Shared Lexical Items as Triggers of Code Switching

Shuly Wintner[†] and **Safaa Shehadi**[†] and **Yuli Zeira**[†] and **Doreen Osmelak**[◇] and **Yuval Nov**[‡]
[†]Department of Computer Science, University of Haifa, Israel
[◇]Department of Language Science and Technology, Universität des Saarlandes, Germany
[‡]School of Public Health, University of Haifa, Israel
shuly@cs.haifa.ac.il, safa.shehadi@gmail.com, yuli.zeira@gmail.com
s9doosme@stud.uni-saarland.de, yuval@stat.haifa.ac.il

Abstract

Why do bilingual speakers code-switch (mix their two languages)? Among the several theories that attempt to explain this natural and ubiquitous phenomenon, the *triggering hypothesis* relates code-switching to the presence of lexical triggers, specifically cognates and proper names, adjacent to the switch point. We provide a fuller, more nuanced and refined exploration of the triggering hypothesis, based on five large datasets in three language pairs, reflecting both spoken and written bilingual interactions. Our results show that words that are assumed to reside in a mental lexicon shared by both languages indeed trigger code-switching, that the tendency to switch depends on the distance of the trigger from the switch point and on whether the trigger precedes or succeeds the switch, but not on the etymology of the trigger words. We thus provide strong, robust, evidence-based confirmation to several hypotheses on the relationships between lexical triggers and code-switching.

1 Introduction

More than half the world's population today is multilingual, yet our understanding of the underlying linguistic and cognitive principles that govern multilingual language is imperfect. It is largely based on controlled laboratory studies, and only recently have psycholinguists begun exploring the extent to which insights from laboratory experiments can be applied in a real-world, communicative setting (Valdés Kroff and Dussias, 2023). Lacking firm theoretical underpinnings, contemporary language technology often does not reflect the ubiquity of multilingual communication.

We focus in this paper on *code-switching (CS)*, the natural tendency of bilingual speakers conversing with each other to switch between two languages, sometimes within a single utterance or even a single word. Our main goal is to explore a specific hypothesis related to CS, namely, that certain words tend to *trigger* CS more than others. The main contribution of this work is theoretical, but we trust that its results will be instrumental for improving future multilingual NLP applications.

Several competing theories try to explain CS, and in particular to identify the factors that contribute to the (typically unconscious) decision of a speaker to code-switch. Speakers are conjectured to code-switch when the concept they are about to utter is more *accessible* in the other language (Heredia and Altarriba, 2001); or more *specific*, lacking precise enough words in the current language (Backus, 2001); or carrying a major *information* load, so that the switch signals to the listener that an important concept is introduced (Myslín and Levy, 2015). The tendency to code-switch is influenced by linguistic factors (e.g., *cognates* are assumed to trigger CS), socio-linguistic factors (e.g., the fluency of the interlocutors in each of the two languages), demographic ones (e.g., the age, gender, or provenance of dialogue participants), and more (Myers-Scotton, 1993, 1998; Auer, 1998; Nilep, 2006).

We focus on the *triggering hypothesis*, whereby ''lexical items that can be identified as being part of more than one language for the speaker [...] may facilitate a transversion from one language to another'' (Clyne, 2003, p. 162). This hypothesis was explored extensively in the past, but earlier studies were limited in scope, were based on limited data, and addressed only spoken language.

This work makes several contributions. First, we investigate a specific type of lexical trigger: We define a category of lexical items (mainly proper names and culturally specific terms) that

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1471–1484, 2023. https://doi.org/10.1162/tacl_a_00613
Action Editor: Micha Elsner. Submission batch: 4/2023; Revision batch: 7/2023; Published 12/2023.

we expect to reside in more than one (or alternatively, in a *shared*) mental lexicon (Section 3). We also pay attention to whether such items originate in one of the two languages or in a third language. Second, unlike previous work, which dealt exclusively with spoken data, we investigate both spoken and written data, in five large datasets that include CS in three language pairs: English–Spanish, English–German, and English–Arabic[1] (Section 4). Third, while we employ the same statistical test that has been used by previous works to assess the association between such shared items and CS, we augment the analysis by also quantifying the *magnitude* of this association as an indication of the strength of the phenomena we observe (Section 5), thereby adding statistical rigor to our analysis.

Our results (Section 6) show strong associations between the presence of shared items (the type of trigger we focus on) and the tendency to code-switch, in all language pairs and datasets. We also provide a thorough and nuanced analysis (Section 7) of the location of the shared item with respect to the switch point, showing that the tendency to switch is lower when the trigger is adjacent to the switch rather than precedes it, and that the association between triggers and CS diminishes as the shared items are more distant from the switch point. Overall, we provide a much fuller, more nuanced picture of the relationships between lexical triggers and CS than was available so far.[2]

2 Related Work

Multilinguality is becoming more and more ubiquitous, to the extent that psycholinguists increasingly acknowledge that bilingualism is the rule and not the exception (Harris and McGhee Nelson, 1992). Grosjean (2010, page 16) stated that ''bilingualism is a worldwide phenomenon, found on all continents and in the majority of the countries of the world'' and Grosjean and Li (2013) assessed that more than half the world's population today is multilingual.

Monolingual and multilingual speakers alike seamlessly adjust their communication style to their interlocutors (Bell, 1984; Pickering and Garrod, 2004; Kootstra et al., 2012; Gallois and Giles, 2015; Fricke and Kootstra, 2016). Specifically, when interlocutors share more than one language, they almost inevitably engage in CS (Sankoff and Poplack, 1981; Muysken, 2000; Clyne, 2003).

Most linguistic research on CS has focused on *spoken* language (Lyu et al., 2010; Li and Fung, 2014; Deuchar et al., 2014, *inter alia*). However, with the rise of social media, *written* CS (Sebba et al., 2012) has become a pervasive communication style (Rijhwani et al., 2017). The spoken language domain is not directly comparable to the written one, and findings on CS in written conversations differ somewhat from those in speech (McClure, 2001; Chan, 2009; Gardner-Chloros and Weston, 2015). The work we present here addresses both modalities.

Various competing theories attempt to explain CS, or at least to propose factors that contribute to the tendency of bilingual speakers to code-switch. Notable among them is the *triggering hypothesis*, which states that specific lexical items that may be included in more than one mental lexicon for the speaker *trigger* switching (Clyne, 2003). Such lexical items include, according to Clyne, *lexical transfers* (i.e., borrowed words and expressions), *bilingual homophones* (including loans from a third language), and *proper nouns*. In this work we focus on a specific type of potential triggers, consisting mainly of proper names but including also culturally specific lexical items that originate in one language and do not have a readily available translation in the other language (e.g., '*taco*', originally from Spanish, in English–Spanish dialogues, or '*muezzin*', originally from Arabic, in English–Arabic conversations).

The triggering hypothesis was explored extensively by Clyne (1967, 1972, 1980, 1987), but these early investigations did not include any statistical analysis. This was first introduced by Broersma and De Bot (2006), who worked with ''a series of transcribed conversations between three Dutch-Moroccan Arabic bilinguals''. This dataset was extremely small by modern standards (it included a few dozen switch points and a few dozen potential triggers). Similarly, Broersma (2009) based her entire analysis on a single 24-minute interview with a single (Dutch–English

[1]The Arabic in this work reflects mostly the dialects of Egypt and Lebanon, and is written in *Arabizi*, an informal writing system that uses the Roman alphabet. See Section 4.

[2]All resources produced in this work, including the annotated datasets and the code, are publicly available on our GitHub repository.

speaking) informant. Still, both were able to find statistically significant associations between triggering and CS. More recently, Soto et al. (2018) extended this investigation to a larger corpus (the Bangor-Miami corpus of Spanish–English [Deuchar, 2009]), but focused only on a predefined list of cognates that they collected. In contrast, we work with much larger datasets that include thousands of switch points and potential triggers, in three different language pairs, and with both spoken dialogues and written social-media interactions.

Broersma and De Bot (2006) (and, subsequently, also Broersma [2009] and Soto et al. [2018]) used the χ^2 test to measure the correspondence between triggering and CS. We use the same measure (more precisely, *Fisher's exact test*, whose significance does not rely on an approximation that is only exact in the limit); but we extend the analysis by considering not only the statistical significance of the test, as determined by its p-value, but also the magnitude of the association between categories as an indication of the strength of the phenomena we observe, as determined by *relative risk* (also known as *risk ratio*). This facilitates a much more nuanced analysis of the results.

3 Goals

Our main goal in this work is to explore the triggering hypothesis more closely, focusing on a class of lexical items that we expect to be shared across the multiple mental lexicons of the multilingual speaker. Extending previous research, we aim at addressing the association between such shared items and CS in multiple datasets reflecting three different language pairs (EN–AR, EN–DE, and EN–ES)[3] and two different modalities (spoken and written).

3.1 *Shared* Lexical Items

Shehadi and Wintner (2022) defined *shared* lexical items as named entities in one language that are not translated to the other, and consequently have a similar form in both languages. They also included terms that lack (or have rare) translation equivalents in the other language.

Following Osmelak and Wintner (2023), we refine the definition of *shared* items by reflecting also the language in which such terms originate. Our motivation is the assumption that a word like '*taco*', which originates in Spanish but is fully adopted by English, may trigger code-switching from English to Spanish but perhaps less so in the reverse direction.

In addition, words in L_1 that do not have a commonly used translation equivalents in L_2, and are hence used in both languages (e.g., '*taxi*', which is commonly used in many Arab-speaking communities) are not considered a code-switch themselves but may trigger code-switching.[4] Specifically, we divide the *shared* category to three subcategories, depending on the origin of the word.

Shared English Named entities shared between two lexicons that originate in English, including person names (e.g., '*Johnson*'), commercial entities (e.g., '*Twitter*', '*Seven Eleven*'), and geographic names that contain English words (e.g., '*Times Square*'). Also included are English-originating cultural terms that are adopted by the other language (e.g., '*taxi*', '*film*') and English acronyms used cross-culturally on social media (e.g., '*lol*').

Shared Arabic/German/Spanish Named entities shared between the two lexicons[5] whose origin is Arabic (e.g., '*Salah*', '*Bahrain*'), German (e.g., '*Merkel*', '*Berlin*'), or Spanish (e.g., '*Carlita*', '*Guatemala*'). Also, culturally dependent terms originating in these three languages that do not have translations in English, e.g., Arabic '*Ramadan*', German '*schnitzel*', or Spanish '*taco*'. This category also includes interjections that are identified with one of these languages, e.g., Spanish '*jajajaja*'; and acronyms that expand to those languages (e.g., Spanish '*PR*' for '*Puerto Rico*' or German '*NRW*' for '*Nordrhein-Westfalen*').

Shared Other Words and terms that are used in both languages, but are not clearly identified with either of them, including named entities or terms that originate in a third language (e.g.,

[3]We use *AR* for Arabic, *DE* for German, *EN* for English, and *ES* for Spanish.

[4]Another deviation from the scheme of Shehadi and Wintner (2022) is that we treat named entities that are specific to a foreign language as words in that language. For example, '*Lebanon*', which is an English-specific variant of the Arabic '*lubnan*', is viewed as an English token.

[5]These categories are defined separately for each language pair. For example, *shared-Arabic* is defined only for the EN–AR datasets. The same holds for *shared-Other*.

‘*Erdogan*’, ‘*Pikachu*’, or ‘*pizza*’); terms whose origin is English but that do not include strong English linguistic features (e.g., ‘*iPod*’); interjections that are commonly used in both languages (e.g., ‘*oh*’ or ‘*wow*’); person names that are common in both languages (e.g., ‘*Lily*’, ‘*Adam*’); and geographical terms that originate in a third language and are written and pronounced similarly in both languages (e.g., ‘*Vietnam*’).

It is important to note that the tagging is context-dependent: Much like named entities, shared items may have different readings (i.e., tags) depending on the context in which they occur. Consider the two examples below. The token ‘*warda*’ is tagged as Arabic in Example 1, but as shared-Arabic in Example 2. Consequently, using lists of shared items (as was done in previous work, e.g., by Soto et al. [2018]), is not a sufficient solution.

(1) *Maynf3sh warda wahda tayep !*
it doesn’t work flower one only !
‘It doesn’t work, only one flower!’

(2) *kan beydafe3 3an amr warda*
was defend about Amr Warda
‘He was defending Amr Warda’

Finally, note that some shared items are multi-word, e.g., ‘*amr warda*’ (a person name) in Example 2. When all tokens have the same origin L, we label the item shared-L; we do the same also when some tokens are shared-L and others are shared-Other. But if one token is shared-L_1 and the other is shared-L_2, we label each token differently. For example, we tag ‘*Nueva York*’ as shared-Spanish followed by shared-English.

3.2 Hypotheses

We pose the following hypotheses:

1. *Shared* lexical items are associated with CS, i.e., they tend to co-occur in the same utterances. This is the main hypothesis investigated intensively by Clyne’s many works and by subsequent research, but we define shared items somewhat differently here, not relying on predefined (or manually annotated) lists of cognates and proper names.

2. Such tendencies are more pronounced when the trigger is closer to the switch point. Previous work investigated ‘‘adjacent words’’, whereas we investigate shared words located up to 6 tokens from the switch point.

3. Triggers that precede the switch point are more strongly associated with CS than those that are adjacent to them. Broersma and De Bot (2006) explain that the trigger can succeed the CS point because language planning does not always work linearly, and the choice of language for words is not necessarily aligned with the linear order of these words in a sentence. They therefore search for ‘‘basic clauses’’ that contain both switches and trigger words, in any order. We do not define basic clauses, resorting instead to a fixed-length window around shared items. But we do check, separately, the case of shared items that precede the CS point, and those that occur on either side of the CS point. We do not separately investigate potential triggers that *follow* the CS point because we expect the association in such cases to be weak. We focus instead on triggers *near* the switch, on either side of it, and compare this situation with triggers that strictly precede the switch.

4. Terms that originate in language L_1 are more likely to trigger a switch from L_2 to L_1 than the other way round. Our rationale here stems from the assumption that shared-L_1 words may be more deeply rooted in the lexicon of L_1 than the lexicon of L_2, even if they are included in both; and hence are more likely to trigger switches *to* L_1 than *from* L_1.

These hypotheses are based on a precise definition of what constitutes a CS point (detailed in Section 5.1). But first, we describe the datasets we use to investigate these hypotheses.

4 Data

We use five different datasets, in three language pairs. The texts are either transcribed dialogues (in the case of Bangor-Miami) or sequences of utterances that constitute a *thread* (in the case of social media). We view a turn of a single author/speaker as a basic unit; if the dataset is not already tokenized, we segment turns to utterances and then to tokens using NLTK (Bird et al.,

2009). Each token is then associated with a language ID tag.

Arabic–English We used the English–Arabizi (Arabic written in the Roman alphabet) dataset compiled and released by Shehadi and Wintner (2022). This corpus includes social media posts from Reddit and Twitter; it contains 2,643 utterances that were manually annotated for language ID (at the word level), which were used to train a highly accurate classifier (the accuracy of identifying words in Arabizi and English was 95%; identifying *shared* items was only 84% accurate, with a precision of 89% and much lower recall). The classifier was then used to automatically annotate additional utterances, resulting in a total of over 865,000 utterances that include CS between English and Arabizi. Each word in this dataset is associated with a unique language ID: Arabizi, English, French,[6] Arabic, Shared, or Other.

We re-annotated the manually annotated data according to our revised definition of shared items and then retrained the classifier and applied it to the entire dataset. We then combined the manually and automatically annotated subsets of each dataset; this resulted in two coherent datasets, one with Reddit posts and the other with Twitter comments. We report results for each dataset separately because they reflect different genres. Table 1 lists statistics for the two EN–AR datasets.[7]

Spanish–English We used two Spanish–English datasets: *Bangor-Miami (BM)* (Deuchar, 2009), a corpus of transcribed Spanish–English bilingual speech; and *SentiMix* (Aguilar et al., 2020), a dataset that was created for investigating sentiment analysis in a code-switched environment (Patwa et al., 2020).

Both corpora include token-level manually annotated language ID tags, but they use different schemes and include ambiguous language tags for named entities and cross-lingual terms. We manually changed the language tags of such tokens to English, Spanish, or a sub-class of *Shared*, according to the scheme of Section 3.1, so that they are consistent throughout the corpus.[8] Table 2 lists statistics for the two EN–ES datasets.[9]

	Reddit	%	Twitter	%
Utterances	205,397		659,958	
Tokens (total)	3,855,900		5,340,658	
Arabizi	585,830	15.2	2,978,070	55.8
English	2,678,442	69.5	1,639,966	30.7
French	7,363	0.2	6,872	0.1
Shared-EN	7,779	0.2	17,849	0.3
Shared-AR	31,992	0.8	19,190	0.4
Shared-Other	20,333	0.5	11,679	0.2
CS (total)	274,200		471,334	
EN→AR	133,642	48.7	233,616	49.6
AR→EN	140,558	51.3	237,718	50.4

Table 1: Statistics of the EN–AR datasets.

	BM	%	SentiMix	%
Utterances	42,854		12,193	
Tokens (total)	277,963		186,585	
English	171,791	61.8	41,290	22.1
Spanish	91,419	32.9	91,419	49.0
Shared-EN	5,659	2.0	1,125	0.6
Shared-ES	2,042	0.7	1,752	0.9
Shared-Other	7,035	2.5	1,585	0.8
MIX	17	0.0	17	0.0
CS (total)	3,923		19,226	
EN→ES	1,669	42.5	8,864	46.1
ES→EN	2,254	57.5	10,362	53.9

Table 2: Statistics of the EN–ES datasets.

German–English We used the Denglisch corpus of mixed English–German Reddit posts compiled and released by Osmelak and Wintner (2023), with its original (‘‘collapsed’’) annotations, which are consistent with our scheme. As in the case of Arabizi, a small subset of this corpus (4,200 sentences) was annotated manually, and the remainder (over 228,000 sentences) was tagged by a classifier trained on the manually annotated subset. The overall word-level accuracy of the classifier was 96.5%, with excellent

[6]We focused only on AR–EN here because the number of French words in the corpus, and consequently the number of French–Arabic CS, was limited. See Table 1.

[7]The percentages do not always sum up to 1 because the datasets may include tokens with other tags (punctuation, emoji, hashtags, etc.) Additionally, the numbers of shared tokens are actually counts of shared *items*: If an item is multi-word, it is counted only once.

[8]No classifier was used on these datasets.

[9]MIX is a category of words that combine morphemes from the two languages; due to the relatively low number of such items, we ignore them in this work.

	German	%
Utterances	36,524	
Tokens (total)	5,429,970	
English	1,826,171	33.6
German	2,281,859	42.0
Shared-EN	26,363	0.5
Shared-DE	24,874	0.5
Shared-Other	52,529	1.0
MIX	4,187	0.1
CS (total)	270,375	
EN→DE	134,478	49.7
DE→EN	135,897	50.3

Table 3: Statistics of the EN–DE dataset.

(97–98%) accuracy for English and German tokens, and 60–66% for shared items (again, with much higher precision than recall). Table 3 lists statistics for this dataset.

Table 4 depicts a few examples of utterances from our datasets, along with their annotation according to the scheme outlined above. Example 3 starts in English but then '*Ahly*' (an Egyptian football club) is mentioned; this token is tagged *shared-Arabic*, and indeed after a few more English tokens, the author switches to Arabic. In Example 4 the reverse pattern is observed: The utterance begins in Spanish, but two proper names tagged as *shared-English* are introduced, and evidently the author switches to English. Finally, Example 5 begins in German and ends in English, perhaps in connection with the use of '*schnitzel*', which is *shared-German*.

5 Methodology

5.1 Definition of CS Points

To check the association between shared items and CS, the latter concept must be carefully defined, which is not always a trivial task (Alvarez-Mellado and Lignos, 2022); previous work has sometimes been careless with this. We consider CS to be a property of a single token, defined as follows: A token w is considered code-switched from L_1 to L_2 when: (i) w is labeled as L_2; (ii) it is preceded (in the same utterance) by a sequence of $n \geq 0$ tokens labeled neither as L_1 nor as L_2; and (iii) this sequence is preceded (in the utterance) by a token labeled as L_1. This definition allows for sequences of shared lexical items (and other tokens, e.g., emoji) to intervene between a token in L_1 and a token in L_2; the CS point is the first L_2 token that follows such a sequence.

Having said that, we exclude some CS points from our analysis: We treat *insertional* switches differently from *alternational* ones. Muysken (1997) defines alternation as ''a true switch from one language to the other, involving both grammar and lexicon''. All three examples in Table 4 are alternational. In contrast, insertion is the embedding of a phrase from one language into an utterance that is otherwise in the other language. Example 6 demonstrates insertional CS: The English token '*technically*' is inserted into an otherwise fully Arabic utterance.

(6) *Gama3a e7na technically fi ramadan*
guys we're in Ramadan
'Guys, we're technically in Ramadan'

It is common to assume that insertional CS like the one in Example 6 involve *a single* trigger, which affects the tendency to switch from L_1 to L_2; the switch back to L_1 is merely an inevitable consequence of the CS being insertional. Therefore, we exclude from our analyses the second switch in case of insertional CS.[10]

We operationalize this as follows: Given a sequence of tokens $w_1w_2w_3$, where w_1 and w_3 are in L_1 and w_2 is in L_2, we only consider w_2, but not w_3, to be a CS point. This does introduce noise occasionally, especially because some insertional switches involve the insertion of two, and sometimes even three tokens, as in Example 7. In such cases, the switch back to Arabic will (erroneously) be taken into consideration in our analyses.

(7) *Mafi good internet b kel lebnen*
there-isn't in all Lebanon
'There's no good internet in all of Lebanon'

5.2 Statistical Analysis

To explore the associations between shared items (as defined in Section 3.1) and CS (as defined above), we ran a multitude of statistical tests.

[10] We experimented also with the alternative approach, namely, treating insertional and alternational switches identically. The results were pretty similar.

(3) *every time I watch an ahly game I get goosebumps fel de2i2a el 74*
Ahly in minute the
EN EN EN EN EN SH-AR EN EN EN EN AR AR AR Other

'Every time I watch an Ahly game I get goosebumps in the 74th minute'

(4) *tu eres scott y yo soy kourtney , had n't we agreed on this ?*
you are Scott and I am Kourtney
ES ES SH-EN ES ES ES SH-EN Other EN EN EN EN EN EN Other

'You are Scott and I am Kourtney, hadn't we agreed on this?'

(5) *Aba sie sagt ja making schnitzel for my husband*
But she says yes
DE DE DE DE EN SH-DE EN EN EN

'But she says yes, making a schnitzel for my husband'

Table 4: Example utterances with their language ID annotations.

		Is *shared*	
Near CS		Yes	No
	Yes	216	17515
	No	659	143299
		24.7%	10.9%

Relative switching propensity: 2.266
p-value: 2.2×10^{-30}

Table 5: Contingency table constructed for the SentiMix corpus, reflecting EN→ES switches that follow shared-English lexical items at distance at most 2 from the switch point.

The tests vary in terms of the dataset used, the type of shared items investigated (the three subclasses of shared items, or all shared items combined), the direction of the CS (from English to the other language or vice versa), whether the shared item *precedes* the CS point or *neighbors* it (given a shared item, we look for CS points following it, but also adjacent to it on either side), and the distance between the two (we look at CS distanced at most 1 to 6 tokens from the shared item).

We now outline the structure of a single such test, where the dataset is the SentiMix corpus, the type of shared item is shared-English, the CS direction is EN→ES, and the CS follows the item, at a distance of at most 2 tokens. Table 5 depicts the data used in this test: It is a 2×2 contingency table, whose columns indicate if the lexical item is shared or not, and whose rows correspond to the presence or absence of CS points near the shared item. The sum of the numbers in the first column is the number of shared items in the dataset, and in the first row, the number of switch points investigated (a single CS point may be counted several times for different shared items). We exclude from the investigation the first and the last token in each utterance (the last token in an utterance cannot trigger a switch following it; and the first is limited in triggering a switch neighboring it).

Across the shared items in the dataset, the proportion of switches (within the specified distance) is $216/(216 + 659) = 24.7\%$, whereas across the non-shared items, this proportion is $17515/(17515 + 143299) = 10.9\%$. Thus, the propensity to switch is $24.7/10.9 = 2.266$ times higher near a shared item, compared to a non-shared item. We refer to the latter ratio as the *relative switching propensity*; mathematically, it is analogous to the well known ''relative risk'' (or ''risk ratio'') from epidemiology and biostatistics (Rothman, 2012). We test whether this ratio differs from 1 in a statistically significant way via a Fisher test, and obtain a p-value of 2.2×10^{-30}, indicating a highly significant increased tendency to switch near a shared item.

Clearly, the relative switching propensity equals 1 if and only if the odds ratio equals 1, and the same statistical test (namely, Fisher's) is appropriate for studying either of these quantities. We prefer to use the relative switching propensity to quantify the magnitude of the association

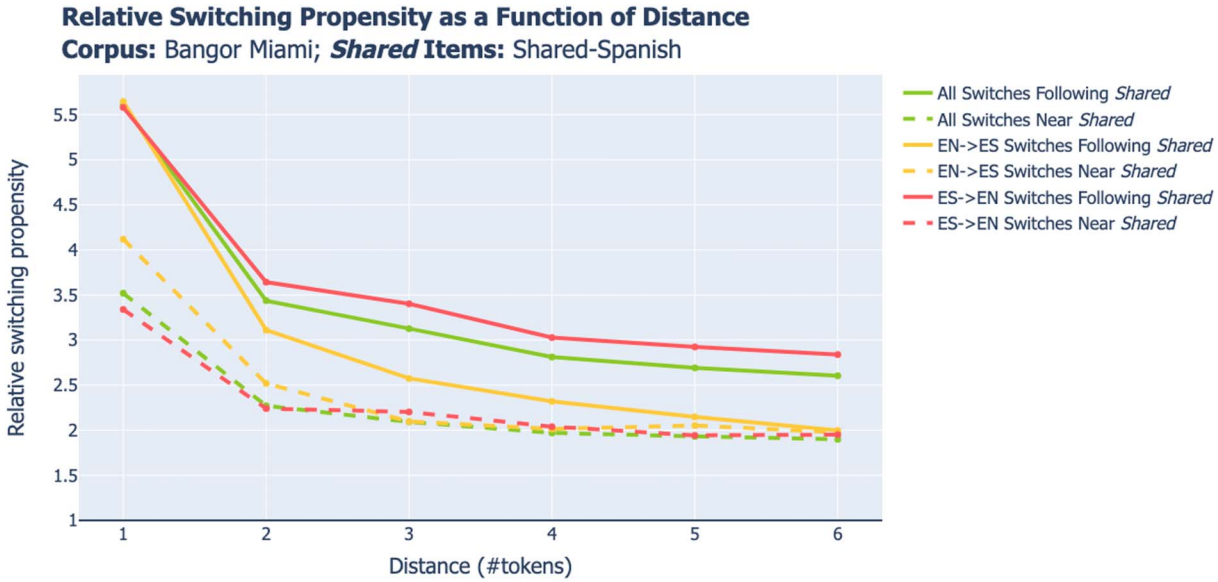

Figure 1: A multi-test plot depicting the results of 36 tests on the Bangor-Miami corpus with shared Spanish items.

between shared items and CS, as it is more readily interpretable: in the above example, the number 2.266 is our estimate for the factor by which switches are more common near shared items, compared to near non-shared items.

6 Results

Section 5.2 defines multiple statistical tests: First, we work with five datasets; for each dataset, we individually explore four types of shared items. For example, with Bangor-Miami we independently explore shared-English items, shared-Spanish, shared-Other, and all shared items combined. We depict the results of each such ‘‘multi-test’’, with a specific dataset and a specific type of shared item, as one plot.

On each plot we depict the results of 36 statistical tests, which differ by the type of switch (EN→ES or ES→EN or both); whether the shared item precedes the CS point or neighbors it; and finally, the distance between the two (1–6). The result of each of these 36 statistical tests is depicted as a point in a graph, where the x axis is the distance (1–6) and the y axis indicates the value of the relative switching propensity for the specific statistical test. This facilitates a clear view of the magnitude of the effect (i.e., the strength of the association) of each test. Additionally, when the p-value of a particular test is greater than 0.05 (the usual threshold for statistical significance), we indicate this as a black diamond marking on the point that corresponds to that test.

Figure 1 shows the multi-test plot reflecting the results on the Bangor-Miami corpus with shared-Spanish items. The 36 points on this plot are connected by lines: Solid lines reflect statistical tests where the shared item precedes the CS point, and dashed lines are for statistical tests where the shared item can occur before or after a CS point. The color of the line reflects the type of switch: EN→ES (yellow), ES→EN (red), or both (green). See the legend to the right of the plot.

Several observations are revealed in Figure 1. First and foremost, with no exceptions, all the tests yield statistically significant results ($p < 0.05$), as there are no black diamonds on the plot. This fundamentally supports our first hypothesis, namely, that there is a clear association between shared items and CS. Furthermore, all the lines are monotonically decreasing, or at least non-increasing, thereby confirming our second hypothesis: The association between shared items and CS is stronger when the two are close, and diminishes as the distance between them increases.

The fact that the solid lines are always above the dashed lines confirms our third hypothesis: The association is stronger with shared items that precede CS points than with shared items that are adjacent to them, on either side. Finally, the solid red line is always above the solid yellow line, indicating that shared-Spanish items are more strongly associated with CS from Spanish to English (red) than with CS from English to Spanish (yellow),

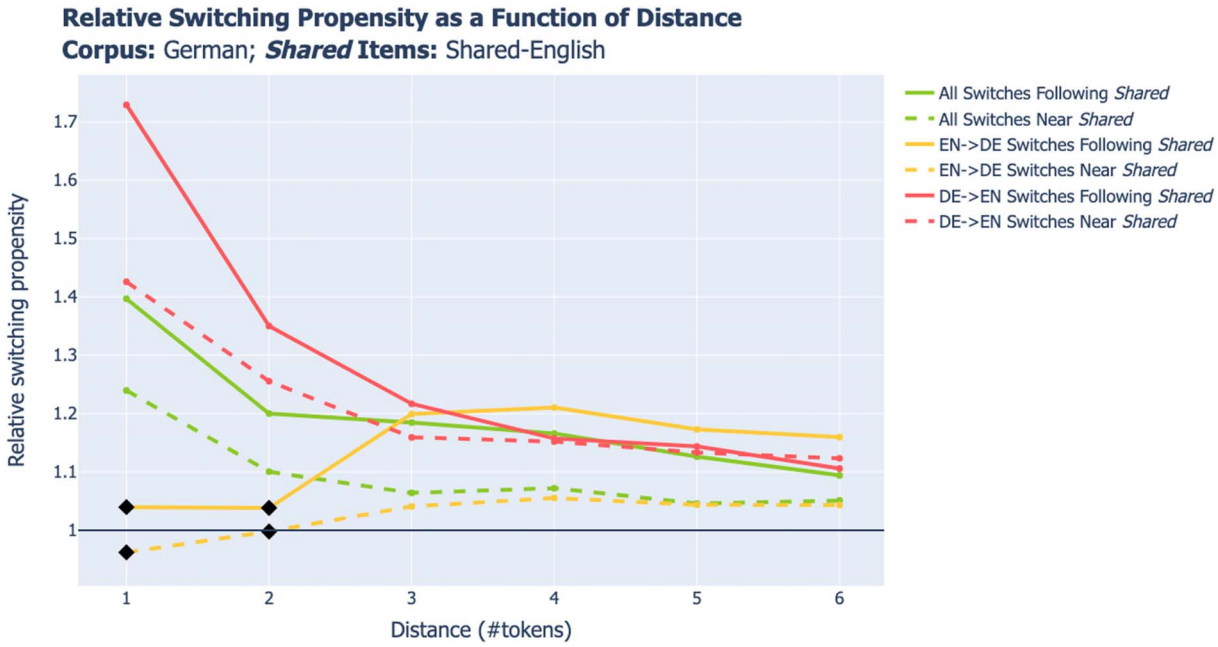

Figure 2: A multi-test plot depicting the results of 36 tests on the EN-DE corpus with shared English items.

in contrast to our hypothesis. This pattern, however, is partly reversed in the dashed lines: The jury is still out on our fourth hypothesis.

While Figure 1 summarizes the results of 36 statistical tests, it is only one out of 20 similar ‘‘multi-tests’’: we have similar plots for five datasets, with four types of shared items per dataset. Space limitations prevent us from presenting all of them here (they will be included in the supplementary materials), but we do show a similar plot of the Denglisch EN-DE corpus, with data reflecting shared-English items, in Figure 2. In addition, we now analyse the aggregate results of all 20 multi-tests in light of our four hypotheses.

7 Analysis

Association between Shared Items and CS. Our first hypothesis was that shared items are indeed associated with CS. To assess the association, we expect the Fisher test to yield a statistically significant result in each of the statistical tests (i.e., no black diamonds in the plots). Not surprisingly, we do find such association. Recall that each plot (such as the one in Figure 1) depicts the results of 36 tests, and that we have 20 such plots. Of the 720 statistical tests, only 10 (1.4%) yield p-values greater than 0.05. We thus overwhelmingly establish the hypothesis that in all our datasets there is significant association between shared items of all kinds and CS, even when the shared item is as far as 6 tokens away from the CS point, and even when the shared item is adjacent to (i.e., may succeed) the CS point. It is interesting to note that of the 10 exceptions, 8 are in the Denglisch corpus. We return to this below.

The Impact of the Distance between Shared Items and the CS Point. Our second hypothesis was that the magnitude of the association diminishes as the shared item and the CS point are more distant. To confirm this hypothesis we need to show that the lines in the plots are decreasing, or at least non-increasing. This is indeed the case for 98 (82%) out of the 120 lines (6 per plot). Furthermore, most of the lines that are not decreasing include only a single point that violates the hypothesized trend. The main issue is, again, with the DE-EN dataset, which is responsible for 13 out of the 22 exceptions.

Shared Items Before or Adjacent to CS Points. We also hypothesized that the shared item ‘‘triggers’’ CS, namely, that such items are more influential when they precede the CS point, as compared to when they are merely adjacent to it (on either side). To establish this hypothesis we need to show that the solid lines are above the dashed ones.

Of the 720 data points in our 20 plots, only 38 did not comply with this condition; in other words, our hypothesis holds for almost 95% of the points. One potential reason for the existence of outliers may be noise in our definition of insertional switches. Recall from Section 5.1 that we try to find triggers for all switches except switches ''back from'' an insertion, but our definition of insertional CS assumes that they consist of exactly one token, whereas in reality some of them are multi-word expressions. We expect that switching ''out of'' such longer insertional switches does not require a trigger, but our analysis nonetheless looks for one. Interestingly, all but one of the outliers involve switches *from* English to the other language (yellow lines). We do not have an explanation for this observation.

The Etymology of the Shared Item and its Relation to the Direction of the Switch. Finally, we hypothesized that shared-L_1 items are more strongly associated with $L_2 \rightarrow L_1$ switches than with $L_1 \rightarrow L_2$ switches. This hypothesis is not supported by the data. For example, in the two AR-EN datasets, switches *to* English were systematically more prominent than switches *from* English, independently of the type of the shared item. In the EN-ES datasets, shared-EN and shared-ES items were associated with switches of both types almost to the same extent; and the DE-EN dataset also showed mixed, inconsistent results.

One potential explanation for this observation has to do with insertional switches. With the exception of Bangor-Miami, our datasets reflect social media interactions on platforms that typically include discussions in English. We conjecture that when a particular discussion is conducted in another languages, insertions of English expressions are highly likely, much more than insertions of phrases in the other language to an otherwise English utterance. As mentioned above, our handling of insertional switches is noisy, which might affect the results.

A more theoretically based explanation of our failure to confirm the fourth hypothesis is grounded in theories of bilingualism which maintain that the two languages of a bilingual are both active simultaneously, and one of them has to be suppressed in order to yield words in the other (e.g., Finkbeiner et al., 2006). If this is indeed the case, then the origin of a shared item does not have to influence its likelihood to trigger a switch in any particular direction.

Summary Previous work on the triggering hypothesis focused solely on spoken dialogues and, consequently, was limited by the data available: typically, a few dozen dialogues spoken by a handful of participants in a single language pair. The extension to written CS, exemplified here, opens the door to investigations with vast amounts of data, but also raises interesting questions on the differences between written and spoken language and how CS is manifested in both modalities. Another interesting question has to do with the differences in the ways CS is manifested in closely-related language pairs (English–German, and to a lesser extent also English–Spanish) vs. in typologically unrelated language pairs (English–Arabic). A third dimension of comparison involves the differences in how CS is related to the status of the two languages involved: whether one of them is a minority language, a heritage language, or a lingua franca.

A thorough investigation of all these issues is beyond the scope of this work; but we do note that among the five datasets used in this research, we did not find major differences between the (spoken dialogue) Bangor-Miami corpus and the (Twitter) SentiMix corpus, which are both in English–Spanish. This suggests that CS in written language, at least as it is used on very informal social media outlets, behaves similarly to CS in spoken language with respect to the triggering hypothesis.

We *did* find significant differences between the Denglisch dataset and all others. Like the Denglisch corpus, one of the Arabizi datasets we studied also consists of Reddit posts, so the peculiarity of the English–German corpus cannot be attributed to the source of the texts it includes. As Osmelak and Wintner (2023) note, this dataset is different from most corpora of bilingual language: German is the official language of Germany, where English is widely understood but is not a minority or heritage language, nor a lingua franca of a sub-community. This may result in a unique pattern of CS, different from the one observed in other language pairs, and might explain the different results we obtain on this dataset. We conjecture that CS between English and German is special because of the status of the two languages in German-speaking countries,

but more research is needed to confirm this conjecture.

8 Conclusion

We investigated the triggering hypothesis using five datasets that reflect bilingual interactions in three language pairs. Employing standard yet powerful statistical methodology, we strongly confirmed three hypotheses: (1) that there is a strong association between code-switching and shared lexical items (proper names, but also culturally specific items that may lack translation equivalents in the other language); (2) that this association is stronger when the shared item precedes the switch point, rather than neighbors it; and (3) that the association diminishes as the shared item is farther away from the CS point.

We were unable to confirm a fourth hypothesis, namely, that shared items originating in language L_1 are more likely to trigger a switch from L_2 to L_1 than the other way round. We do not know whether this is due to noise in our datasets or a bona fide property of bilingual language, rooted in cognitive-theoretical explanations; we leave this for future investigation.

While the data used to establish the above results are unprecedented in terms of their size and diversity, at least in the psycholinguistic literature, we believe that they do not tell a full story. We would very much like to extend our datasets to more language pairs, to have sufficiently large datasets that would facilitate a comparative analysis of spoken vs. written data, and also enough data to compare CS between etymologically close languages vs. unrelated language pairs. We leave such investigations for future research.

Ethical Considerations

This research was approved by the University of Haifa institutional review board. We used previously collected data that are freely available for research purposes, and redistribute those data according to their original licenses. All data are anonymized and we anticipate very minimal risk of abuse or dual use of the data.

Limitations

Our datasets are by no means representative, and any conclusion resulting from their processing is limited to the population of speakers they reflect. However, the magnitude of the data we used here, especially compared to the sizes of corpora used previously to derive theories of code-switching, is sufficient to guarantee the replicability of our findings on further data.

Acknowledgments

We are grateful to Melinda Fricke and Anat Prior for many discussions related to this work, and excellent ideas. We also thank the three anonymous TACL reviewers for their valuable feedback and suggestions. This work was supported in part by grant no. 2019785 from the United States-Israel Binational Science Foundation (BSF), and by grants 2007960, 2007656, 2125201, and 2040926 from the United States National Science Foundation (NSF).

References

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Elena Alvarez-Mellado and Constantine Lignos. 2022. Borrowing or codeswitching? Annotating for finer-grained distinctions in language mixing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3195–3201, Marseille, France. European Language Resources Association.

Peter Auer. 1998. *Code-Switching in Conversation: Language, Interaction and Identity*. Routledge.

Ad Backus. 2001. The role of semantic specificity in insertional codeswitching: Evidence from Dutch–Turkish. In Rodolfo Jacobson, editor, *Codeswitching Worldwide II*, pages 125–154. De Gruyter Mouton, Berlin, New York. https://doi.org/10.1515/9783110808742.125

Allan Bell. 1984. Language style as audience design. *Language in Society*, 13(2):145–204. https://doi.org/10.1017/S004740450001037X

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Mirjam Broersma. 2009. Triggered codeswitching between cognate languages. *Bilingualism: Language and Cognition*, 12(4):447–462. https://doi.org/10.1017/S1366728909990204

Mirjam Broersma and Kees De Bot. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and Cognition*, 9(1):1–13. https://doi.org/10.1017/S1366728905002348

Brian Hok-shing Chan. 2009. English in Hong Kong Cantopop: Language choice, code-switching and genre. *World Englishes*, 28(1):107–129. https://doi.org/10.1111/j.1467-971X.2008.01572.x

Michael G. Clyne. 1967. *Transference and Triggering; Observations on the Language Assimilation of Postwar German-speaking Migrants in Australia*. Ph.D. thesis, Monash University, The Hague, Netherlands.

Michael G. Clyne. 1972. *Perspectives on Language Contact: Based on a Study of German in Australia*. Hawthorn Press.

Michael G. Clyne. 1980. Triggering and language processing. *Canadian Journal of Psychology*, 34:400–406. https://doi.org/10.1037/h0081102

Michael G. Clyne. 1987. Constraints on code switching: How universal are they? *Linguistics*, 25(4):739–764. https://doi.org/10.1515/ling.1987.25.4.739

Michael G. Clyne. 2003. *Dynamics of Language Contact: English and Immigrant Languages*. Cambridge Approaches to Language Contact. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511606526

Margaret Deuchar. 2009. The Miami Corpus: Documentation file. Unpublished manuscript.

Margaret Deuchar, Peredur Davies, Jon Herring, M. Carmen Parafita Couto, and Diana Carter. 2014. Building bilingual corpora. *Advances in the Study of Bilingualism*, pages 93–111. https://doi.org/10.21832/9781783091713-008

Matthew Finkbeiner, Tamar H. Gollan, and Alfonso Caramazza. 2006. Lexical access in bilingual speakers: What's the (hard) problem? *Bilingualism: Language and Cognition*, 9(2):153–166. https://doi.org/10.1017/S1366728906002501

Melinda Fricke and Gerrit Jan Kootstra. 2016. Primed codeswitching in spontaneous bilingual dialogue. *Journal of Memory and Language*, 91:181–201. https://doi.org/10.1016/j.jml.2016.04.003

Cindy Gallois and Howard Giles. 2015. Communication accommodation theory. In Karen Tracy, Cornelia Ilie, and Todd Sandel, editors, *The International Encyclopedia of Language and Social Interaction*, pages 1–18. Wiley Online Library. https://doi.org/10.1002/9781118611463.wbielsi066

Penelope Gardner-Chloros and Daniel Weston. 2015. Code-switching and multilingualism in literature. *Language and Literature*, 24(3):182–193. https://doi.org/10.1177/0963947015585065

François Grosjean. 2010. *Bilingual: Life and Reality*. Harvard University Press. https://doi.org/10.4159/9780674056459

François Grosjean and Ping Li. 2013. *The Psycholinguistics of Bilingualism*. Wiley-Blackwell.

Richard Jackson Harris and Elizabeth Marie McGhee Nelson. 1992. Bilingualism: Not the exception any more. In Richard Jackson Harris, editor, *Cognitive Processing in Bilinguals*, volume 83 of, *Advances in Psychology*, pages 3–14. North-Holland. https://doi.org/10.1016/S0166-4115(08)61485-5

Roberto R. Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science*, 10(5):164–168. https://doi.org/10.1111/1467-8721.00140

Gerrit Jan Kootstra, Janet G. Van Hell, and Ton Dijkstra. 2012. Priming of code-switches in sentences: The role of lexical repetition,

cognates, and language proficiency. *Bilingualism: Language and Cognition*, 15(4):797–819. https://doi.org/10.1017/S136672891100068X

Ying Li and Pascale Fung. 2014. Language modeling with functional head constraint for code switching speech recognition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 907–916. https://doi.org/10.3115/v1/D14-1098

Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, and Haizhou Li. 2010. SEAME: A Mandarin–English code-switching speech corpus in South-east Asia. In *INTERSPEECH 2010*, pages 1986–1989.

Erica McClure. 2001. Oral and written Assyrian–English codeswitching. In Rodolfo Jacobson, editor, *Codeswitching Worldwide II*, volume 126 of *Trends in Linguistics. Studies and Monographs [TiLSM]*, pages 157–191. Mouton de Gruyter Berlin. https://doi.org/10.1515/9783110808742.157

Pieter Muysken. 1997. Code-switching processes: Alternation, insertion, congruent lexicalization. In Martin Pütz, editor, *Language Choices: Conditions, Constraints, and Consequences*, pages 361–380. Benjamins, Amsterdam. https://doi.org/10.1075/impact.1.25muy

Pieter Muysken. 2000. *Bilingual Speech: A Typology of Code-mixing*. Cambridge: Cambridge University Press.

Carol Myers-Scotton. 1993. *Social Motivations for Codeswitching: Evidence from Africa*. Clarendon Press, Oxford.

Carol Myers-Scotton. 1998. *Codes and Consequences: Choosing Linguistic Varieties*. Oxford University Press, Oxford.

Mark Myslín and Roger Levy. 2015. Code-switching and predictability of meaning in discourse. *Language*, 91(4):871–905. https://doi.org/10.1353/lan.2015.0068

Chad Nilep. 2006. ‘‘Code switching’’ in sociocultural linguistics. *Colorado Research in Linguistics*, 19. https://doi.org/10.25810/hnq4-jv62

Doreen Osmelak and Shuly Wintner. 2023. The Denglisch corpus of German-English code-switching. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 42–51. https://doi.org/10.18653/v1/2023.sigtyp-1.5

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.100

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190. https://doi.org/10.1017/S0140525X04000056

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Sekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1971–1982. https://doi.org/10.18653/v1/P17-1180

Kenneth J. Rothman. 2012. *Epidemiology: An Introduction*. Oxford University Press.

David Sankoff and Shana Poplack. 1981. A formal grammar for code-switching. *Research on Language & Social Interaction*, 14(1):3–45. https://doi.org/10.1080/08351818109370523

Mark Sebba, Shahrzad Mahootian, and Carla Jonsson. 2012. *Language Mixing and Code-switching in Writing: Approaches to Mixed-language Written Discourse*. Routledge. https://doi.org/10.4324/9780203136133

Safaa Shehadi and Shuly Wintner. 2022. Identifying code-switching in Arabizi. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 194–204. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wanlp-1.18

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The role of cognate words, POS tags, and entrainment in code-

switching. *Proceedings of Interspeech 2018*, pages 1938–1942. https://doi.org/10.21437/Interspeech.2018-1099

Jorge R. Valdés Kroff and Paola E. Dussias. 2023. Production, processing, and prediction in bilingual codeswitching. In *Psychology of Learning and Motivation*, volume 78 of *Psychology of Learning and Motivation*, chapter 6. Academic Press. https://doi.org/10.1016/bs.plm.2023.02.004