An enhanced model of gemination in spelling: Evidence from a large corpus of typing errors

Christopher R. Hepner (chepner3@jhu.edu)
Department of Neurology, Johns Hopkins University
1629 Thames Street, Suite 350, Baltimore, MD 21213, USA

Svetlana Pinet (spinet1@jhmi.edu)
Department of Neurology, Johns Hopkins University
1629 Thames Street, Suite 350, Baltimore, MD 21213, USA

Nazbanou Nozari (nozari@jhu.edu)
Department of Neurology; Department of Cognitive Science, Johns Hopkins University
1629 Thames Street, Suite 350, Baltimore, MD 21213, USA

Abstract

Geminates (or double letters) are a feature of many languages, including English. Studies of the spelling errors produced by individuals with orthographic working memory deficits have provided evidence that geminates are not produced as two independent instances of the same letter. Instead, there must be a special mechanism in the orthographic system that produces geminates. Several theories have attempted to model such mechanisms. However, in most cases, the predictions of such theories have been tested using data from single-case neuropsychological studies. In the current study, we re-evaluate these theories using the largest corpus of geminate errors in typing collected to date, and show that no theory can explain all the findings. We then propose an enhanced model of gemination that can.

Keywords: double letters; geminates; typing; orthographic working memory; graphemic buffer

Introduction

To type a word (e.g., *broom*) to dictation, a sequence of phonemes (/bɹum/) must be converted into a sequence of letters (B-R-O-O-M). In neurotypical adults, this can be accomplished by either serially mapping phonemes to graphemes using the sublexical route, or by retrieving the whole letter sequence in parallel from long-term memory (LTM) using the lexical route. In the lexical route, hearing a word activates its representation in phonological LTM, which activates the word meaning in the lexical semantic system. This, in turn, activates the word spelling (BROOM) in orthographic LTM (O-LTM). Orthographic information is then processed by orthographic working memory (O-WM, often referred to as the *graphemic buffer*), responsible for maintaining the orthographic representation and selecting its letters in sequential order to pass them on to effector-specific motor plans, e.g., sequences of key presses in typing.

In many languages, including English, spellings sometimes contain double letters or *geminates* (e.g., O in BROOM). Findings from neuropsychological studies of spelling disorders suggest that gemination is more than just two independent instances of the same letter. For example, Fischer-Baum and Rapp (2014) describe a patient who produced more geminate additions in non-geminate words (e.g., MARK → MARRK) after spelling a geminate word (e.g., BROOM) than a non-geminate word (e.g., BROAD). The perseveration of the gemination independently of the letter identity implies the existence of a special geminate feature (see also Caramazza & Miceli, 1990). Other models, e.g., McCloskey et al. (1994), also propose a special mechanism for gemination, but without proposing a geminate feature.

The majority of the data on which models of gemination are based come from case studies of individuals with selective damage to O-WM. These individuals produce errors across spelling modalities (writing, typing, and spelling out loud) which increase in frequency as a function of word length. While extremely valuable in principle, the utility of neuropsychological data can be limited by the relatively small number of errors of interest, as well as individuals' idiosyncrasies. This is perhaps the reason why, despite several elegant proposals, no consensus has been reached on this topic. In this study, we have created a large corpus of geminate errors from 100 neurotypical adults each typing 800 geminate words. Using this corpus, we test current theories of gemination and demonstrate that none of them is sufficient to explain all of the findings. We then propose an enhanced model which accounts for both current and previous findings on geminate errors.

Theoretical Accounts of Gemination

Any model of gemination must accommodate two basic assumptions: (a) It must include a representation of letter order in addition to letter identity; otherwise, words such as DOG and GOD would be indistinguishable. A full review of models of segment sequencing in language production is beyond the scope of this paper, but it is important to note that chaining models and their variations do not provide a satisfactory explanation of O-WM errors. (b) It must have a special mechanism for geminate production, beyond treating geminates as two independent instances of the same letter. Generally speaking, two classes of models have been proposed: *geminate feature* models (Caramazza & Miceli,

1990; Fischer-Baum & Rapp, 2014) which represent the geminate as an independent feature, and *geminate links* models (McCloskey et al., 1994) which represent the geminate by linking two adjacent slots in the positional frame to the same letter identity.

In the current paper, we pick two models that meet both of the criteria above and are each representative of one class of gemination models. The first model, McCloskey et al. (1994), henceforth referred to as M1994, is described above. The second model, Glasspool and Houghton (2005), henceforth referred to as G&H2005, combines a geminate feature model with a *competitive queuing* mechanism for sequencing segments. This mechanism uses Initiate and End nodes to dynamically establish a gradient of activation such that activation is highest for the letter in the current position and progressively lower for subsequent letters. This gradient is implemented in an Item layer akin to a positional frame[1] separate from, but connected to, the letter identity representations. Letters are selected by a competitive winner-take-all process (implemented in the model as a *competitive filter*), and the produced letter is temporarily inhibited to prevent perseveration. The geminate feature is represented by a separate node, which, like the letter identity nodes, receives activation from a single slot in the Item layer (i.e., the starting geminate position; e.g., 3 in BROOM). If the geminate feature's activation passes a threshold, it sends a signal to output production processes to repeat the production of the last segment, after which it is inhibited just like letter identity representations.

We test the predictions of these two models on our geminate error corpus to evaluate whether either, or both, can account for all the findings.

Methods

One hundred native English speakers (56 females, M_{age} = 34, age range: 18–67 years), who had passed spelling and typing proficiency pretests, participated for payment through Amazon Mechanical Turk. Participants completed two sessions of a timed typing to dictation task. The stimuli were 600 words, 5-16 letters long, comprising 400 experimental words with a single geminate and 200 filler words without a geminate. All 600 words were presented auditorily in each session in randomized order, and participants typed them before a deadline (300 ms + 180 ms per letter) with an ITI of 1000 ms, with breaks after every 50 trials.

Results

No response was produced on 78 experimental trials. Of the remaining 79,922 responses, 18,865 (23.60%) contained at least one error. Of those, 3,894 (20.64%) consisted of a single error affecting the geminate. This "clean" set was used in the analyses. All the error types obtained in this study, with the exception of Splits (e.g., BOROM; discussed in the Error Categories section) have also been reported in handwriting studies, making it unlikely that we are looking at typing-specific errors. Moreover, both the length effect (more errors on longer words) and the position effect (more errors in the middle positions) that are typical of O-WM deficits were evident in our data; $t = -51.07, p < .001$ for the length effect, and $t = -4.03, p < .001$ for the position effect. We can thus conclude with reasonable confidence that the errors in our corpus are representative of the same cognitive processes that have been investigated by previous studies of gemination.

Error Categories

Table 1 presents the error types of interest, their definitions, and examples.

Geminate Deletions. M1994 explains these errors by assuming that one of the geminate links has been lost and a repair process has removed the corresponding slot in the positional frame. G&H2005 explains them by assuming that the geminate feature has failed to reach the activation threshold. Thus both accounts explain basic deletions. However, a closer look at the data show that the probability of a geminate deletion is much higher if the target geminate letter appears in the wrong position than if any other letter appears in the wrong position. In the set of 1,283 errors containing letter movements but no letter additions, deletions, or substitutions, 34.62% of responses (36 out of 104) with the target letter in the wrong position had geminate deletions, compared to only 11.37% (134 out of 1,179) with a non-target letter in the wrong position, $\chi^2 = 42.94,\ p < .001$. This finding, which implies interdependence between the letter identity and the gemination process, is not expected from either account.

Geminate Additions. M1994 accounts for additions through "reloading", a mechanism by which a degraded representation can be refreshed by retrieving it again from O-LTM. If the degraded and the newly-loaded representations have geminates in different positions, the result is an addition. According to G&H2005, additions happen when the geminate feature reaches the activation threshold in more than one position. We report three empirical findings regarding geminate additions and evaluate the two models in light of each. The first is the distribution of geminate additions around the target geminate position. Figure 1a plots this distribution for 568 geminate additions in the current dataset. The probability of geminate additions drops quickly the farther the position gets from the target position. In fact, the only position where the rate of geminate addition is higher than chance is position −1, with 263 errors observed compared to 107.49 expected, $\chi^2 = 95.38, p < .001$. Since the process for geminate additions proposed by M1994 involves a geminate shift, it can account for the increased likelihood of additions closer to the target position using the same mechanism (see below). G&H2005 predict

[1] G&H2005 view this layer as also coding some information about abstract letter identities, although the nature of such information has not been clearly specified.

Table 1: Error categories, definitions, and examples with their respective error counts.

Error type	Definition	Example	Count
Geminate deletions	Only one copy of the geminate letter has been produced.	BROOM → BROM	1,853
Geminate additions	Both the original geminate letter and another letter in the target spelling have been doubled.	BROOM → BRROOM	568
Geminate shifts	Another letter in the target spelling was doubled instead of the original geminate letter.	BROOM → BRROM	846
Substitutions	The geminate has been substituted by two copies of a different letter, either from within the sequence or from outside.	BROOM → BRBBM BROOM → BRXXM	38
Exchanges	The original geminate letter has swapped positions with another letter. The response contains a double letter, which may or may not be the same as the geminate letter in the target.	BROOM → BORRM BROOM → BOORM	19
Pseudosubstitutions	One of the two copies of the geminate letter has been replaced by another letter, either from within the sequence or from outside.	BROOM → BROBM BROOM → BROXM	390
Splits	One of the two copies of the geminate letter has exchanged with an adjacent letter, splitting the geminate.	BROOM → BOROM	180

that geminate additions should be most likely in positions adjacent to the target position because, due to the gradient of activation across positions, adjacent positions have the next highest activation after the target position, so they activate the geminate feature more strongly than other non-target positions. Thus both accounts predict that geminate additions should occur more often in positions closest to the target geminate position.

The second finding is related to the first one: while the probability of a geminate addition at position −1 is significantly higher than chance, the same probability at position +1 is significantly *lower* than chance, with 33 errors observed compared to 88.47 expected, $\chi^2 = 27.34, p < .001$. M1994 does not have a mechanism to account for this. In G&H2005, the geminate feature gets inhibited after it has affected production. In order to be activated again, it needs to overcome this suppression. When the first production occurs before the target position, the chance of recovering from inhibition at the target position is still good, because that position has a link to the geminate feature that can directly activate it. However, if the first production occurred at the target position, it is unlikely that the noise alone can overcome the inhibition enough to produce the geminate feature again in the next (i.e., +1) position. G&H2005 can thus account for this finding.

Finally, the data suggest that, in words with an additional copy of the target geminate letter (e.g., C<u>O</u>COON), geminate additions are much more likely on that additional copy than on any other letter, e.g., p(COCOON → COOCOON) > p(COCOON → COCOONN). In the 267 geminate addition errors in which there was another copy of the target letter, 79 (29.59%) of the additions occurred on that additional copy compared to 51.07 expected by chance, $\chi^2 = 7.37, p = .007$. This finding, which suggests a link between the letter identity and the geminate feature, is especially intriguing because second copies of the target letter must, by definition, be more than one position away from the target geminate position. Thus the propensity for geminate additions to occur on the same letter identity seems to override the strong tendency for additions to occur close to the target position. Since neither M1994 nor G&H2005 have any mechanisms to bind the geminate feature to letter identity, neither model can account for this finding.

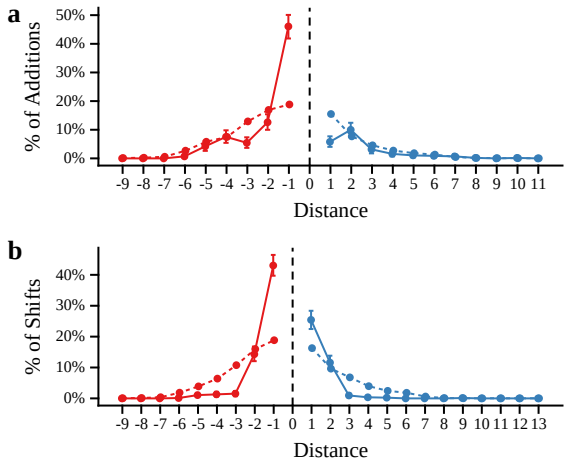

Figure 1. Distributions of geminate (a) additions and (b) shifts. Solid lines are observed proportions and dashed lines are proportions expected by chance. Error bars are 95% CIs.

Geminate Shifts. According to M1994, these errors occur when at least one of the two geminate links is detached and a repair process reconnects the broken links to the wrong letter. In G&H2005, shifts are caused by the activation of the geminate feature in the wrong position. We examine the same three patterns we reported above for geminate additions, this time for geminate shifts. Figure 1b plots the distribution of the 846 geminate shifts in the current dataset. Similar to additions, the probability of geminate shifts drops

quickly the farther the position gets from the target position. Only positions −1 and +1 around the target geminate show significantly higher than chance probability of a geminate shift; position −1: 364 observed vs. 159.33 expected, $\chi^2 = 114.76, p < .001$; position +1: 215 observed vs. 137.69 expected, $\chi^2 = 20.86, p < .001$. M1994 predicts this pattern, because more distant movements require more links to be broken and reattached. G&H2005 also predicts this finding because the gradient of activation across positions causes adjacent positions (and thus the contents of those positions) to have more activation than distant positions.

The strong asymmetry between −1 and +1 positions observed in geminate additions is not visible here. Both positions show higher than chance probability of shifts, a finding that both M1994 and G&H2005 can account for (see the explanation of the same finding for geminate additions). Finally, in words with an additional copy of the target geminate letter (e.g., C<u>O</u>COON), we examined whether a geminate is more likely to shift to that additional copy than to any other letter, e.g., p(COCOON → COOCON) > p(COCOON → COCONN). Unlike geminate additions, this comparison did not reveal a special status for the additional copy of the target geminate in geminate shifts. In the 358 geminate shift errors in which there was another copy of the target letter, 49 (13.69%) occurred on that additional copy compared to 65.49 expected by chance, $\chi^2 = 2.49, p = .943$. Thus, both M1994 and G&H2005 can explain these findings.

Substitutions. Substitutions happen when a letter is replaced by another letter from either within or outside the target sequence. This is easily explained by both M1994 and G&H2005 (and any other theory that views letter representations as separate from positions). A closer examination of the data revealed that the target geminate letter is much less likely to participate in substitutions than any other letter in the word: out of the 1,204 responses in which the only error was a substitution, only 3.16% (38 out of 1,204) affected the target geminate letter, despite geminates accounting for 13.43% (1,204 out of 8,963) of the opportunities for these errors, $\chi^2 = 103.58, p < .001$[2]. In M1994, the double links between adjacent positions and the letter identity could provide a mechanism for binding the letter more tightly to its position. In G&H2005, on the other hand, there is no mechanism to account for this pattern.

Exchanges. Two distinct patterns of exchanges are of particular interest to us: *position-preserving* errors, in which the target geminate letter has been exchanged (i.e., swapped positions) with another letter, but the original position of the geminate has been preserved (e.g., BROOM → BORRM) and *identity-preserving* errors, in which a similar swap between the target geminate and a non-geminate letter has happened, but this time, the geminate has remained attached to the target geminate letter rather than to its target position (e.g., BROOM → BOORM). Our corpus contains 65 exchanges without additional letter insertions, deletions, or substitutions. In this set, both patterns occur more often than expected by chance: 23 position-preserving errors compared to 10.51 expected by chance, $\chi^2 = 5.31, p = .011$, and 35 identity-preserving errors compared to 12.69 expected, $\chi^2 = 15.04, p < .001$. Moreover, *identity-preserving* errors were significantly more common than *position-preserving* errors, $\chi^2 = 3.77, p = .026$. The propensity for exchanges to preserve geminate position is predicted by both M1994 and G&H2005, because the representation of the geminate is connected to position in both of these models. However, neither model would predict higher than chance probability of identity-preserving errors or its greater probability than position-preserving errors, because there is no mechanism for binding the geminate representation directly to letter identities in either model.

Pseudosubstitutions. In M1994, pseudosubstitutions occur when one of the two links to the target geminate letter is broken and a repair process fills the empty position with a different letter. G&H2005, or any theory that proposes a single slot for the geminate letter in the positional frame, can only explain pseudosubstitutions as a combination of two independent errors: a geminate deletion and a letter insertion adjacent to the geminate. If that were the case, pseudosubstitutions should be less common than *either* of those errors individually. However, there are significantly more pseudosubstitutions (390) than single-letter insertions adjacent to the geminate (258) in our set of 18,865 incorrect geminate word spellings, $\chi^2 = 26.95, p < .001$, ruling out the double-error explanation.

Splits. These errors (N = 180) made up 4.52% of all the geminate errors in our corpus, which is more than any studies of handwriting. We thus suspect that splits might be specific to typing. In keeping with this assumption, splits happened more often when the target geminate and the intruding letter were typed with different hands than the same hand, both for anticipations (e.g., BROOM → BROMO) $\chi^2 = 23.19, p < .001$, and perseverations (e.g., BROOM → BOROM) $\chi^2 = 12.05, p < .001$. We thus conclude that these errors most likely arise during motor programming specific to the typed modality, which is outside of the scope of theories discussed in this study.

Table 2 provides a summary of the empirical results reported for each error category and indicates whether M1994 and/or G&H2005 can account for that finding. The two models successfully explain a wide range of empirical findings on geminate errors, but neither model in its current form can account for all of the empirical findings. Three classes of issues can be identified: (a) Cases that can be accounted for by M1994, but not G&H2005. The common origin of these is the fact that M1994 proposes two slots for

[2] The effect of gemination is reliable even after position is taken into account.

Table 2: Comparison of empirical findings to predictions of previous models.

Finding	M1994	G&H2005
1. Basic geminate deletions	✓	✓
a- Higher probability of geminate deletions when target geminate letter moves	×	×
2. Basic geminate additions	✓	✓
a- Positional distribution	✓	✓
b- Suppression in +1 position	×	✓
c- Higher probability of gemination of another copy of the target geminate letter	×	×
3. Basic geminate shifts	✓	✓
a- Positional distribution	✓	✓
b- Significantly more geminate shifts in both −1 and +1 positions	✓	✓
c- No increased probability of gemination of another copy of the target geminate letter	✓	✓
4. Basic substitutions	✓	✓
a- Lower probability of substitutions affecting the target geminate than other letters	✓	×
5. Basic exchanges	✓	✓
a- Higher than chance probability of position-preserving errors	✓	✓
b- Higher than chance probability of identity-preserving errors	×	×
c- Higher probability of identity-preserving than position-preserving errors	×	×
6. Basic pseudosubstitutions	✓	×

the geminate letter in the positional frame, but G&H2005 proposes only one. (b) Cases that can be accounted for by G&H2005, but not by M1994. The single instance of this (2b in Table 2) stems from the presence of an inhibition mechanism on the geminate in G&H2005 after the geminate feature affects production, which is absent in M1994. (c) Finally, there are cases where both models fail to explain the finding. The common feature of these cases is that they suggest an interdependence between the letter identity and the gemination process that is absent in both M1994 and G&H2005. In the next section, we propose a model that integrates these three features into the basic framework of G&H2005, and show that the enhanced model can account for all the empirical findings reported here.

The Enhanced Geminate Model

We maintain the general architecture of G&H2005 (Figure 2), with three modifications: (1) The positional frame contains two slots instead of one for the geminate letter, similar to M1994 (feature 1 in Figure 2). (2) While in the original G&H2005 model the geminate feature affects output processes (e.g., repeating the motor program for producing the last letter), the enhanced model proposes that the main function of the geminate feature is to block the inhibition of letters after their production (feature 2 in Figure 2). As can be seen in the figure, a single geminate node sends an inhibitory signal to all the self-inhibitory connections to letter identities; however, only the letter that has been just produced would have an activated self-inhibitory connection. Thus the geminate feature has a focal effect on that particular letter. The novel mechanism we have proposed here for the operation of the geminate feature has an important advantage over G&H2005: in that model, when the letter in the geminate position is reached, e.g., the first O in BROOM, a signal is sent to the output processes to repeat the production of the O. The geminate feature must then be suppressed, otherwise more copies of O will be produced. As acknowledged by the authors, this suppression makes it hard for the model to account for double geminates, e.g., BALLOON, as well as the many adjacent geminate addition errors, e.g., BRROOM, observed in our data. The enhanced model, on the other hand, has no problem with double geminates. When the first L in BALLOON is produced, the geminate feature inhibits the L's self-inhibition, thus keeping it activated for re-selection in the next position (i.e., the second L). The geminate feature is inhibited for the next selection step so that extra copies of L are not produced, but it is released from that inhibition afterwards, allowing it to repeat the process when the first copy of O is selected. Because there are no words in English with three consecutive identical letters (e.g., BROOOM), this simple rule of "inhibit the geminate feature for one step after it has exerted its effect" can account for all gemination patterns in English. Finally, (3) the enhanced model differs from G&H2005 in that it proposes a link between the target geminate letter and the geminate feature, such that the letter can activate the geminate feature directly (feature 3 in Figure 2).

When BROOM is to be produced, the operation of the system is similar to G&H2005 until the third positional slot is reached. Unlike G&H2005, not only the slot, but also the target geminate letter O sends activation to the geminate feature. This double source of activation ensures that the geminate feature passes the threshold in most cases where gemination is required. When O wins the competition among the letters, it is produced in the third position. Normally, it would be immediately inhibited after production, but the activated geminate feature inhibits this inhibition process, whereby allowing O to win the competition again when it receives activation from the fourth positional slot.

The geminate feature itself undergoes inhibition once it has exerted its influence. Thus, under normal circumstances, it would not pass the activation threshold again, even though the newly activated O in the fourth position tries to reactivate it (this would lead to BROOOM-like errors). Since the geminate feature is unlikely to pass the threshold here, O will undergo post-production inhibition and the next letter in the sequence will win the competition for the next slot.

Because the enhanced model preserves all the important basic features of G&H2005, it is expected to account for all the findings that that model accounts for. These are explained earlier in the paper and we will not reiterate them here.[3] Instead, we focus on the findings that were not explained by one or both models. Findings 4a and 6 in Table 2 were accounted for M1994, but not G&H2005, whereas 2b was only explained by G&H2005. The enhanced model can account for 4a and 6, because it adopts M1994's assumption of double slots for the geminate, as well as 2b by virtue of the post-production inhibition mechanism from G&H2005. Table 2 shows four additional cases where neither M1994 nor G&H2005 could account for the data (1a, 2c, 5b and 5c). All of these cases point to a connection between the target geminate letter and the geminate feature, which is specified in the enhanced model's new feature 3. Since these findings and the specific feature in the model that is proposed to account for them are new, we unpack the mechanism for each one below.

Higher probability of geminate deletion when the target geminate letter moves than when non-target letters move (1a). In the enhanced model, in the absence of noise, the geminate feature can only pass the threshold of activation necessary for its operation if its input is the summed activation of both the positional slot and the target letter. When the target letter is activated in the wrong position, the geminate feature no longer receives the summed activation, which causes a geminate deletion. During the movement of non-target letters, the geminate feature still receives the summed activation, making deletions less likely. It is important to note that under noisy circumstances, activation of either the letter identity or the positional slot may be enough to push the geminate feature above the threshold, just not as robustly as when the summed input is received.

Higher probability of a geminate addition on another copy of the target geminate letter than on any other letter (2c). Since the enhanced model includes a connection from the target letter identity to the geminate feature, the geminate feature will receive activation from the letter whenever it is selected, even in the non-target position. It would thus be more likely for noise to push the activation of the geminate feature over the threshold, compared to when it is receiving no activation from the letter representations.

Higher probability of identity-preserving exchanges compared to chance (5b) and compared to position-preserving exchanges (5c). Both of these findings also point to the fact that the target geminate letter directly activates the geminate feature, even when it appears in the wrong position.

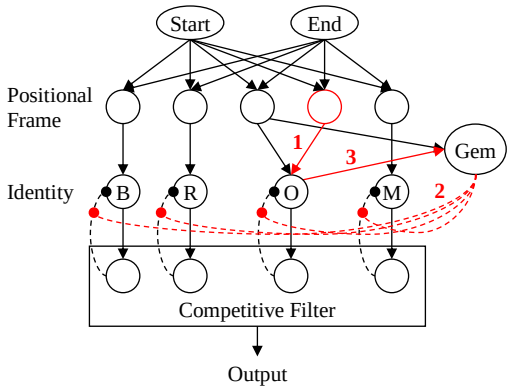

Figure 2: The enhanced geminate model.

Conclusions

The enhanced model proposed in this paper integrates key insights from previous gemination models, but views the gemination process as primarily consisting of inhibiting the self-inhibition of the most recently produced letter. This function, together with a direct link between the target geminate letter and the geminate feature, allows the model to account for all of the empirical data that, to our knowledge, have been reported on geminate errors.

References

Caramazza, A., & Miceli, G. (1990). The structure of graphemic representations. *Cognition*, *37*(3), 243–297.

Fischer-Baum, S., & Rapp, B. (2014). The analysis of perseverations in acquired dysgraphia reveals the internal structure of orthographic representations. *Cognitive Neuropsychology*, *31*(3), 237–265.

Glasspool, D. W., & Houghton, G. (2005). Serial order and consonant–vowel structure in a graphemic output buffer model. *Brain and language*, *94*(3), 304-330.

McCloskey, M., Badecker, W., Goodman-Schulman, R. A., & Aliminosa, D. (1994). The structure of graphemic representations in spelling: Evidence from a case of acquired dysgraphia. *Cognitive Neuropsychology*, *11*(3), 341–392.

[3] We have verified that the changes in the enhanced model do not affect its ability to explain the patterns accounted for in the past models. Due to space constraints, however, we were forced to limit the discussion to cases not accounted for by other models.