ACCEPTED MANUSCRIPT
To appear in the Quarterly Journal of Experimental Psychology

What Reading Aloud Reveals about Speaking: Regressive saccades implicate a failure to monitor, not inattention, in the prevalence of intrusion errors on function words

Elizabeth R. Schotter[1], Chuchu Li[2], & Tamar H. Gollan[2]

1. University of South Florida

2. University of California, San Diego

Author Note

Elizabeth Schotter is now at the Department of Psychology, University of South Florida.

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders (011492), the National Institute of Child Health and Human Development (079426) and the National Science Foundation (BCS1457159). This research was also funded by the Microsoft Gift Funds, the Atkinson Endowed Chair awarded to Keith Rayner.

We thank Yumeng (Rainy) Gu for help with stimulus development and data collection. We are immensely grateful to Keith Rayner for his support and guidance in the development of this project; he passed away in the middle of data collection. We thank Rob Hartsuiker and two anonymous reviewers for helpful feedback on previous versions of this manuscript.

Correspondence to:
Elizabeth R. Schotter
eschotter@usf.edu
Department of Psychology
University of South Florida
4202 E. Fowler Ave, PCD 4118G
Tampa, FL 33620
Running head: BILINGUAL LANGUAGE INTRUSIONS

Abstract

Bilinguals occasionally produce *language intrusion errors* (inadvertent translations of the intended word), especially when attempting to produce function word targets, and often when reading aloud mixed-language paragraphs. We investigate whether these errors are due to a failure of attention during speech planning, or failure of monitoring speech output by classifying errors based on whether and when they were corrected, and investigating eye movement behavior surrounding them. Prior research on this topic has primarily tested alphabetic languages (e.g., Spanish-English bilinguals) in which part of speech is confounded with word length, which is related to word skipping (i.e., decreased attention). Therefore, we tested 29 Chinese-English bilinguals whose languages differ in orthography, visually cueing language membership, and for whom part of speech (in Chinese) is less confounded with word length. Despite the strong orthographic cue, Chinese-English bilinguals produced intrusion errors with similar effects as previously reported (e.g., especially with function word targets written in the dominant language). Gaze durations did differ by whether errors were made and corrected or not, but these patterns were similar for function and content words and therefore cannot explain part of speech effects. However, bilinguals regressed to words produced as errors more often than to correctly produced words, but regressions facilitated correction of errors only for content, not for function words. These data suggest that the vulnerability of function words to language intrusion errors primarily reflects automatic retrieval and failures of speech monitoring mechanisms from stopping function versus content word errors after they are planned for production.

Bilinguals have a remarkable ability to use an intended language largely without interference from their other language (Gollan, Sandoval, & Salmon, 2011; Poulisse, 1999), even when volitional control over language selection is taken away (i.e., in cued language switching paradigms; Gollan, Kleinman, & Wierenga, 2014; Meuter & Allport, 1999). Although bilinguals rarely produce words in an unintended language, they do occasionally produce *language intrusion errors* – saying a translation-equivalent word that refers to the intended meaning, but not in the intended language. Such errors do not occur randomly, but instead appear to reflect cognitive mechanisms underlying bilingual language selection (Poulisse, 1999) and therefore provide a useful way to study how bilinguals generally achieve successful language control.

Bilinguals produce language intrusion errors systematically when reading aloud mixed-language paragraphs, and these errors follow similar patterns as spontaneously occurring intrusions. For example, similar to intrusion errors produced in spontaneous speech (Poulisse & Bongaerts, 1994), intrusions produced in the read-aloud task are more common for function words than content words (e.g., saying *pero* when the written word was *but*; Gollan & Goldrick, 2016, 2018a; Gollan, et al., 2014; Gollan, Stasenko, Li, & Salmon, 2017; Kolers, 1966; Ratiu & Azuma, 2017). This similarity suggests the source of the errors in the read aloud task is similar to the source in spontaneous speech (i.e., a failure in speech planning and/or monitoring; Gollan et al., 2014; Gollan & Goldrick, 2016) rather than a source specific to reading (i.e., a failure to allocate attention to to-be-said words when reading that would lead to mis-perception errors: Goodman & Goodman, 1977). The strongest evidence for this argument comes from the fact that many intrusions occur on non-cognates, which do not share

orthography or phonology across languages (e.g., *pero—but*). However, previous studies cannot entirely rule out an attention-based account because they tested alphabetic languages, in which function words are shorter than content words and might lead them to receive less visual attention (i.e., to be skipped) during reading.

The current study tests Chinese-English bilinguals (see also Li & Gollan, 2018) to provide further evidence against an inattention account. Any language intrusion errors from English (written with the Roman alphabet, e.g., *book*) into Chinese (written with logographic characters, e.g., 书, pronounced *"shū"* meaning *book*), or vice versa, would be very unlikely to be misperception errors because of striking differences in orthography. This argument is consistent with a recent study which showed robust language switching costs even when bilinguals read inherently univalent stimuli (i.e., Chinese characters that are orthographically distinct from English words; Slevc, Davey, & Linck, 2016). Unlike pictures or Pinyin[1], which could cue a response in either language, Chinese characters should only elicit one response (i.e., in Chinese) and therefore, if possible, should limit the need for response selection. However, the finding of switch costs even with univalent stimuli suggests that strong visual cues cannot function to replace language control mechanisms, and thus should not be enough to eliminate language intrusion errors either.

Additional evidence that a lack of visual attention during reading is not the source of language intrusion errors is the observation that bilinguals often produced intrusion errors even when their eyes directly fixated on the to-be-said target words as they were producing the error (Gollan et al., 2014). Thus, speech (correct or otherwise) and eye movements may be

[1] A Roman alphabetic system that transcribes the pronunciation of Chinese characters.

planned at similar times, and it may be that attention to planned speech, or monitoring thereof, is what determines whether the speech is produced as an error, and if so, whether the error is caught and corrected (Hartsuiker, Corley, & Martensen, 2005; Hartsuiker, Catchpole, de Jong, & Pickering, 2008; Nooteboom & Quené, 2008; for review see Postman, 2000). Speech monitoring may occur externally, when or after speakers hear their own speech (Declerck et al., 2016; Huettig & Hartsuiker, 2010; Laver, 1980), internally, focusing on planned speech before it is produced (Blackmer & Mitton, 1991; Motley, Camden, & Baars, 1982), or both (i.e., there are multiple monitors; Hartsuiker et al., 2005; Hartsuiker, 2006; Hartsuiker & Kolk, 2001; Nooteboom & Quené, 2008, 2017; Oomen, Potsma, & Kolk, 2001). Eye tracking measures can help to pinpoint both the source of errors in the read-aloud task, whether these arise primarily in planning and monitoring of speech, in turn revealing the likely source of intrusion errors in more naturalistic bilingual speech production.

The few studies that have used eye tracking to investigate the amount of direct attention to targets of speech production errors during reading aloud provided an incomplete picture of how these errors pattern. For example, Ratiu & Azuma (2017) found no difference in *gaze duration* (i.e., the duration of time spent looking directly at the target word before leaving it) between language intrusions and correct productions – leaving the source of errors a mystery. Similarly, a study of monolinguals reading aloud in a single language found no differences in skipping or gaze duration on correct productions and substitution errors (e.g., saying *unusable* instead of *usable*; Paulson, 2002). But neither of these studies distinguished between errors that were never corrected and those that were spontaneously detected and self-corrected. Prior research with bilinguals classified intrusion errors as *complete* (i.e., the

entire language-translation word was produced) or *partial* (i.e., the bilinguals caught and corrected themselves before completely producing the intrusion; Gollan et al., 2014; Gollan & Goldrick, 2016; cf. Ratiu & Azuma, 2017 did not specify subtypes of intrusions), but did not also partition out intrusions that were completely produced and then subsequently corrected (i.e., *late corrections*; cf. Gollan & Goldrick, 2018b, who did investigate self-corrections in an aging study with monolinguals reading aloud). These prior studies found that type of error interacts with part of speech: complete intrusions more often involved function than content targets whereas partial intrusions were more common for content than function words. The difference between complete (i.e., uncorrected) intrusions and late corrections may shed light on the role of monitoring.

The predominance of function word targets in bilingual intrusion errors, in both spontaneous speech (Poulisse, 1999) and reading aloud (Gollan & Goldrick, 2016, 2018; Gollan et al., 2017; Kolers, 1966), might relate to their tendency to be shorter, higher frequency, and more predictable than content words; in fact, these same properties of function words lead them to receive less visual attention (i.e., to be *skipped* more) during silent monolingual reading (O'Regan, 1979; Saint-Aubin & Klein, 2001). However, previous studies revealed that many language intrusion error targets were directly fixated (i.e., skip rates for function words are overall low; i.e., 22.2%: Gollan et al., 2014; see also Ratiu & Azuma, 2017).

Skipping rates decrease overall when monolinguals proofread (i.e., monitor for errors; Schotter, Bicknell, Howard, Levy, & Rayner, 2014), but errors are detected less for function words relative to content words, even when controlling for word length (Haber & Schindler, 1981) and whether the word was skipped (Saint-Aubin, Kenny, & Roy-Charland, 2010).

Repetitions of entire words are also less likely to be detected for function words (e.g., "… jumped off *the the* swing and…") relative to content words (e.g., "… jumped off the *swing swing* and…"), even when both instances of the repeated word are fixated (Staub, Dodge, & Cohen, 2018). These studies all involved tasks that specifically required readers to check the text for errors, which suggests that what leads function words to be error prone is failures in monitoring. Because part of speech effects in alphabetic languages are correlated with word length (Gollan & Goldrick, 2018a), effects of skipping are hard to dissociate from part of speech effects (i.e., because word length has a strong influence on skipping; see Drieghe, Brysbaert, Desmet, & De Baecke, 2004). Therefore, Chinese-English bilinguals are an important test case because the reduced variability in word length (i.e., most Chinese words are 1- or 2-characters), and high similarity in length between function and content words, allowed us to unconfound effects of part of speech and word length. Furthermore, the eyes not only fixate on words as they are being encoded, but also make regressions to words when the reader is unsure of what they have read (Bicknell & Levy, 2011; Booth & Weger, 2013; Schotter, Tran, & Rayner, 2014) and these regressions increase when monitoring is increased (i.e., during proofreading; Schotter, Bicknell, et al., 2014). Thus, the current study provides critical information about how and when speech errors are monitored by comparing the eye movements for intrusions errors that are caught and corrected (i.e., either by an internal or external monitor) and uncorrected errors (i.e., those for which both monitors failed).

In summary, we investigated subtypes of language intrusion errors and their relationship to eye movements in reading aloud for Chinese-English bilinguals, which allowed us to investigate (1) whether strong visual cues to language membership suppress language

intrusion errors, and if not, (2) the cognitive mechanisms underlying part of speech effects on intrusion errors (e.g., inattention versus monitoring failures) independent of word length effects. Eye movement recordings allow us to investigate temporal relationships between production and monitoring processes; higher skipping rates and shorter gaze durations would implicate attention prior to or during reading and planning of speech production, whereas regressions would implicate monitoring processes applied after errors are planned for production. To provide further evidence about processes related to production and repair of errors, we compared eye movement measures across error subtypes that differed with respect to success or failure of speech monitoring (i.e., corrected vs. complete errors).

Method

Subjects

Forty-one Chinese-English bilingual undergraduates at the University of California, San Diego participated for course credit. Three of them were excluded for speaking Cantonese rather than Mandarin, 4 were excluded due to poor calibrations of the eye tracker, and 5 were excluded for being English-dominant rather than Mandarin-dominant (dominance was determined by performance on the Multilingual Naming Test (MINT); Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2012; Sheng, Lu, & Gollan, 2014). As a result, twenty-nine subjects (22 Female, 7 Male) remained in the final analyses. Table 1 shows these subjects' characteristics.

Table 1. Subjects' Characteristics.

Characteristic	*M*	*SD*
Age at time of testing	20.8	3.4
Age of Acquisition of English	6.9	3.1
Years lived in English-speaking country	3.1	2.4
Years lived in Mandarin-speaking country	17.7	3.4

	English		*Chinese*	
	M	*SD*	*M*	*SD*
Self-rated spoken language proficiency[a]	5.2	0.9	6.9	0.3
Percent of language use in childhood[b]	14.2	15.8	83.8	20.1
Current percent of language use[b]	57.0	19.8	41.3	19.2
Primary caregiver spoken language proficiency[a]	2.4	1.5	6.9	0.4
Secondary caregiver spoken language proficiency[a]	2.3	1.8	6.5	1.5
Multilingual Naming Test score[c]	24.5	2.9	30.8	1.0

Note: Two subjects were from Taiwan and 27 were from mainland China; the age of acquisition for Mandarin was at birth for all subjects.

a. Proficiency rating scale range: 1 (little to no knowledge) to 7 (like a native speaker).

b. For some subjects, these numbers summed to less than 100% because they spoke more than two languages; for one subject the numbers summed to more than 100% (because of either a clerical error or confusion about the question—these numbers were unaltered).

c. Every other item from the MINT (Multilingual Naming Test; Gollan et al., 2012) was administered, so the maximum possible score was 33.

Materials & Design

Sixteen paragraphs with 106 words, on average, were selected from novels published in both English and Chinese. A Chinese-English bilingual experimenter created two mixed-language versions, approximating mixing frequency in the examples published by Gollan et al. (2014; Kolers, 1966) by starting with a single language version of the paragraph and replacing words with their translation equivalents. These replacements maintained position in the sentence such that one mixed-language paragraph maintained English as the *default language* and the other maintained Chinese as the default language (Gollan & Goldrick, 2018a). Each subject read 16 paragraphs, four in each condition: (a) English only (M_{length} = 109 words, *SD* =

3.29), (b) Chinese only (M_{length} = 101 words, *SD* = 10.38), (c) mixed language with English as default (M_{length} = 111 words, *SD* = 5.23), (d) mixed language with Chinese as default (M_{length} = 104 words, *SD* = 9.42) (see Appendix for examples). Conditions were rotated between subjects and paragraphs in a Latin-Square design and the paragraph order for each subject was randomized. For the mixed language paragraphs, the percent of switch words was 50.04% (SD = 7.62%) and 48.37% (SD = 6.82%) in conditions (c) and (d), respectively (Table 2).

Table 2. The characteristics of the words in each mixed language condition.

		English default language		Chinese default language	
		function	content	function	content
English targets	Percentage	34.97%	28.65%	25.03%	18.27%
	Mean word length	3.09	5.66	3.48	5.50
Chinese targets	Percentage	22.56%	13.82%	24.67%	32.03%
	Mean word length	1.45	1.71	1.36	1.90

Note. The word length for English targets refers to the number of letters of each word, while it refers to the number of characters for Chinese targets

Apparatus

Eye movements were recorded via an SR Research Ltd. Eyelink 1000 Plus eye-tracker in desktop mode, with a temporal resolution of 1000 Hz. A forehead rest was used to minimize head movements, but the chinrest was removed to avoid interference with speech production. After calibration, eye position error was less than .5°. Subjects were seated 60 cm from a 20-inch HP p1230 CRT monitor with a pixel resolution of 1280 x 1024. Although viewing was binocular, only movements of the right eye were recorded. Vocal responses were recorded with a Cyber Acoustics 3.5mm Computer Desktop Unidirectional Noise Cancelling Microphone connected to the display computer via an ASIO driver.

Procedure

After signing a consent form, subjects completed a language history questionnaire to

provide subjective measures of language proficiency, and history of language use and caregiver language use. Following the eye tracking experiment, they named half of the pictures from the Multilingual Naming Test (MINT; Gollan et al., 2012; Sheng, et al., 2014) in each language to provide an objective measure of language proficiency.

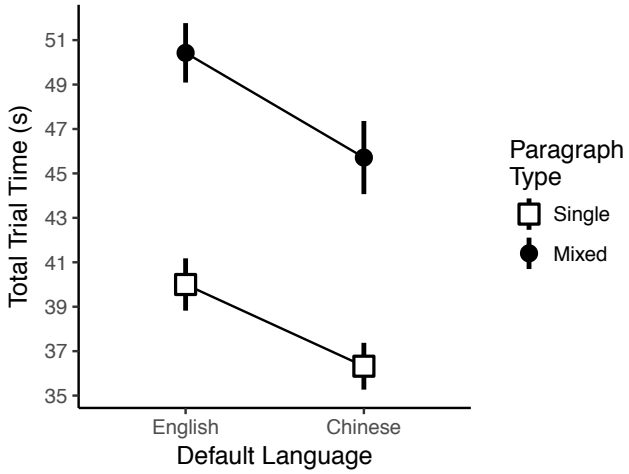
At the start of the experiment, subjects completed a 9-point calibration procedure to allow monitoring of both horizontal and vertical eye movements. At the start of each trial, a black box (65 x 85 pixels) appeared in the top-left corner of the screen, where the paragraph would start. When a fixation was detected, the box disappeared, the paragraph appeared, and the audio recording started. Paragraphs were presented in black on a white background in 20-point Calibri font in a different randomized order for each subject. So that the languages were presented as naturally as possible, English words were separated by spaces but Chinese words were unspaced. Subjects were instructed to read the paragraphs aloud as accurately as possible at a comfortable pace. Vocal responses were recorded separately for each trial; a timestamp was written to the eye tracker file to mark the start of the audio recording. Vocal responses were coded after the experimental session by a Chinese-English bilingual experimenter and the latency from the start of the audio file to the start of the intrusion error and correction phrase were marked manually by the experimenter using Cool Edit software.

Results

To start, we assessed the difficulty of the task by analyzing total paragraph reading time (in milliseconds) with linear mixed effects models with default language (English vs. Chinese), language mixing (single vs. mixed language paragraph) and their interaction entered as centered, crossed fixed effects and subjects and items as random effects with the maximal

random effects structure. Regression coefficients (b) represent effect size in milliseconds, a *t* statistic greater than or equal to 1.96 indicates an effect significant at the .05 level. There was a main effect of default language (bilinguals read Chinese paragraphs faster than English paragraphs; *b* = -4454.2, *SE* = 1231.4, *t* = 3.62), which confirms that these subjects were Chinese-dominant. There was also a main effect of language mixing (bilinguals read mixed language paragraphs more slowly than single language paragraphs; *b* = 9799.1, *SE* = 1016.3, *t* = 9.64), which confirms that mixed language paragraphs are more difficult to read, and there was no interaction (*t* < .46; Figure 1).

Figure 1. Total paragraph reading as a function of default language (English vs. Chinese) and language mixing (Single vs. Mixed language paragraph). Error bars represent +/- 1 SEM.



Error production measures

Language intrusion errors, which only occurred in mixed-language paragraphs (see also Gollan et al., 2014), were identified by a fluent Chinese-English bilingual and were coded for the direction of the intrusion error as follows: (a) *Chinese intrusion*, saying the Chinese translation of a word written in English, (b) *English intrusion*, saying the English translation of word written in Chinese. These errors were further classified into subtypes: (a) *partial intrusions* were

intrusions spontaneously caught by the subject in mid-utterance and self-corrected before they were fully produced, (b) *late corrections* were intrusions that were self-corrected only after they were completely produced, and (c) *complete intrusions* were fully produced intrusion errors subjects never self-corrected. We first analyzed intrusion error rates (collapsing all subtypes) to test whether Chinese-English bilinguals produced the same patterns as previously found with the read-aloud task, and then analyzed relationships between different subtypes and eye movement behavior.

Reversed dominance effects. In previous studies (Gollan & Goldrick, 2016, 2018a; Gollan et al., 2014; 2017; Li & Gollan, 2018; Ratiu & Azuma, 2017), bilinguals more often mistakenly replaced a word written in the dominant language with its nondominant language translation equivalent than the opposite, an effect we refer to as *reversed dominance*. We replicated this pattern; Chinese-dominant bilinguals were more likely to produce English intrusions, saying the English translation instead of the written Chinese word, than vice versa. For all bilinguals, the majority of intrusion errors were English intrusions, and many bilinguals (15 of the 29) produced *no* Chinese intrusions (Table 2; Figure 2)[2]. We tested this with a logistic regression (intrusion was coded as 1 and correct productions were coded as 0), with language of the word as a fixed factor, subject as a random factor with intercepts and the slope for language, and item (individual word) entered as a random effect with intercepts only (models with more complex random effects structures failed to converge) and the effect of language was significant ($z = 3.50$, $p < .001$). In addition, the language that intrudes is more likely to match the default language (Gollan & Goldrick, 2018a), so we conducted a logistic regression

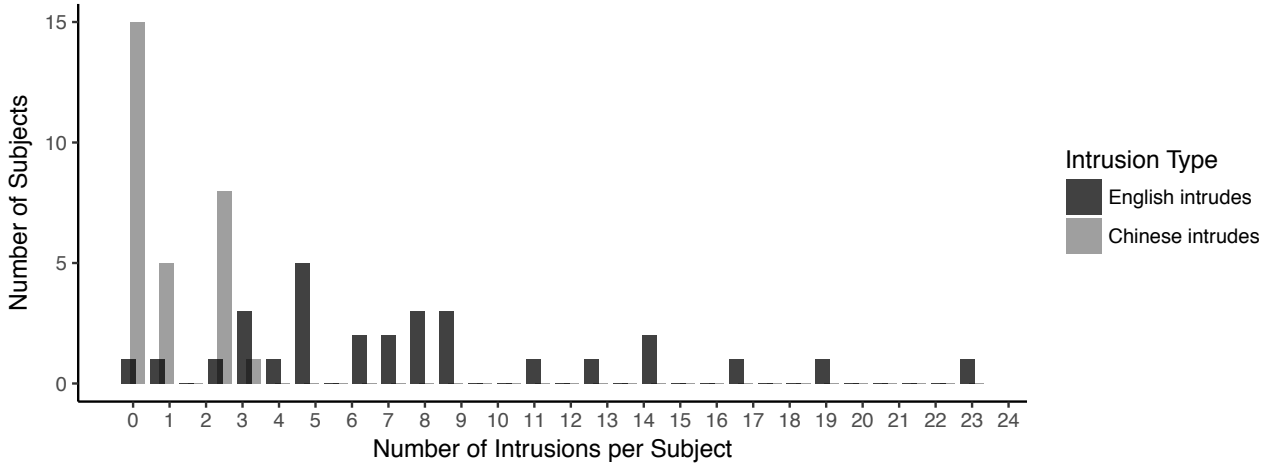
[2] There was only one bilingual who made no intrusions errors in either language.

with fixed factors of the language of the word, the default language of the paragraph, and their interaction as fixed effects with random intercepts only for subject and paragraph number. The effect of language of the word was still significant (z = 11.14, p < .001), as was the effect of default language of the paragraph (i.e., intrusion errors were more likely in paragraphs with English as the default language; z = 3.84, p < .001), as was the interaction (i.e., English intrusions were more likely in English default paragraphs than Chinese default paragraphs – 134 and 87, respectively – but there was no difference for the rare Chinese intrusions – there were 16 in both types of paragraphs; z = -8.39, p < .001). Because of the rarity of Chinese intrusion errors (i.e., Chinese being spoken instead of written English words), the following analyses were performed only on written Chinese words, but we present both in tables and figures for completeness.

Table 2. Average number of intrusions as a function of language of intrusion (aggregated by subject) and percent of English intrusions.

	English Intrudes	Chinese Intrudes	Percent English Intrusion
Mean	7.90	0.83	90.5
SD	5.45	0.97	

Figure 2. Histogram of the number of intrusions per subject in the two languages.



Part of speech effects. Second, we investigated whether Chinese-English bilinguals were more likely to produce intrusion errors on function than on content word targets. We tested this with a logistic regression, with part of speech of the word entered as a fixed factor, subject as a random factor with intercepts and slope for part of speech, and item (word) as a random effect with intercepts only. Of English intrusion errors, the majority were on function words (66.9%) rather than content words, a significant effect of part of speech ($z = 2.02$, $p < .05$).

Because word length in Chinese is less variable than alphabetic languages, we are able to investigate part of speech effects among words of equivalent lengths (see Table 3). To control for the fact that function and content words vary in length, we ran an additional analysis with word length (number of characters, entered as a centered variable) and its interaction with part of speech into the analysis as a fixed effect and random effects for subjects. Of great interest, in this model, the effect of part of speech was still significant ($z = 2.15$, $p < .05$), the effect of word length was not statistically significant ($z = -1.67$, $p = .09$), and neither was the interaction ($z = 1.10$, $p = .27$). Moreover, the fact that the part of speech effect is observed independent of length can be seen in Table 3: intrusion errors are more common on function words than on content words both for 1-character words, for which there are more function words than content words in these paragraphs (509 and 256 words, respectively), and for 2-character words, for which there are more content words than function words in these paragraphs (278 and 407 words, respectively). To consider why function words are more vulnerable to intrusions we compared subtypes of intrusion errors, based on whether and when bilinguals noticed and corrected the errors, and examined eye movement data to consider how monitoring processes might be influencing the observed patterns.

Self-correction Subtypes. When we compared complete intrusions (i.e., errors that were never corrected) to partial intrusions (i.e., errors that were caught and then corrected part-way through) and late corrections (i.e., completely produced errors that were caught and corrected a word or two downstream) a clear difference in the patterns for function and content targets was observed, replicating differences between partial and complete intrusions for different parts of speech reported in other studies (Gollan et al., 2014; Gollan & Goldrick, 2016; Poulisse & Bongaerts, 1994). We analyzed the subject-level rate of English intrusions using a generalized linear model with a Poisson link (to account for the fact that count data are not normally distributed, but rather highly skewed; Agresti, 2012) with fixed effects of part of speech, type of intrusion, and their interaction, and random effects for subject with intercepts and slopes for type of intrusion. Overall, as reported above, intrusions on function word targets outnumbered those on content words ($z = 4.58$, $p < .001$), and there was no difference between late corrections and complete intrusions either in overall rate or with respect to an interaction with part of speech (both ps $> .67$). In contrast, partial intrusions were less common than complete intrusions overall ($z = 2.41$, $p < .05$), and there was a significant interaction with part of speech ($z = 3.11$, $p < .005$; Figure 3), in which partial intrusions exhibited reversed part of speech effects (i.e., bilinguals were more likely to produce partial intrusions on content than on function words). Thus, in addition to being more susceptible to intrusion errors, once planned as errors, function word intrusions are also less likely to be corrected mid-utterance relative to content words. We include partial intrusions in the figures below for completeness and so that they can be compared to prior studies in which they were reported, but given the sparsity of these errors for these bilinguals, we excluded partial intrusions from the analyses.

Figure 3. Average intrusion rate (aggregated by subject) as a function of part of speech (content vs. function), language of intrusion (i.e., English vs. Chinese), and type of intrusion (partial, late correction, and complete) for all word lengths. Error bars represent +/- 1 SEM.

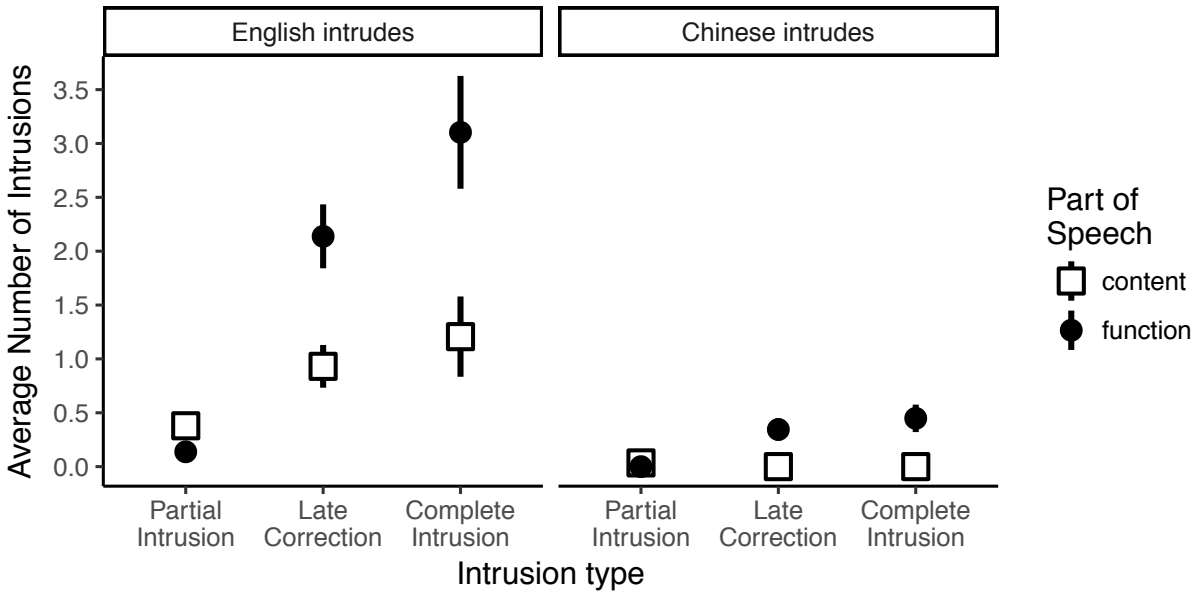


Table 3. Total number of intrusions produced in the experiment by 29 Chinese-English Bilinguals as a function of part of speech (content vs. function word), language of intrusion (i.e., English vs. Chinese), and type of intrusion (partial, late correction, and complete) reported separately for all words (top rows), Chinese words that are one character (middle rows) and Chinese words that are two-characters (bottom rows).

	English Intrudes			Chinese Intrudes		
	Partial intrusion	Late correction	Complete intrusion	Partial intrusion	Late Correction	Complete intrusion
All word lengths						
Function (808)[a]	4	61	89	0	10	13
Content (777)	11	28	36	1	0	0
One-character words						
Function (509)	1	42	64	—	—	—
Content (256)	4	10	20	—	—	—
Two-character words						
Function (278)	1	19	25	—	—	—
Content (407)	7	14	13	—	—	—

[a] *Note:* Numbers in the left-hand column refers to number of Chinese words in one set of the mixed language paragraphs. The numbers for "all word lengths" also include some three-character and four-character words, for this reason the number of intrusions below sums to smaller number than presented in these rows.

Eye movement measures

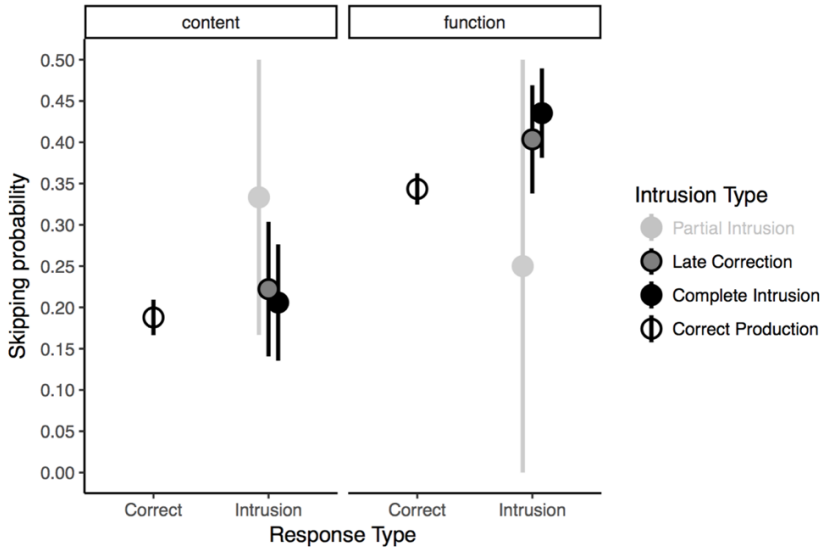
The following analyses contain only words that were susceptible to errors (i.e., for which at least one intrusion error was produced) and Chinese target words, which elicited the majority of intrusion errors (i.e., words for which English translation equivalents were produced instead of written Chinese targets). For gaze duration, values shorter than 50 or longer than 2000 ms were excluded from the analysis. Regressions were classified irrespective of whether the target word was initially skipped.

***Skipping and Gaze duration*.** Based on prior research, we expected that function words would be skipped more than content words. The critical question is whether skipping (i.e., a complete lack of visual attention for function words) caused part of speech effects on intrusion errors. If inattention, or a decrease in attention, to written words were to explain this effect, skipping rate (the likelihood that a word does not receive a fixation before subsequent words are fixated) and gaze durations (the sum of all fixations on a word that is not skipped before the reader leaves it) would show the following patterns. Particularly for function word targets, skip rates would be lowest and gaze durations would be longest on correctly produced targets, then in ascending order of skip rate and descending order of gaze duration we should see partial intrusions, late corrections, and complete intrusions. If not, skip rates and gaze durations would not differ across type of production and/or part of speech.

We analyzed skipping rate with a logistic mixed effects regression. The fixed effects included response type (correct production, late correction, complete intrusion) coded with correct productions as the baseline and three contrasts that compared each of the intrusion types to it independently, part of speech of the word (function vs. content) entered as a

centered predictor, and their interactions. The random effects included intercepts for words crossed with intercepts and the slope for part of speech for subjects. Only the effect of part of speech was significant (function words were skipped more often than content words; $z = 3.37$, $p < .001$); none of the contrasts for the type of response, nor any interactions were significant (all *p*s > .24; Figure 4).

Figure 4. Skipping probability as a function of production (correct, partial intrusion, late correction, or complete intrusion) and part of speech of the word (Chinese words only). Partial intrusion errors, plotted for reference, were not included in the analysis. Error bars represent +/- 1 SEM.

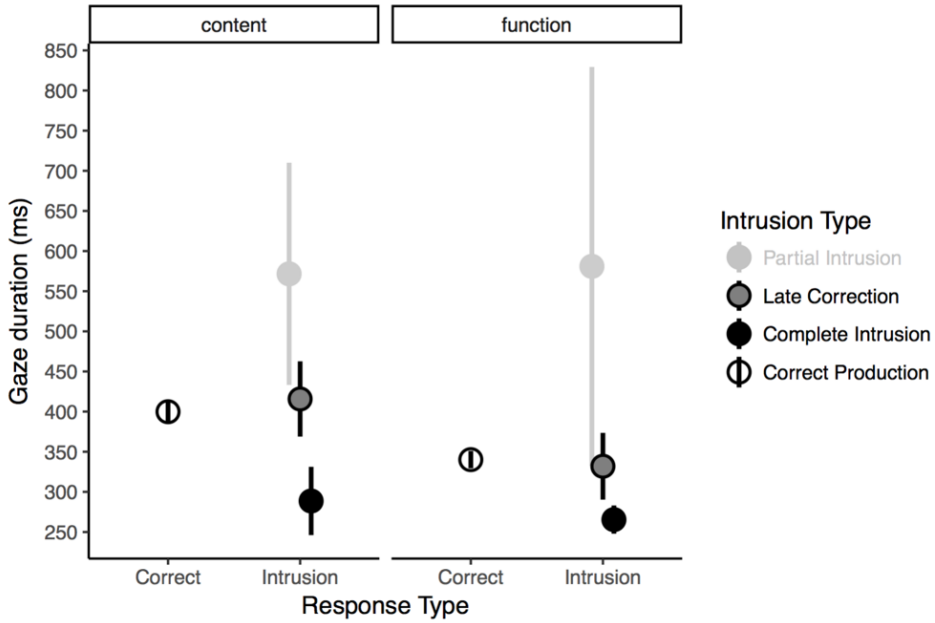


Although function words may indeed be skipped more, these data suggest that is not why they are more susceptible to intrusion errors in the read aloud task. This pattern differs from the interaction between skipping and part of speech reported by Gollan et al. (2014) because word length effects were confounded with part of speech in the alphabetic languages they tested (i.e., for the Spanish-English bilinguals) but are not confounded in Chinese (see also the analysis of word length effects on intrusion error rates, above). To test whether word length was driving the effects of part of speech, we ran an additional analysis that also included word length (for 1- and 2-character words) and its interaction with part of speech in the fixed

effects and random slope for the interaction for subjects. Although this analysis revealed a significant effect of word length (z = -6.43, p < .001) and marginally significant interaction with part of speech (z = -1.71, p = .08), the effect of part of speech remained significant (z = 2.44, p < .05), suggesting that function words were skipped more than content words above and beyond the effect of word length. Critically, the interactions between part of speech and production type remained non-significant (both ps > .3).

We analyzed gaze duration with a linear regression with the same effects structure as the skipping analysis: we used correct productions as the baseline and compared late corrections and uncorrected intrusions to it separately, along with the main effect of part of speech and its interactions with these contrasts (see above). There was a significant effect of part of speech: correctly produced function words received shorter gaze durations than correctly produced content words (b = -57.83, SE = 21.20, t = 2.73). The contrast for response type was significant for complete intrusions (i.e., they were fixated for less time than correctly produced words: b = -91.05, SE = 27.67, t = 3.29) but not for late corrections (i.e., they were fixated for a similar amount of time as correctly produced words: b = 9.60, SE = 29.26, t = .33). However, there were no interactions between part of speech and response type (both ts < .42; Figure 5). Visual attention may be partially related to production of language intrusions, but it cannot explain part of speech effects because function and content words exhibited the same pattern with respect to error subtypes. In the following section we investigate whether regression behavior after moving past the target of the intrusion error can help further explain how failures in monitoring produce different subtypes of intrusion errors.

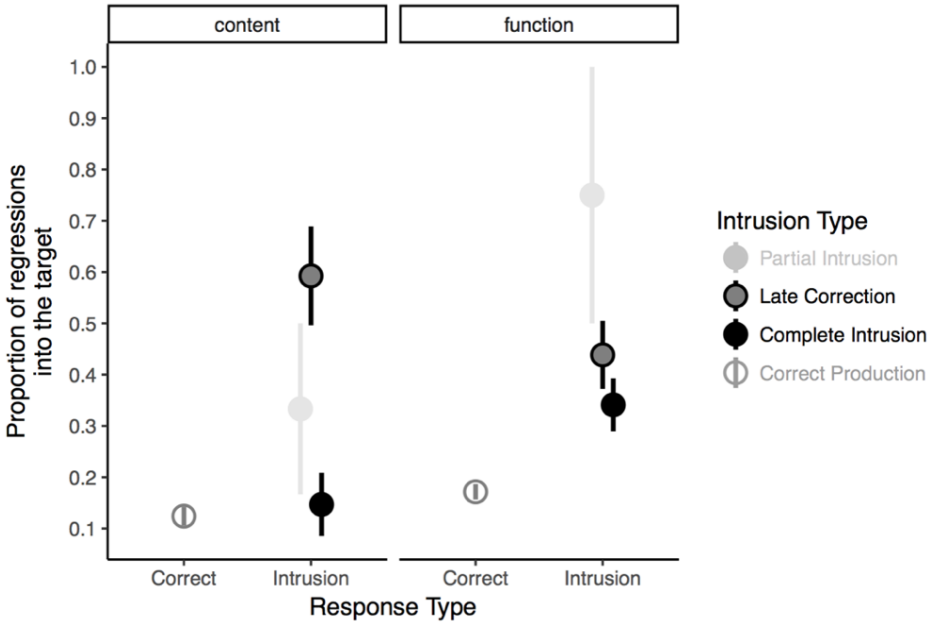
Figure 5. Gaze duration as a function of production type (correct, late correction, or complete intrusion) and part of speech of the written word (Chinese words only). Partial intrusion errors, plotted for reference, were not included in the analysis. Error bars represent +/- 1 SEM.



***Regressions into the target*.** If regressions reflect monitoring of planned or produced errors, they might be associated with self-correction behaviors; regressions should increase for late corrections relative to complete intrusions. Because monitoring processes have nothing to flag for correctly produced utterances, we excluded correctly produced targets from the analysis, but include them in the figure. A logistic regression comparing late corrections (baseline) and complete intrusions, part of speech effects (centered), and their interaction revealed significantly fewer regressions to complete intrusions than to late corrections (z = -3.56, p < .001), no part of speech effect (z = -1.31, p = 1.89), and a significant interaction (z = 2.41, p < .05; Figure 6). Bilinguals made regressions to function word targets words equally often when they were corrected than when they were uncorrected (z = -1.17, p = .24), but were less likely to make regressions to content word targets when they were uncorrected compared

to when they were corrected (z = -2.20, p < .05). These data suggest that errors on function words are more difficult to catch even when overt monitoring behaviors are present (i.e., increased regressions).

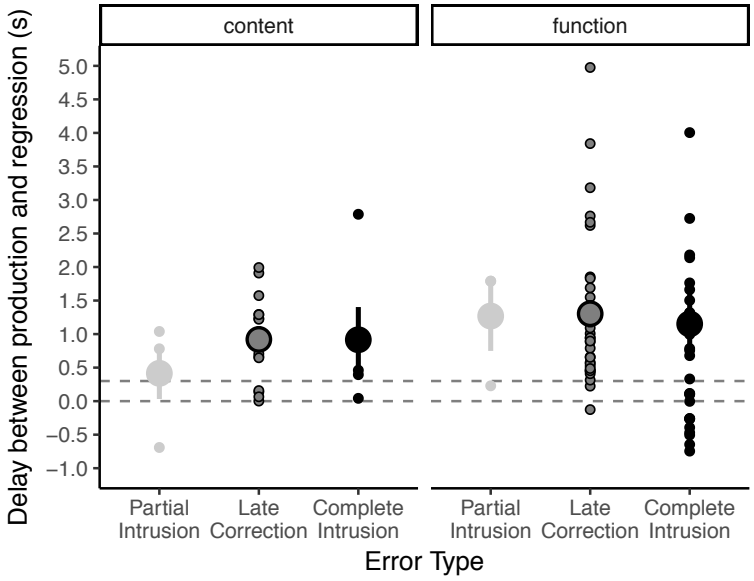
Figure 6. Rate of making a regression into the target as a function of production type (complete intrusion or late correction) and part of speech of the written word (Chinese words only). Correct productions and partial intrusions, plotted for reference, were not included in the analysis. Error bars represent +/- 1 SEM.



Finally, to address whether these monitoring behaviors reflect monitoring of planned versus overtly produced speech, we investigated the time-course of regressions relative to the onset of error production. As shown in Figure 7, most regressions occurred after the onset of overt production of the intrusion error. In fact, the latency between the onset of error production and the regression was almost always longer than 300 ms, which Huettig and Hartsuiker (2010) suggested is the earliest interval during which one can expect to observe eye movements initiated by self-monitoring of overtly produced speech (i.e., listening to oneself). Moreover, virtually all regressions that *did* occur prior to production of the error (thereby possibly involving monitoring of planned speech) involved function word targets for errors that

were never corrected (i.e., complete intrusions). This finding implies an inability to monitor and stop planned production of intrusion errors on function words even when overtly re-inspecting the intended target prior to producing the error[3].

Figure 7. Latency from intrusion error onset to regression onset as a function of error correction type (partial intrusion, late correction, or full intrusion) and part of speech of the word for English intrusions on Chinese words. Small circles represent individual data points, large circles represent the condition mean, error bars represent +/- 1 SEM. Dashed lines at 0 illustrates the onset of speech, and .3 s (300 ms) represents the likely lower bound for onset of monitoring of externally produced speech.



Discussion

The results reported here, particularly the eye movement data, implicate failures of speech planning and monitoring mechanisms, rather than misperception during reading, in the production of language intrusion errors in the read-aloud task. In addition to replicating several previously reported findings with a group of bilinguals with distinct orthographies, these data also provide critical new insights about the vulnerability of function words to monitoring

[3] We checked whether these early regressions (i.e., prior to production) were to correct for unintended skipping of the word, which was not the case. Of all the regressions prior to production, only one occurred for a word that was skipped – this was on a function word target that was a late correction.

failures in general and language selection failures in bilingual speech production. Summarizing the replications of previously reported findings, we observed (a) longer reading times for mixed-language than single-language paragraphs, (b) longer reading times for non-dominant (English) than dominant (Chinese) language paragraphs, (c) reversed dominance effects on intrusion errors, (d) both late corrected and complete (i.e., uncorrected) intrusions more often involved function than content word targets, but partial intrusions exhibited the opposite pattern, and (e) function words were skipped more and received shorter gaze durations than content words.

We also reported a number of novel findings, which provided new information about the vulnerability of function word targets to intrusion errors. Specifically, complete intrusions elicited shorter gaze duration than correctly produced targets and late corrected intrusions. However, despite what an inattention account would predict, gaze durations were *not* longest for correctly produced targets, they were equivalent for correct productions and late corrected intrusions, and furthermore, partial intrusions elicited the longest gaze durations. These gaze duration patterns were also found to the same extent for function and content words. In contrast, regressions showed a different pattern for function and content words, suggesting that monitoring after speech is planned leads function words to be more vulnerable than content words to language intrusion errors. Regressions to target words were more common for late corrected than for complete intrusions on content word targets, whereas regressions were equally likely for late corrected and complete intrusions for function word targets. Finally, when regressions did occur, almost all of them were initiated after overt production of the error with the exception of function word intrusions that were never corrected (i.e., complete

intrusions). We discuss the implications of each of these results in turn.

First, it is striking that Chinese-English bilinguals produced language intrusion errors in the read-aloud task even though visually distinct orthographies provided clear cues to language membership, and intrusions followed similar patterns as were previously reported in Spanish-English bilinguals. Both these findings reduce the likelihood that inattention during reading leads to intrusion errors in the read-aloud task. Previously we suggested that reversed dominance and default language effects reflect global inhibition of the dominant language to allow language mixing (Gollan et al., 2014; Gollan & Goldrick, 2018; Raitu & Azuma, 2017); the replication of this result in a different population of bilinguals implies broad applicability of this bilingual control mechanism (see also Li & Gollan, 2018). The persistent susceptibility of function words to intrusion errors is informative because in English, Spanish, and French function words tend to be shorter than content words. In the present study, intrusion errors were more common for function than content words, even when length was unconfounded with part of speech, an analysis that was only possible in this study due to the characteristics of the Chinese writing system. This suggests that function words are vulnerable to intrusion errors above and beyond the fact that they tend to be very short, skipped in the read-aloud task, and poorly attended in general in speech production (Poulisse & Bongaerts, 1994).

Even though function words were skipped more than content words overall (see also Angele & Rayner, 2013; Gautier, O'Regan, & Le Gargasson, 2000; O'Regan, 1979), the probability of skipping did not differ between correctly produced words and intrusion errors (cf. Gollan et al., 2014 with Spanish-English bilinguals for whom word length and part of speech were confounded). Gollan and Goldrick (2018) did statistically control for word length in a study

of Spanish-English bilinguals, but given the almost non-overlapping distributions of length between the two parts of speech in those languages the comparison in Chinese is important to demonstrate that this effect is observed independent of length effects. This echoes the pattern seen in proofreading among monolinguals, which involves an increase in monitoring and more visual attention to the text (Schotter, Bicknell, et al., 2014): errors are more likely to be missed in function than in content words (Haber & Schindler, 1981; Staub et al., 2018), even when comparing among fixated or skipped words (Saint-Aubin et al., 2010) and when controlling for word length (Saint-Aubin & Klein, 2001). Thus, while inattention does appear to increase intrusion errors, part of speech effects on either language intrusion or proofreading errors do not appear to be caused by decreased overt attention; as we explain below, instead, part of speech effects might reflect more automatic planning of function words and failures of monitoring mechanisms *even when increased attention is applied*.

Eye movement behaviors coupled with division of intrusions subtypes based on failure or success of monitoring processes revealed further insights. Gaze durations were shorter for words that were produced as complete intrusions (i.e., that were never corrected) but were longer for partial intrusions (i.e., errors that were corrected mid-utterance). The latter result must be interpreted with caution (given the very low number of partial intrusion errors), but provides some indication that increased attention can play a role in determining which errors are successfully monitored *prior to overt production of intrusion errors*. However, these patterns were similar for function and content words, suggesting that decreased attention cannot explain part of speech effects (i.e., the susceptibility of function words to intrusion errors). Perhaps eye movements and production (errors) are planned separately, with a

potential to influence each other through monitoring; longer gaze provides an opportunity to monitor and rapidly correct an error, if produced (i.e., partial intrusions), and shorter gaze provides little opportunity for monitoring processes to catch and correct errors. The lack of a difference between gaze duration on correctly produced and late corrected error targets may suggest that some types of monitoring are not, or cannot be, applied until after production is overt (i.e., that monitors of internally planned and externally produced speech differ in how they operate). For similar arguments about distinct cognitive mechanisms underlying early interrupted versus later repaired speech errors in other paradigms see Hartsuiker et al. (2005, 2008; Nooteboom & Quené, 2008). Note that the decrease in gaze duration that we reported for complete intrusions relative to late corrections differs from the non-significant difference reported by Ratiu and Azuma (2017; Paulson, 2002). However, in that study, there was no distinction between errors that were corrected and those that were not. In our study, late corrections had equivalent gaze durations as correctly produced words but complete intrusions had shorter (and partial intrusions had longer) gaze durations than correctly produced words: collapsing these types of errors would obscure these differences and lead to the null effects.

In contrast to skipping and gaze duration, regressions back to the target did differ by part of speech. For content word targets, regressions were more likely for late corrections than for uncorrected errors (i.e., complete intrusions), whereas for function word targets, regressions were as likely for late corrections as for uncorrected errors. Because most regressions occurred after onset of production of the intrusion error in a time-course that suggests monitoring of overtly produced speech (i.e., regressions rarely occurred prior to production of the error, and almost only did so *for errors on function words that were never*

corrected), any internal monitor that was applied was not very effective. This is particularly surprising given that language intrusion errors (at least for content word switches) stand out quite obviously, and are identified more quickly than semantic errors within a single language in explicit monitoring tasks (i.e., in which participants press a button when they identify the error; Ivanova, Ferreira, & Gollan, 2017). Taken together, the patterns we observed are consistent with proposals that speakers monitor their speech both *before* they produce an error (i.e., via an internal monitor – as suggested by gaze durations), and *after* they produce an error (i.e., via an external monitor – as suggested by regressions). As such, the current study joins other work in implicating monitoring at multiple points during speech production (Hartsuiker & Kolk, 2001), and in multiple stages of bilingual speech production (Gollan et al., 2014).

Even though these data implicate monitoring mechanisms in bilingual intrusion errors, which may be analogous to other types of monitoring failures (e.g., in proofreading), it is still an open question as to why function words are more susceptible to *being produced as intrusion errors*. Function words may be more likely to be targets of intrusion errors primarily because they are more likely to be automatically retrieved, possibly due to their predictability (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009) or ability to be inferred from context, which would require less perceptual processing of their surface form (Staub et al., 2018). Even when equated for orthographic length and overt attention, function words were still more likely to be produced as errors. As Staub et al. (2018) suggested, the failure of monolinguals to detect errors of function word repetitions or omissions may be more easily attributed to eye movement control error. In contrast, these language intrusion errors are unlikely to be attributed to that because of their distinct orthographic form. However, recent research

suggests that readers may sometimes fixate a word when they had intended to skip it based on parafoveal processing (but could not because the saccade program was past the point of cancellation; Schotter & Leinenger, 2016), during this *forced fixation* they do not encode the fixated word, but instead "skip it with their mind" (Schotter, Leinenger, & von der Malsburg, 2018; see also Schotter, von der Malsburg, & Leinenger, 2018).

Unlike previous work on proofreading (in monolinguals), in the present study intrusion targets and errors (for both content and function words) were equated for meaning (i.e., the intrusion errors are translation equivalents of the intended target words); thus, despite being production errors, they did not violate expected meaning or grammatical class. Although we suggested part of speech effects primarily reflect failures of late-operating monitoring mechanisms in catching these errors, we note that the persistent finding part of speech effects in production of speech errors is consistent with previous suggestions that bilinguals select a default language at the level of syntax (Gollan & Goldrick, 2018), there are different retrieval mechanisms for function versus content words (Garrett, 1975, 1982), and that bilingual control mechanisms operate differently over word classes and are sensitive to the default language (e.g., the Matrix Language Framework; Myers-Scotton & Jake, 2009). Once selected for production, retrieval of a function word may be both more automatic than that of a content word (Gollan & Goldrick, 2018; 2019), and more ballistic so that it cannot be stopped once planned - even in the presence of overt behavior suggesting attempted monitoring.

References

Angele, B., & Rayner, K. (2013). Processing the in the parafovea: Are articles skipped automatically? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 649-662.

Bicknell, K., & Levy, R. (2011). Why readers regress to previous words: A statistical analysis. *Proceedings of the Cognitive Science Society*.

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*, 173-194.

Booth, R. W., & Weger, U. W. (2013). The function of regressions in reading: Backward eye movements allow rereading. *Memory & Cognition*, *41*, 82-97.

Branzi, F. M., Della Rosa, P. A., Canini, M., Costa, A., & Abutalebi, J. (2015). Language control in bilinguals: monitoring and response selection. *Cerebral Cortex*, *26*(6), 2367-2380.

Declerck, M., Lemhöfer, K., & Grainger, J. (2016). Bilingual language interference initiates error detection: Evidence from language intrusions. *Bilingualism: Language and Cognition*.

Garrett, M. F. (1975). The analysis of sentence production. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 9, pp. 133–177). New York, NY: Academic Press.

Garrett, M. F. (1982). Production of speech: Observations from normal and pathological language use. In A. Ellis (Ed.), *Normality and Pathology in Cognitive Functions* (pp. 19–76). London, United Kingdom: Academic Press.

Gautier, V., O'Regan, J. K., & Le Gargasson, J. F. (2000). The-skipping revisited in French: programming saccades to skip the article les. *Vision Research*, *40*, 2517-2531.

Gollan, T.H., & Goldrick, M. (2016). Grammatical constraints on language switching: Language control is not just executive control. *Journal of Memory and Language*, 90, 177-199.

Gollan, T. H., & Goldrick, M. (2018a). A Switch is Not a Switch: Syntactically-Driven Bilingual Language Control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44,* 143-156.

Gollan, T. H., & Goldrick, M. (2019). Aging deficits in naturalistic speech production and monitoring revealed through reading aloud. *Psychology and Aging*. in press

Gollan, T. H., Kleinman, D., & Wierenga, C. E. (2014). What's easier: Doing what you want, or being told what to do? Cued versus voluntary language and task switching. *Journal of Experimental Psychology: General*, *143*, 2167-2195.

Gollan, T. H., Sandoval, T., & Salmon, D. P. (2011). Cross-language intrusion errors in aging bilinguals reveal the link between executive control and language selection. *Psychological Science, 22,* 1155–1164.

Gollan, T.H., Schotter, E.R., Gomez, J., Murillo, M., & Rayner, K. (2014). Multiple levels of bilingual control: Evidence from language intrusions in reading aloud. *Psychological Science, 25*, 585-595.

Gollan, T. H., Stasenko, A., Li, C., & Salmon, D.P. (2017). Bilingual language intrusions and other speech errors in Alzheimer's disease. *Brain & Cognition*, 118, 27-44.

Gollan, T.H., Weissberger, G., Runnqvist, E., Montoya, R.I., & Cera, C.M. (2012). Self-ratings of spoken language dominance: A multi-lingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism: Language and Cognition, 15,* 594-615.

Goodman, K.S., & Goodman. Y.M. (1977). Learning about psycholinguistic processes by analyzing oral reading. *Harvard Educational Review, 47*, 317-333.

Haber, R.N., & Schindler, R.M. (1981). Error in proofreading: Evidence of syntactic control of letter processing? *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 573-579.

Hartsuiker, R. J. (2006). Are speech error patterns affected by a monitoring bias? Language and *Cognitive Processes, 21*, 856–891.

Hartsuiker, R. J., Catchpole, C. M., de Jong, N. H., & Pickering, M. J. (2008). Concurrent processing of words and their replacements during speech. *Cognition*, *108*(3), 601-607.

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related reply to Baars et al. (1975). *Journal of Memory and Language*, *52*(1), 58-70.

Hartsuiker, R. J., & Kolk, H. H. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive psychology*, *42*(2), 113-157.

Holbrook, M. B. (1978). Effect of subjective verbal uncertainty on perception of typographical errors in a proofreading task. *Perceptual and Motor Skills*, *47*, 243-250.

Huettig, F., & Hartsuiker, R. J. (2010). Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes*, *25*, 347-374.

Ivanova, I., Ferreira, V.S., & Gollan, T.H (2017). Form overrides meaning when bilinguals monitor for errors. *Journal of Memory and Language*, 94, 75-102.

Jared, D., Levy, B. A., & Rayner, K. (1999). The role of phonology in the activation of word meanings during reading: evidence from proofreading and eye movements. *Journal of Experimental Psychology: General*, *128*, 219-264.

Kolers, P. (1966). Reading and talking bilingually. *American Journal of Psychology*, *3*, 357–376.

Laver, J. D. (1980). Monitoring systems in the neurolinguistic control of speech production. *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand*, 287-305.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 33*, 41–103.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Li, C., & Gollan, T. H. (2018). Cognates interfere with language selection but enhance monitoring in connected speech. *Memory & Cognition*, online first.

Meuter, R. F. I., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language, 40,* 25–40.

Motley, M. T., Camden, C. T., & Baars, B. J. (1982). Covert formulation and editing of anomalies in speech production: Evidence from experimentally elicited slips of the tongue. *Journal of*

Verbal Learning and Verbal Behavior, *21*, 578-594.

Myers-Scotton, C., & Jake, J. (2009). A universal model of code-switching and bilingual language processing and production. In B. Bullock & A. Jacqueline Toribio (Eds.), *The Cambridge handbook of linguistic code-switching* (pp. 336–357). New York, NY: Cambridge University Press.

Nooteboom, S., & Quené, H. (2008). Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors. *Journal of Memory and Language*, *58*(3), 837-861.

Nozari, N., Dell, G. S., & Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive psychology*, *63*(1), 1-33.

Nozari, N., & Novick, J. (2017). Monitoring and control in language production. *Current Directions in Psychological Science*, *26*, 403-410.

Oomen, C. C., Postma, A., & Kolk, H. H. (2001). Prearticulatory and postarticulatory self-monitoring in Broca's aphasia. *Cortex*, *37*, 627-641.

O'Regan, J.K. (1979). Eye guidance in reading: Evidence for the linguistic control hypothesis. *Perception & Psychophysics, 25,* 501-509.

Paulson, E.J. (2002). Are oral reading word omissions and substitutions caused by careless eye movements? *Reading Psychology*, *23*, 45-66.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*, 97-132.

Poulisse, N. (1999). Slips of the tongue: Speech errors in first and second language production. Amsterdam, The Netherlands: John Benjamins.

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics, 15,* 36–57.

Ratiu, I., & Azuma, T. (2017). Language control in bilingual adults with and without history of mild traumatic brain injury. *Brain and Language*, 166, 29–39.

Saint-Aubin, J., Kenny, S., & Roy-Charland, A. (2010). The role of eye movements in the missing-letter effect revisited with the rapid serial visual presentation procedure. *Canadian Journal of Experimental Psychology 64*, 47-52.

Saint-Aubin, J., & Klein, R.M. (2001). Influence of parafoveal processing on the missing-letter effect. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 318-334.

Schotter, E.R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition*, *131*, 1-27.

Schotter, E.R., & Leinenger, M. (2016). Reversed preview benefit effects: Forced fixations emphasize the importance of parafoveal vision for efficient reading. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 2039-2067.

Schotter, E.R., Leinenger, M., & von der Malsburg, T. (2018). When your mind skips what your eyes fixate: How forced fixations lead to comprehension illusions in reading. *Psychonomic Bulletin & Review*, online first.

Schotter, E. R., Tran, R., & Rayner, K. (2014). Don't believe what you read (only once) comprehension is supported by regressions during reading. *Psychological Science*, *25*, 1218-1226.

Schotter, E.R., von der Malsburg, T., & Leinenger, M. (2018). Forced fixations, trans-saccadic integration, and word recognition: Evidence for a hybrid mechanism of saccade triggering in reading. *Journal of Experimental Psychology: Learning Memory & Cognition,* online first.

Sheng, L., Lu, Y., & Gollan, T. H. (2014). Assessing language dominance in Mandarin–English bilinguals: Convergence and divergence between subjective and objective measures. *Bilingualism: Language and Cognition*, *17*, 364-383.

Staub, A., Dodge, S., & Cohen, A. L. (2018). Failure to detect function word repetitions and omissions in reading: Are eye movements to blame? *Psychonomic Bulletin & Review*, online first.

Appendix

An example paragraph in each of the four experimental conditions. Each paragraph was presented only once to each subject, counterbalanced across conditions, in a within-subjects Latin square design.

(a) The train slows as it reaches the heart of the city, and I sit up to watch the smaller buildings that overlook the marsh. I hold the handle and lean out just enough to see where the tracks go. They dip down to street level just before they bend to travel east. I breathe in the smell of wet pavement and marsh air. They train dips and slows, and I jump. My legs shudder with the force of my landing, and I run a few steps to regain my balance. I walk down the middle of the street, heading south, toward the marsh.

(b) 到达城中心后，火车慢了下来，我坐起身，看着原本渺小的建筑物一点一点清晰变大，在那里可以俯瞰沼泽。我抓住车厢把手，探身出去，想看清轨道去往哪里。它们先下行到与街面齐平，然后一路蜿蜒向东。我在街面和沼泽地散发的潮湿的气味中呼吸着。火车开始往下行驶，速度也慢了下来，我趁机跳下车。因为落地时的冲撞，两腿有些发抖，我往前跑了几步，才恢复了平衡。我走在大街中间，转向南，朝沼泽的方向出发。

(c)The train slows 当它 reaches the 心属于 the city, and 我坐 up 来 watch the smaller buildings 那俯瞰 marsh. I 握住 the handle 再 lean 出去 just 够去看 where the tracks 去往. They dip 下 to 街面齐平 just 前 they 弯 to travel 东. I 呼吸进 the smell of wet pavement 和 marsh 空气. The train dips 和慢, and I 跳. 我的腿 shudder 因为 the 力属于 my landing, and 我跑些 steps 来 regain my 平衡. I 走 down the 中 of the street, 向 south, toward that 沼泽。

(d) 到达城 heart after，火车 slows 下来，我 sit 起 body，看着 overlook 沼泽 that 渺小的建筑物。我 hold of train 把手，探身 out 去，enough 想 look 清轨道 travel where。它们 first 下行到 with 街面齐平，然后一路弯曲 travel 向 east。我 at 街面和沼泽地 emit 的潮湿的气味中 breath 着。火车 begin 往 down 行驶，速度 also slow 下来，我趁机 jump 下车。with 落地时的冲撞，legs 有些发抖，I 往前 run a few 步，才恢复了 balance。我 walk 在大街 middle of，turn 向南，towards marsh 的 direction 出发。