolving new language resource' in Proceedings of the XVI EURALEX International Congress: The User in Focus, pp. 377–87. Bolzano/Bozen. 15–19 July 2014. EURAC research.

Moerth, K., D. Schopper and O. Siam. 2017. 'Linking instead of lemmatising: enriching the tunico corpus with the dictionary of tunis arabic' in V. Ritt-Benmimoun (ed.) Tunisian and Libyan Arabic Dialects: Common trends – Recent Developments – Diachronic Aspects, pp. 218–38. Prensas de la Universidad de Zaragoza.

Muchlinski, D., D. Siroky, J. He and M. Kocher. 2016. 'Comparing Random Forest with logistic Regression for Predicting Class-Imbalanced Civil War Onset Data', Political Analysis 24 (1), pp. 87–103.

Myers-Scotton, C. 1995. Social Motivations for Codeswitching: Evidence from Africa. (First issued in paperback.) Oxford: Clarendon Press.

Myers-Scotton, C. and J. Jake. 2009. 'A universal model of code-switching and bilingual language processing and production' in B.E. Bullock and A.J.

Toribio (eds) The Cambridge Handbook of Linguistic Code-Switching, pp. 336–57. Cambridge: Cambridge University Press.

Myslín, M. and R. Levy. 2015. 'Code-switching and predictability of meaning in discourse', Language 91 (4), pp. 871–905.

Oommen, T., L.G. Baise and R.M. Vogel. 2011. 'Sampling bias and class imbalance in maximum-likelihood logistic regression', Mathematical Geosciences 43 (1), pp. 99–120.

Piccinini, P. and M. Garellek. 2014. 'Prosodic cues to monolingual versus code-switching sentences in English and Spanish' in Conference on Speech Prosody 2014. Dublin, Ireland. 20–23 May 2014.

Poplack, S. 1980. 'Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching', Linguistics 18 (7–8), pp. 581–618.

Poplack, S. 2001. 'Code-switching (linguistic)' in N.J. Smelser and B. Baltes (eds) International Encyclopedia of the Social and Behavioral Sciences, pp. 2062–5. Elsevier.

Poplack, S. and M. Meechan. 1998. 'Introduction: how languages fit together in codemixing', International Journal of Bilingualism 2 (2), pp. 127–38.

Poplack, S., S. Robillard, N. Dion and J. Paolillo. 2020. 'Revisiting phonetic integration in bilingual borrowing', Language 96 (1), pp. 126–59.

R Core Team. 2021. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Sebba, M. 1998. 'A congruence approach to the syntax of codeswitching', International Journal of Bilingualism 2 (1), pp. 1–19.

Sebba, M. 2009. 'On the notions of congruence and convergence in code-switching' in B.E. Bullock and A.J. Toribio (eds) The Cambridge Handbook of Linguistic Code-Switching, pp. 40–57. Cambridge: Cambridge University Press.

Shen, A., S. Gahl and K. Johnson. 2020. 'Didn't hear that coming: effects of withholding phonetic cues to code-switching', Bilingualism: Language and Cognition 23 (5), pp. 1020–31.

Shenk, P.S. 2006. 'The interactional and syntactic importance of prosody in Spanish-English bilingual discourse', International Journal of Bilingualism 10 (2), pp. 179–205.

Singer, H.-R. 1984. Grammatik der arabischen Mundart der Medina von Tunis ('Grammar of the Arabic dialect of the Medina of Tunis'). Berlin and New York: de Gruyter.

Smith, N.J. and R. Levy. 2013. 'The effect of word predictability on reading time is logarithmic', Cognition 128 (3), pp. 302–19.

Sridhar, S.N. and K.K. Sridhar. 1980. 'The syntax and psycholinguistics of bilingual code mixing', Canadian Journal of Psychology/Revue Canadienne de Psychologie 34 (4), pp. 407–16.

Stolcke, A. 2002. 'SRILM – an extensible language modelling toolkit' in Proceedings of the 7th International Conference on Spoken Language

Processing (ICSLP 2002), pp. 901–4. Denver, Colorado, USA. 16–20 September 2020. ISCA.

Tagliamonte, S.A. and R.H. Baayen. 2012. 'Models, forests, and trees of York English: *was/were* variation as a case study for statistical practice', Language Variation and Change 24 (2), pp. 135–78.

Torres Cacoullos, R. 2020. 'Code-switching strategies: prosody and syntax', Frontiers in Psychology 11: 2130.

Treffers−Daller, J. 1992. 'French–Dutch codeswitching in Brussels: social factors explaining its disappearance', Journal of Multilingual and Multicultural Development 13 (1–2), pp. 143–56.

Van Hell, J.G. and A.M.B. De Groot. 1998. 'Conceptual representation in bilingual memory: effects of concreteness and cognate status in word association', Bilingualism: Language and Cognition 1 (3), pp. 193–211.

Verreyt, N., E. Woumans, D. Vandelanotte, A. Szmalec and W. Duyck. 2016. 'The influence of language-switching experience on the bilingual executive control advantage', Bilingualism: Language and Cognition 19 (1), pp. 181–90.

Walters, K. 2011. 'Gendering French in Tunisia: language ideologies and nationalism', International Journal of the Sociology of Language 2011 (211).

Weiner, E.J. and W. Labov. 1983. 'Constraints on the agentless passive', Journal of Linguistics 19 (1), pp. 29–58.

Your short guide to the EUP Journals Blog http://euppublishingblog.com/

A forum for discussions relating to Edinburgh University Press Journals

1. The primary goal of the EUP Journals Blog

To aid discovery of authors, articles, research, multimedia and reviews published in Journals, and as a consequence contribute to increasing traffic, usage and citations of journal content.

2. Audience

Blog posts are written for an educated, popular and academic audience within EUP Journals' publishing fields.

3. Content criteria - your ideas for posts

We prioritize posts that will feature highly in search rankings, that are shareable and that will drive readers to your article on the EUP site.

4. Word count, style, and formatting

- Flexible length, however typical posts range 70-600 words.
- Related images and media files are encouraged.
- No heavy restrictions to the style or format of the post, but it should best reflect the content and topic discussed.

5. Linking policy

- Links to external blogs and websites that are related to the author, subject matter and to EUP publishing fields are encouraged, e.g.to related blog posts

6. Submit your post

Submit to ruth.allison@eup.ed.ac.uk

If you'd like to be a regular contributor, then we can set you up as an author so you can create, edit, publish, and delete your *own* posts, as well as upload files and images.

7. Republishing/repurposing

Posts may be re-used and re-purposed on other websites and blogs, but a minimum 2 week waiting period is suggested, and an acknowledgement and link to the original post on the EUP blog is requested.

8. Items to accompany post

- A short biography (ideally 25 words or less, but up to 40 words)
- A photo/headshot image of the author(s) if possible.
- Any relevant, thematic images or accompanying media (podcasts, video, graphics and photographs), provided copyright and permission to republish has been obtained.
- Files should be high resolution and a maximum of 1GB
- Permitted file types: *jpg, jpeg, png, gif, pdf, doc, ppt, odt, pptx, docx, pps, ppsx, xls, xlsx, key, mp3, m4a, wav, ogg, zip, ogv, mp4, m4v, mov, wmv, avi, mpg, 3gp, 3g2*.