Precision of phonological errors in aphasia supports resource models of phonological working memory in language production

Jenah Black & Nazbanou Nozari

Carnegie Mellon University, Pittsburgh, PA, USA
Center for the Neural Basis of Cognition (CNBC), Pittsburgh, PA, USA

Corresponding Author: Jenah Black, Carnegie Mellon University, Dept. of Psychology, 5000 Forbes Avenue, Pittsburgh, PA 15213. Email: jenahb@andrew.cmu.edu

Abstract

Working memory is critical for many cognitive functions and language production is no exception. Despite their differences, all accounts of working memory agree on its capacity limitation. Two dominant models have been proposed to account for such capacity limitation: slot models and resource models. In recent years, resource models have found support in both visual and auditory perception. An important question is whether they also extend to *production*. We investigate this issue by analyzing sublexical errors from four individuals with aphasia. Using tools from computational linguistics, we first define the concept of "precision" of sublexical errors. We then demonstrate that such precision decreases with increased working memory load, i.e., word length, as predicted by resource models. Finally, we rule out alternative accounts of this effect, such as articulatory simplification. These data provide the first evidence for the applicability of the resource model to production and further point to the generalizability of this account as a model of resource division in working memory.

Working memory is critical to many cognitive functions, including language production (Jacquemot & Scott, 2006). It is fairly obvious that speakers must be able to maintain a coherent message in memory as they produce sentences, but more intriguing is the existence of a working memory process sensitive specifically to information regarding sounds (as opposed to concepts and words). This is often referred to as the "phonological buffer". In brief, the role of the phonological buffer is to maintain the activation of phonological representations until they are utilized in further processes, such as articulatory mapping for spoken production or phoneme/grapheme conversion for writing (Caramazza et al., 1986; Goldrick & Rapp, 2007).

A central feature of working memory, including the phonological buffer, is its limited capacity (Baddeley, 2003). For example, shopping for a large number of groceries without a list may result in leaving the store without a vital ingredient. Similarly, an increase in word length increases the probability of sublexical errors (e.g., Caramazza et al., 1986). But how is capacity determined and how is this limited resource divided among the items in working memory? There are two primary accounts of resource limitation in working memory: *slot models* and *resource models*. Slot models (e.g., Cowan, 2001; Cowan et al., 2012) assume a fixed upper bound on the number of items that can be stored in memory. Items are stored perfectly if they get a slot and not stored at all if they do not get a slot. In contrast, resource models (e.g., Ma et al., 2014) view working memory as a continuous resource without a fixed number of items as its limit. Any increase in the number of items causes some decrease in the precision of memory, even if small.

Resource models have been supported by evidence in visual perception as well as auditory perception (e.g., Bays, Gorgoraptis, Wee, Marshall, & Husain, 2011; Hepner & Nozari, 2019). But they have not been tested in language *production*. This paper aims to fill this gap. There is, however, a nontrivial methodological problem that must be resolved first. Recall that resource models predict a gradual decrease in accuracy as a function of increased working memory load. Such a decrease may be too small to be captured with binary accuracy (i.e., correct/incorrect). Rather, the empirical test of resource models relies on the notion of error "precision", or the distance between the error and the target. For example, mistakenly recalling a green patch as a blue patch is a more "precise" error than recalling the patch as red, because green is closer to blue on the color spectrum than red. While similarity as a continuous measure is intuitive in some domains, such as color (e.g., most people can easily imagine a spectrum of colors between green and blue), it is much less intuitive in the domain of phonological production (e.g., most people cannot readily imagine a spectrum of sounds between /b/ and /p/). The goal of the paper is thus two-fold: (a) to develop a suitable measure of precision in the phonological domain, and (b) to use this measure to test the prediction of slot vs. resource models of working memory in language production by analyzing errors from four individuals with aphasia.

1.1. Slot versus Resource Models of Working Memory

As briefly mentioned above, two general accounts have been proposed to explain the capacity limitation of working memory: slot models and resource models. Slot models (e.g, Cowan, 2001) posit that working memory is composed of a number of slots, each of which can hold one to-be-remembered item. The upper bound of this number, *K*, is the capacity of working memory. The value of *K* has historically been examined using span tasks. Similar to remembering your grocery list, a span task requires participants to remember a series of items (such as numbers or words) and report them back after a delay. For example, Miller (1956) summarized many of these classic experiments on span, concluding that humans could maintain 7 ± 2 items in working memory. Critically, this estimate held across items of varying complexity, such as digits (e.g., "3, 2, 8"), letters (e.g., "M, Q, E"), and words (e.g., "cat, tree, wrench"), suggesting that the quantity of items, rather than the content of the item, was the key factor for limits on information storage. The fact that memory was unaffected by the complexity of items aligned with data demonstrating the utility of "chunking" in memory (e.g., Gobet & Simon, 1998; see Gobet et al., 2001 for a review). If a higher-order chunk could be stored in memory (e.g., a word) then the lower-order units need not be stored (e.g., the individual phonemes in the words). Thus, across the population and item complexity, *K* would be 7 ± 2 according to Miller (see also Cowan, 2001, who considers *K* to be closer to 4). If the number of to-be-remembered items is less than *K*, then all items can be remembered accurately. However, if the number of items exceeds *K*, then any excess items are completely forgotten and can only be guessed. Thus, a slot-based model of working memory assumes that performance will be at ceiling for small sets but will decrease as the number of items increases beyond *K.*

In contrast to slot models, *resource models* (e.g., Ma et al., 2014) do not propose a concrete cap on the number of items that can be held in working memory. Rather, they view working memory as a continuous resource with no fixed upper bound, divided between all items. Naturally, as the set size increases, each item receives less of the resource and thus will be remembered less *precisely*. Measuring "precision" was not common in classic memory paradigms. For example, after studying a number of items, participants were usually asked to recall an item (with or without a cue), or to pick the item among several alternatives. They either succeeded or failed, providing only a binary measure of behavior. This changed with the proposal of the *continuous reproduction* paradigms (Wilken & Ma, 2004) in the visual domain. In an example paradigm, a number of color swatches are presented in various locations on the screen. After a short delay, a cue appears in one of the locations, followed by a continuous color wheel. Participants click the point on the color wheel that they believe most closely matches the color swatch in the cued location. Precision is indexed by how close participants' choices are to the target colors on the wheel; the closer to the target, the more precise the response. Empirical findings using the continuous reproduction paradigms provided strong support for the resource models of working memory in visual processing (Bays et al., 2011; Oberauer & Lin, 2017; van den Berg et al., 2012; Wilken & Ma, 2004).

More recently, Hepner and Nozari (2019) adapted the continuous reproduction paradigm to speech perception. For each of the four pairs of syllables, /bɑ/–/dɑ/, /kɑ/– /gɑ/, /ɹɑ/–/lɑ/, and /sɑ/–/ʃɑ/, recorded from natural speech, they created five equally spaced intermediate artificial syllables by changing the acoustic properties of the initial consonant, while leaving the vowel unchanged. For each pair, a different acoustic property was manipulated to minimize interference in working memory due to featural overlap (see Hepner & Nozari, 2019, for details). This procedure resulted in four continua, each containing seven syllables. After a familiarization phase, participants were presented with 1, 2 or 4 syllables from different continua on each trial. After a short delay, a visual slider appeared for one of the continua presented on that trial (e.g., /b/-/d/). Participants used the continuous slider to mark where a particular syllable that they heard on that trial was located between the two endpoints. Similar to the findings of the continuous paradigm in visual processing, the authors observed a monotonic increase in the deviation between the remembered and the actual syllable as the set size increased (see also Joseph et al., 2015). This increase was evident even when the set size changed from a single sound to two sounds, which is aligned with prediction of resource models, but not slot models, since K in most neurotypical individuals is greater than 2.

Collectively, the results of visual and auditory studies discussed above support the resource models over the slot models. Moreover, they show that despite the very different nature of representations in these two domains, the same computations apply to both. However, all the findings discussed so far are in *perception*. Do resource models also apply to *production*? We answer this question by testing the prediction of resource models in the domain of phonological working memory in language production.

1.2. Phonological buffer and the notion of precision of sublexical errors

The language production system can be generally understood as a set of hierarchical levels, moving from abstract representations (e.g., lexical items) to more grounded representations (e.g., motor codes). There are at least two distinct stages of processing (Dell, 1986): in the first stage, semantic features of an object or concept activate the corresponding lexical item, or lemma (Levelt, 1989), along with the other lexical items that share some of those features. The second stage begins by selecting a lemma and proceeds by mapping it on to representations which translate it into sound. This stage itself includes multiple layers of representation, including the phonological level, the motor planning level, and the motor execution level (Goldrick & Rapp, 2007). At the phonemic level, phonemes activate their respective phonetic features and phonological rules are applied, generating allophonic variations that are sensitive to context. Finally, phonetic representations are translated into a plan of motor movements, which are executed to produce speech.

Reflecting the general organization of the language production system, language errors fall under two general categories, lexical and sublexical (e.g., Fromkin, 1971; Garrett, 1975). A lexical error reflects a problem in the first stage of processing and is marked by selecting the incorrect lexical item (e.g., “dog” for “cat”). In contrast, a sublexical error reflects a problem in the second stage of processing and is marked by selecting an incorrect segment for a correct lexical target (e.g., “gat” for “cat”). The language production system exhibits two additional properties, cascading and interactivity (Dell, 1986). Cascading refers to the continuous influence of higher over lower layers of production. Interactivity refers to the influence of the lower layers of representations on higher layers (Nozari & Dell, 2009; Pinet & Nozari, 2018; see Kurowski & Blumstein, 2016 for cascading activation in aphasia; see Dell et al., 2014 for review; see Rapp & Goldrick, 2000, and Pinet & Nozari, 2023, for the limits of interactivity). These properties affect error patterns. For example, if a /k/ is mistakenly produced as a /g/, the erroneous /g/ has a longer VOT than a /g/

that was correctly produced for a /g/ target, a finding that indicates partial activation of the phoneme /k/ and its cascading into the phonetic feature space, even though it was not fully selected (Goldrick & Blumstein, 2006). Similarly, sublexical errors are more likely to be real words than nonwords, because of feedback between the segments (e.g., phonemes) and lexical items (e.g., Nozari & Dell, 2009). The same feedback makes two lexical items more likely to exchange segments if they already have a segment in common (e.g., "top" and "fog" are more likely to interact than "tip" and "fog"; Dell, 1984; Pinet & Nozari, 2018). These findings show that errors reflect the underlying properties of the hierarchical system from which they arise.

Models of language production have not always integrated the idea of the phonological buffer. Similarly, theories of memory have not traditionally considered speech errors and their properties as falling within the scope of memory. However, the evidence suggests that memory and language production are closely entwined. Early evidence in favor of a phonological buffer was provided by Baddeley (1984), who showed that phonological overlap between stimuli decreases recall accuracy. This finding was later complemented by neuropsychological evidence from individuals who showed a similar pattern of producing phonologically related errors in reading, writing, and auditory word repetition (Bisiacchi et al., 1989; Bub et al., 1987; Caramazza et al., 1986; Shallice et al., 2000). Interestingly, such errors were more common on nonword than word targets (McCarthy & Warrington, 1984; Nozari et al., 2010). As nonwords do not enjoy the top-down semantic-lexical support that words do, their production is more strongly dependent on the intact operation of a phonological buffer (Nozari & Dell, 2013). This also highlights the critical role of the phonological buffer during language acquisition (e.g., Brooks & Macwhinney, 2000), when most input is initially perceived as nonwords.

As implied by its name, the phonological buffer is a post-lexical buffer, i.e., it operates on phonemes, and not higher-level representations, such as words (e.g., Caramazza et al., 1986; Goldrick & Rapp, 2007). However, the sensitivity of processing in patients with the buffer damage to lexical status and in some cases to lexical frequency (e.g., Bub et al., 1987; Shallice et al., 2000) suggests an influence of the lexical representations on the operations of the buffer. Critical for our purpose is the link between the phonemes and the phonetic features, as this relationship is the key factor in determining the "precision" of a sublexical error. Specifically, precision can be defined in terms of the number of phonetic features that two phonemes have in common. For example, producing a /p/ for the target /b/ is a more precise error than a /s/ for the same target, because /p/ and /b/ share more phonetic features, and are thus closer in the articulatory phonetic space.

There are a number of reasons to believe that sublexical errors are sensitive to phonetic features. For example, phonemes that share phonetic features are more likely to interact (The phonetic similarity effect; Caramazza et al., 1986; MacKay, 1970; Shattuck-Hufnagel & Klatt, 1979). This results from the cascading activation between phonemes and phonetic features. Figure 1 shows a schematic of this interaction when the lemma "cat" activates its component phonemes. When a given phoneme, e.g., /k/ is activated, it activates its articulatory phonetic features, e.g., velar and stop, via cascading activation. Those features in turn activate other relevant phonemes through feedback. In this case, both features also converge on phoneme /g/, while only the stop feature activates /d/. Thus /g/, which is more similar to /k/ due to more shared features, gains greater activation and is thus more likely to be erroneously selected for production than the more dissimilar phoneme /d/.

The phonetic similarity effect suggests that phonological errors can indeed be sensitive to the articulatory phonetic features. This allows us to assess how precise they are, by measuring the distance between the error and the target in the phonetic space. Changes to precision of sublexical errors can then be used to test the predictions of resource models of working memory in language production.

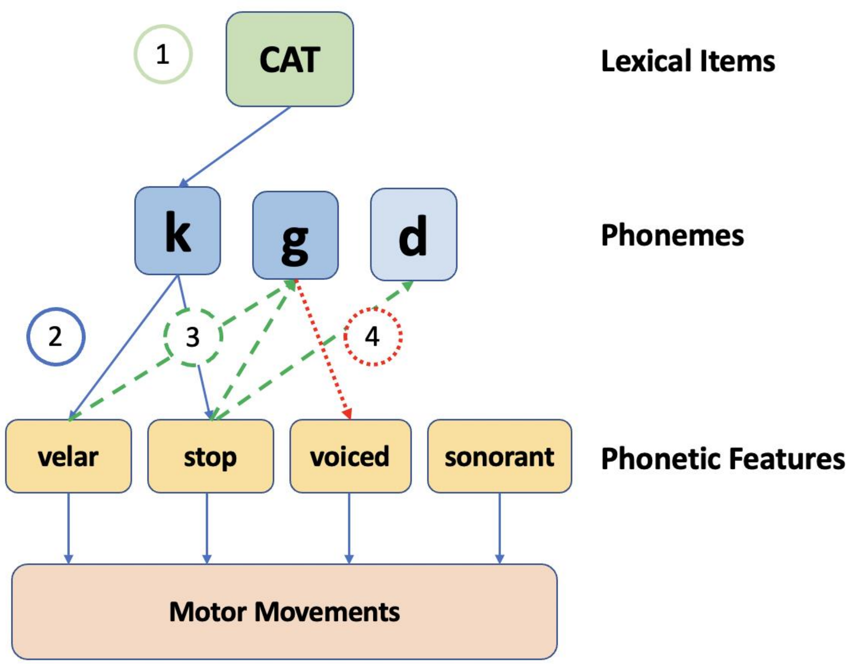

Figure 1. The mechanism underlying the phonetic similarity effect. This example shows why the onset /g/ is a more likely error for target /k/ than the onset /d/. (1) Activation of the 'cat' lexical item, (2) Feedforward activation of phonetic features via the phoneme /k/, (3) Feedback activation of the phonetically related phoneme /g/ as well as the competitor phoneme /d/. While the /d/ phoneme is partially activated by the 'stop' feature, the /g/ phoneme is activated more strongly because it shares more phonetic feature with /k/. (4) Feedforward activation of the phonetic features of /g/, including the erroneous 'voiced' feature, leading to the production of /g/ instead of /k/.

1.3. The Current Study

The aim of the current study was to use the precision of sublexical errors to test the prediction of resource vs. slot models of working memory in language production. If a resource model is applicable to the phonological buffer, the precision of the sublexical error should systematically decrease as the working memory load increases. In the context of word production, an increase in the working memory load is defined as an increase in the word length. Thus, the specific prediction of the resource model is that sublexical errors in longer words should be phonetically more distant than sublexical errors in shorter words. On the other hand, the notion of "precision" does not apply to slot models. A phoneme is either encoded correctly or is not encoded at all and is pulled randomly. Therefore, if the phonological buffer is slot-based, precision should not show a systematic relationship with load.

Precision was measured as the ALINE distance (Kondrak, 2000). The ALINE algorithm takes two strings of phonemes—a *target* (e.g., "toothbrush", /tuθbɹʌʃ/) and a *response* (e.g., /puftʌs/)—and finds an alignment that minimizes the total cost of the edit operations (i.e., the sequence of deletions, insertions, and substitutions) needed to get from the target to the response. The ALINE *distance* between each pair of aligned phonemes is then calculated, which is the sum of the differences in their features, weighted by the perceptual salience of those features. Figure 2 shows the alignment selected for /tuθbɹʌʃ/ → /puftʌs/ and the ALINE distances between the aligned phonemes. As can be seen, the distance score is lower when the target and error are closer in the phonetic space (/θ/ and /f/) and larger when they are farther apart (e.g., /b/ and /t/). The prediction is then straightforward: if phonological working memory in production is resourced-based, we would expect an increase in ALINE distances as word length increases. If, on the other hand, phonological working memory is slot-based, there should be no systematic relationship between length and ALINE distances. To test these predictions, we targeted individuals with aphasia who have a phonological working memory deficit.

Target:	t	u	θ	b	ɹ	ʌ	ʃ
Response:	p	u	f	t		ʌ	s
Distance:	6	0	2	16		0	4

Figure 2. An example of an alignment of a target and a response selected by the ALINE algorithm, with the ALINE distance between each pair of aligned phonemes shown below. In this example, the response /puftʌs/ was produced for the target /tuϑbɹʌʃ/ ("toothbrush").

The general similarity in the pattern of errors between individuals with aphasia and neurotypical speakers argues for similar underlying process in the generation of such errors (e.g., Dell et al., 1997). This, in turn, implies that aphasic errors can be used to test hypotheses that are generalizable to cognitive principles in neurotypical systems. Moreover, the much greater quantity of errors produced by individuals with aphasia gives us the statistical power to test cognitive hypotheses that are difficult to test in neurotypical speakers due to the sparsity of errors in the speech of such individuals. A subset of individuals with aphasia show sublexical errors that reflect a specific impairment of phonological working memory. Individuals with phonological or graphemic buffer deficit typically show a length effect (more sublexical errors on longer words) and a decreased capacity of phonological working memory measured independently (Buchwald & Rapp, 2009; Caramazza et al., 1986, 1987b; Gathercole et al., 1994;

Rapp & Dufor, 2011). This group is particularly well-suited for testing the predictions of slot versus resource models. Therefore, we tested four such individuals on a large picture naming task (Nozari, 2019), collecting 1000+ sublexical errors. We then computed ALINE distances for these errors and investigated the relationship between word length and such distances.

Observing a correlation between length and ALINE distance would suggest that phonological working memory is underpinned by a continuous resource. But such results could also be the result of articulatory simplification. Articulatory simplification refers to changing certain sounds to others in a manner that reduces articulatory complexity and effort, e.g., increasing the degree of constriction. Galluzzi et al. (2015) demonstrated that among a group of people with phonological deficits, those who were additionally apraxic showed significantly more simplifications than those without apraxia. This is in line with the broader literature showing that individuals with apraxia generally make errors on marked phonemes, resulting in a subsequently unmarked production (Cera & Ortiz, 2010; Klich et al., 1979; Marquardt et al., 1979). It is important to note that uncovering the origin of simplifications is not always straightforward. For example, Buchwald and Miozzo (2011, 2012) described two speakers diagnosed with non-fluent aphasia and apraxia of speech (participant D.L.E. and participant H.F.L.). Both individuals tended to simplify word-initial consonant clusters (e.g., /sp/ to /p/). In Buchwald and Miozzo (2011), D.L.E. showed a robust pattern of accommodation such that a voiceless stop became aspirated after the initial fricative was deleted, abiding by the phonotactic rules of English. However, H.F.L. did not demonstrate accommodation, and the new onset remained unaspirated. In Buchwald and Miozzo (2012), the authors examined the simplification of consonant clusters made up of /s/-nasal onsets and nasal singletons (e.g., ‘small’ and ‘mall’). Prototypically, nasal singletons are produced with a longer duration compared to nasals in a cluster. D.L.E produced simplified clusters with a longer duration, similar to how the nasal would be produced in a nasal singleton. H.F.L. produced the shorted duration expected from a consonant cluster. Overall, D.L.E. produced a pattern of errors that would be expected if the errors were generated at the phonological level, while H.F.L. produced a pattern of errors that would be expected if the errors were generated during motor planning.

These results show that simplifications that appear similar on the surface do not always have a common origin. To add to this, there is still debate on what exactly qualifies as articulatory simplification (see Romani et al., 2017; Berent, 2017; Buchwald, 2017; Goldrick, 2017; Pouplier et al., 2017; Ziegler, 2017). Nevertheless, there is agreement that at least some simplifications reflect processing at a level below phonemes and may thus be less relevant to the theories tested in this paper. We extracted a number of simplifications with greater consensus in the literature and conducted a series of control analyses to test whether any length effects observed in the main analyses could alternatively be accounted for by articulatory simplification. Specifically, we used changes in markedness as simplification criteria (Galluzzi et al., 2015; Romani et al., 2017; Buchwald, 2017; Berent, 2017; cf., Ziegler, 2017). We used three primary dimensions along which consonants become less marked to determine simplifications. Table 1 lists these dimensions. An error was considered a simplification along a particular dimension if the error changed a more marked articulation to a less marked articulation. For example, a change from /s/ to /t/ would be a simplification as the manner of articulation changed from fricative to stop.

It is possible that frequency or severity of articulatory simplifications could increase as word length increases, making it unclear whether changes in precision as a function of length should be attributed to working memory or articulatory challenges. However, if articulatory complexity is the primary driver of sublexical errors, it should be the case that increased word length should result in errors with *simplified* articulations, rather than errors that make articulation more complex or neutral. If this directional change exists and increases as a function of word length, then it would be likely that decreased precision in longer words is the result of articulatory challenges, rather than inadequate resources in working memory. Conversely, if simplification is not impacted by word length, then changes in precision as a function of word length as likely the result of the increased memory load.

Table 1. Three primary dimensions and their relative markedness used to assess simplifications.

Articulatory dimension	*More Marked*	*Less Marked*
Manner	Affricate, Fricative	Stop
Place	Velar	Coronal (e.g., alveolar), Labial
Voice	Voiced	Voiceless

2. Methods

2.1 Participants

Four native English-speaking individuals with post-stroke chronic aphasia (1 female, 3 males; ages 32 – 64 years) participated in the study. Three were recruited from the Snyder Center for Aphasia Life Enhancement (SCALE; https://www.leagueforpeople.org/scale), while one (P2) was recruited from the Western Pennsylvania Patient Registry (WPPR). The basic attributes of the participants are described in Table 2. All participants were consented under protocols approved by the Institutional Review Board (IRB) for Johns Hopkins University or Carnegie Mellon University and received monetary compensation for their participation.

Table 2. Participants' demographic information.

Participant	*Age*	*Gender*	*Premorbid Handedness*	*Education*	*Years Post-Stroke*
P1	59	Male	Right	high school	7
P2	47	Male	Right	unknown	16
P3	32	Female	Right	associates	3
P4	64	Male	Right	high school	2

2.1.1 Background Tests and Inclusion Criteria

Participants were selected to meet the following inclusion criteria: (a) good semantic-lexical comprehension so they would be able to follow task instructions and have a strong knowledge of labels for concepts, (b) impaired phonological processing in production, so they would produce the data suitable for the planned analyses, and (c) evidence of a phonological working memory impairment, so they would be sensitive to the manipulation of working memory load.

Semantic-lexical abilities were measured with two auditory-word-to-picture matching tasks, one easy (175 pictures from the Philadelphia Naming Test [PNT, Roach et al., 1996] presented as target with unrelated foils) and one more difficult (30 pictures presented as targets along with semantically and phonologically related foils), previously normed and used (e.g., Nozari, 2019). Accuracy was above 98% for each participant in the easy task (Except P2 for whom this score was unavailable) In the difficult task, scores were above 93%, except for P2, who scored 77%. These scores show good semantic-lexical abilities for P1, P3, and P4. P2 shows some deficit in this respect, which we will unpack more when discussing the results (Table 3).

Phonological processing in production was measured by the number of lexical or nonlexical errors that were phonologically related to the target in picture naming using the PNT and in auditory word repetition using the repetition version of the PNT, the Philadelphia Repetition Test (PRT). The coding scheme used for these tests can be found in Appendix 1. Participants' overall production abilities ranged from 39-71% accurate on the PNT and slightly higher, 47-89% on the PRT. Importantly, in each of the four participants, at least 25% and 60% of the commission errors in naming and repetition, respectively, were phonologically related to the target (see Table 3), marking some level of impairment in phonological processing. The full breakdown of PNT errors by error type can be found in Appendix 2.

Table 3. Results of the background tests used as inclusion criteria, including the Philadelphia Naming Test (PNT) and the Philadelphia Repetition Test (PRT). See text for the description of tests. The coding scheme for the PNT and PRT can be found in Appendix 1.

	Semantic-lexical comprehension		PNT		PRT		Phonological working memory
Participant	*Accuracy on the easy task*	*Accuracy on the difficult task*	*Overall accuracy*	*Phonological errors (% out of commission errors)*	*Overall accuracy*	*Phonological errors (% out of commission errors)*	*Rhyme probe scores*
P1	100%	100%	41%	82 (81%)	47%	87 (95%)	2.58
P2	-	77% (20% semantic, *3% phonological)*	39%	29 (31%)	50%	85 (98%)	1.58
P3	100%	93% *(7% semantic)*	71%	11 (24%)	89%	13 (77%)	0.50
P4	98%	90% (3% semantic, *7% phonological)*	49%	35 (47%)	65%	39 (64%)	1.41

Finally, phonological working memory was assessed using a modified version of the rhyme probe task (Freedman & Martin, 2001). In this task, participants hear a list of words followed by a test probe. They indicate whether or not the test probe rhymed with any of the words in the proceeding list. The list length gradually increases from 1 to 6 items (levels 1-6). Each level has multiple trials and participants must pass at least 75% of the trials within a level to move on to the next level. The final score indexes the last level that the participant successfully passed, with the decimal indicating the percentage of items in the next level in which they could not reach the criterion of 75% to pass. For example, a score of 2.25 indicates that the participant answered 75% or more of the trials with a list of 2 correctly, but their answer was correct on only 25% of trials with a list of 3 words. A control group of 10 participants' performance showed a range of 4.6–6, with a mean of 5.19. Participants' performance on the rhyme probe task ranged between 0.5 and 2.58, indicating that the highest number of items that any of the participants could reliably hold in working memory was less than 3. Table 3 presents the full results of the background tests.

2.2 Material and Procedures

A large-scale picture naming task was used (Nozari, 2019). The word set consisted of 444 items (Appendix 3), comprised of color photographs obtained from online repositories. The entire set was administered twice, in two different pseudorandom orders, for a total of 888 trials. To reduce the semantic blocking effect (Schnur, 2014), items from the same semantic category were at least 12 items apart. Participants were given 20 seconds to name each picture and no feedback was given, except if the participant misidentified the target word (e.g., the participant responds "hand" to a picture of a finger, and the experimenter prompts: "what part?"). Testing was broken into as many sessions as was needed to complete the task. Sessions were recorded and saved for offline transcription.

2.3 Segmental Error Coding and ALINE Distance Calculation

All targets, along with the first lexical response on each trial, was transcribed into IPA from the recorded audio. If the participant committed a segmental error on an identifiable alternative target, the target was considered to be the lexical error. For example, if in response to the picture of an orange, the participant responded /æbl/, the target was considered to be the lexical item "apple". The data were transcribed and coded by two independent coders. Inter-rater reliability was high for consonants (k = .88), but much lower for vowels (k < .60), in part due to regional variations in vowel pronunciations and coders' different perception of the produced vowel and its mapping to the target vowel. For this reason, we only used consonants in the current study. Transcriptions and codings for the consonants were checked and discrepancies were reconciled by a third coder. If an alternative target could not be recognized by both raters, that item was removed (9.2% of all items, 12% of P1, 8.1% of P2, 10.2% of P3, 7.5% of P4). All coders were native speakers of American English.

To measure precision of the representations stored in phonological working memory, we used the ALINE distance between the aligned phonemes. Although only consonants were included in the statistical analyses, vowels were retained during alignment to ensure the best possible fit between the target and the response. The phonemes in each pair of target and response transcriptions were aligned using PyAline (Huff, 2010), an implementation of the ALINE algorithm (Kondrak, 2000) in Python. The ALINE algorithm consists of two suboperations: one to calculate the distance between the target and the response phonemes and one that uses the distance measure to generate the optimal alignment. The constants used in these calculations can be found in Appendix 4. The difference score is determined by the sum of the differences between the feature vectors of the target and response phonemes. These vectors consist of multivalued features (for *place*, *manner*, *height*, and *back*) and binary features (for *nasal*, *retroflex*, *syllabic*, *long*, *lateral*, *aspirated*, *round*, and *voice*.) For instance, the place feature is coded as 1.0 if the phoneme is bilabial and .1 if glottal, with intermediate values assigned to phonemes with places of articulation in the middle of the mouth. Each feature difference is weighted by the salience of that feature[1]. For example, the *place* feature is more salient than the *aspiration* feature and is thus considered to a greater degree when differences are summed. Different sets of features are used for vowels (*syllabic*, *nasal*, *retroflex*, *high*, *back*, *round*, and *long*) and consonants (*syllabic*, *manner*, *voice*, *nasal*, *retroflex*, *lateral*, *aspirated*, and *place*.) If one of the phonemes is a vowel and the other a consonant, the consonant features are used. The equation for calculating phoneme distance is as follows:

$$distance_{target,response} = \sum_{f \in features} |(target_f - response_f)| * \ salience_f$$

The distance metric is then used to calculate the optimal alignment of the target and response phoneme sequences. Traversing incrementally across both sequences, all possible phoneme substitutions, expansions, or skips are considered by subtracting the distance measure between the phonemes from a base value for a potential substitution, expansion, or skip (Appendix 4), resulting in a matrix of possible alignments in which higher values indicate better alignments. The optimal alignment is then selected by recursively searching the matrix, following the path of minimal alignment 'cost'. The precise algorithms used to determine the alignment can be found in the OSF repository and are detailed in Kondrak (2000). The final ALINE distance metric is computed based on the optimally aligned phonemes of the target and response, using the above distance equation.

The resulting distance metric can be thought of as a measure of phonetic precision, where a larger distance corresponds to a less precise representation of the target. If no error was made (i.e., the target and response phonemes are the same), the distance is 0. If the response included a phoneme omission, a distance could not be computed since there was no target phoneme to align with the response phoneme. Figure 2 illustrates an example of an aligned target and response word with the corresponding distances between aligned phoneme pairs, including an omitted response phoneme.

[1] The question of how to quantitatively measure perceptual salience is still contested. Here, we adopt the values from Kondrak (2000), which weight manner and place as the most salient features, while other features (like aspiration) as weighted less heavily.

2.3.2 Articulatory Simplification Coding

The markedness rankings used in our analysis were developed using criteria identified in Galluzzi et al. (2015) and Romani et al. (2017). Since markedness is more clearly defined for consonants than for vowels, only consonants will be considered. There are three primary dimensions along which consonants become less marked: voicing (consonants tend to become voiceless, excluding nasals which are not contrasted in English), place (velar consonants tend to be coronal, e.g., alveolar), and manner (fricative and affricates tend to become stops). See Table 1.

For place, if the place of articulation for the target was velar but the response phoneme was either coronal or labial, the error was coded as 1 (a simplification). If the target was coronal or labial and the response was velar, the error was coded as -1 (a complication). If there was no change in place or if the error involved a swap between coronal and labial places of articulation the error was coded as 0 (neither a simplification nor a complication). A similar coding scheme was used for manner where the error was coded as 1 if the target was a fricative or an affricate and the response was a stop, -1 if the target was a stop and the response was a fricative or an affricate, and 0 if there was no change in manner simplicity. For voicing, a response was coded a 1 if the target was voiced and the response was voiceless, -1 if the target was voiceless and the response was voiced, and 0 if there was no change in voicing.

2.4 Statistical analysis

Unless stated otherwise, analyses were conducted using linear mixed effects models with the lmerTest (v3.1.3, Kuznetsova et al., 2017) package in R (v3.5.1, R Core Team, 2018). In the main analyses, ALINE distance and word length were the dependent and independent variable, respectively. The log of the ALINE distance for each phoneme in a word was used so that the distance distribution approximates normality. Word length was determined as the number of phonemes in a word and was centered and scaled. To ensure that the effect of length of ALINE distance was not altered by the phoneme's position in the word, the analyses were repeated with a control variable, the relative position of a phoneme to the word boundaries, henceforth referred to simply as "position". The motivation for including this variable comes from a large body of literature on positional effects in tasks that require serial recall from memory. In such tasks, items at the beginning and the end of a list are typically remembered better than items in the middle positions, creating a U-shaped distribution (e.g., Atkinson & Shiffrin, 1971). A similar effect is observed in orthographic tasks, where accuracy in usually highest for the first and last letters than the middle letters in a word (Buchwald & Rapp, 2006; Caramazza et al., 1987a), mimicking the same U-shaped distribution and showing the clear application of memory principles to segmental encoding. Position was calculated as the distance to the nearest boundary (i.e., number of phonemes to the start or end of the word, whichever is closest) divided by the total length of the word. For example, in /tɹʌmpət/, the /p/ phoneme is two positions away from the end of the word (versus four position from the start) and the word length is seven, so its relative to a boundary is .29. Onset and coda /t/ each have a distance of 0 to the nearest edge, and so on and so forth, capturing the U-shaped distribution referred to above. Finally, the random effect structure was tailored to the model: in the aggregate analyses, both the random intercept of items and subjects were included. In the individual analysis, only the random intercept of items was included.

The effect of length on place and manner simplification was examined using the same linear mixed effects models, except with the dummy-coded place, manner, and voicing changes as the dependent variable. Only errors were included. Error that involved approximants (P1: 71, P2: 28, P3: 2, P4: 34) and glottals (P1: 6, P2: 23, P3: 0, P4: 5) were removed prior to analysis because there is not broad consensus in the literature on their markedness relative to other consonants. The 'clmm' function in 'ordinal' package (v2019.12.10, Christenson, 2019) was used in order to perform linear mixed effects models on ordinal data. The models were fit to the manner and place simplifications separately, and p values were corrected for multiple comparisons using the Bonferroni method.

The data, analysis code and full tables of results can be found here:
https://osf.io/5dwzb/?view_only=b9f61e67cf984513b980ef04b2695c35.

3. Results

A total of 1798 phonemes with errors were obtained for analysis. Table 4 displays the distribution of words and errors that were used to calculate ALINE distances.

Table 4. The number of words and consonant errors elicited from each participant used to calculate ALINE distances.

Participant	Words	Words with a Consonant Error	Consonants	Consonants with an Error
P1	765	410	2101	634
P2	1156	302	3053	374
P3	797	39	2384	47
P4	881	149	2470	170

3.1 Effect of Word Length on ALINE Distance

First, we examined the effect of word length on ALINE distance to determine if increasing the number of phonemes in working memory resulted in greater distance between error phonemes and target phonemes. Figure 3a illustrates the average ALINE distance calculated over all errors for the four participants, which grows monotonically as a function of word length.

The first-pass model with only word length as the independent variable showed a significant increase in ALINE distance as a function of increasing word length (β = 0.183, t = 13.23, p < .001). The second-pass model added the control variable position and its interaction with length. The effect of length remained significant in this model (β = 0.183, t = 13.01, p < .001). There was not a significant effect of position (β = -0.009, t = -0.84, p = 0.401). However, there was an interaction between length and position (β = 0.042, t = 4.08, p < .001). This interaction shows that the impact of length on ALINE score increases the farther the phoneme is from the edges. To ensure that any effect of length is not due to the differences in the number of correct phonemes in shorter vs. longer words, we repeated this analysis with only errors included. The length effect remained significant in both the first pass (β = 0.137, t = 4.103, p < .001) and the second pass (β = 0.123, t = 3.418, p < .001) models.

To confirm this pattern was not an artifact of aggregating the data, the same analysis was carried out on each individual (Fig. 3b). In the first-pass model, the effect of word length on ALINE distance was significant in three and marginal in the fourth participant (Table 5), showing a common trend in all participants. The inclusion of the position variable and its interaction with length in the second-pass model did not change this pattern. Appendix 5 presents the full results of these analyses.

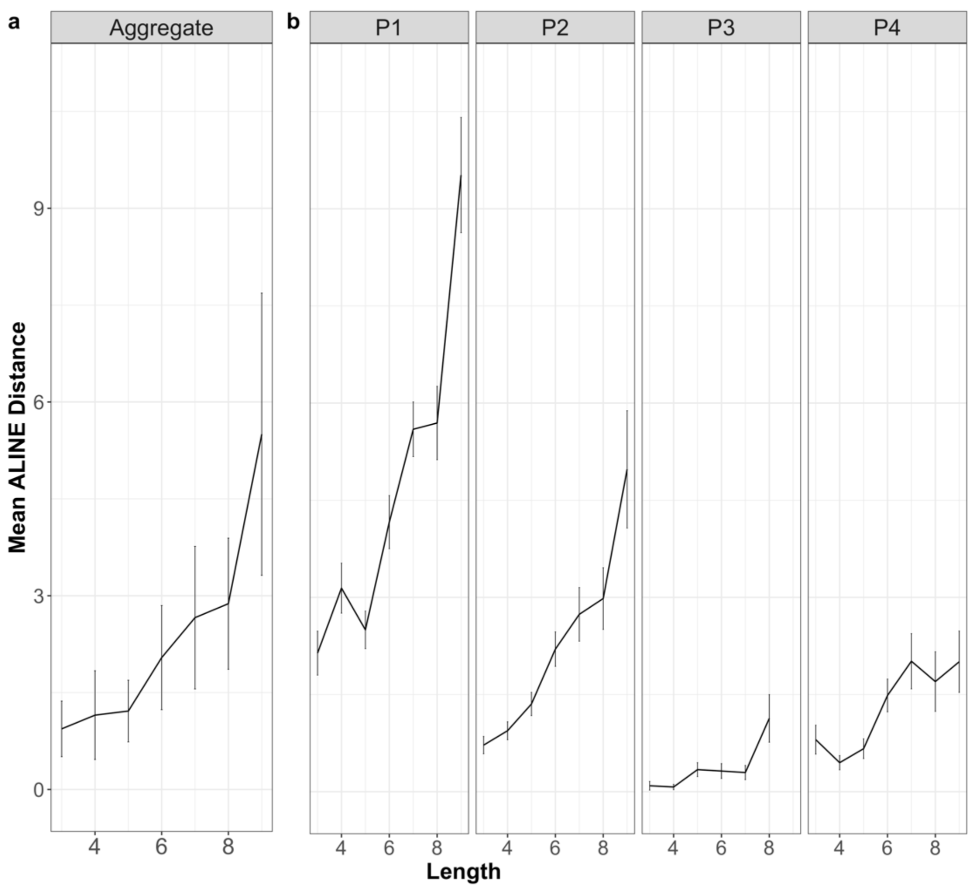

Figure 3. ALINE distance calculated over all errors as a function of word length in aggregate data (a) and individual participants (b). Error bars represent standard error. A minimum of 10 errors was required to include a given length for a participant.

Table 5. The effect of length on ALINE distance for individual participants over all errors.

Participant	Coefficient	SE	t	p value

P1	0.24	0.03	9.48	< .001
P2	0.23	0.03	8.89	< .001
P3	0.08	0.03	2.35	0.019
P4	0.13	0.03	5.25	< .001

Recall that we also included lexical errors on targets in our analysis. This brings up the question, is the length effect different for the lexically substituted targets? This question is especially relevant to participants who made a large number of such lexical substitutions. Of our four patients, P2 had a lower score on semantic-lexical tasks and produced the greatest number of lexical substitutions in our task (756, compared to P1: 182, P3: 289, and P4: 325). We, therefore, conducted two additional analyses, one at the group level and one specifically on P2. These analyses included target status (target or lexical substitution), length, and the interaction between the two as its fixed effect. The random effect structure was similar to the models reported earlier. Table 6 reports the results of these two analyses. The critical findings across both analyses were that the length effect was preserved and that it did not interact with target status, ensuring that the main finding was not contaminated by the target status.

There was also an unexpected finding: a significant negative relationship between the presence of a lexical error and ALINE distance was observed, indicating that errors were more precise when a lexical error was made. This was unexpected because Buchwald and Falconer (2014) compared letter errors when a lexical item was retrieved correctly compared to when a semantically related lexical item was erroneously retrieved. They found an increase in the letter errors, arising from the graphemic buffer, in the latter case. In other words, committing a lexical error further increased the chance of a sublexical error. The results of the current analysis seem to suggest the opposite. However, we believe that this may be an artifact of a selection bias. The trial was only coded as a lexical error when the coders could clearly identify an item as the participant's new target. If the response was less clear, the trial was more likely to be excluded from the analyses. We therefore do not attach a theoretical interpretation to this finding. Rather, the purpose of reporting these analyses was to demonstrate that the effect of length on ALINE distance is not an artifact of including lexical errors and is not modulated by such inclusion.

Table 6. The effect of lexical errors and length on ALINE distance for all participants (aggregate) and P2.

	Factor	*Coefficient*	*SE*	*t*	*p value*
Aggregate	Length	0.183	0.016	11.354	< .001
	Target status	-0.098	0.023	-4.288	< .001
	Length x Target status	-0.022	0.022	-0.982	.326
P2	Length	0.213	0.035	6.018	< .001
	Target status	-0.132	0.041	-3.234	.001
	Length x Target status	-0.024	0.041	-0.585	.558

To summarize, we found that as length increased, so did the ALINE distance between the target phoneme and the error phoneme. In other words, an increase in the working memory load led to less precise errors in language production. This effect was not an artifact of the different number of phonemes without errors in short and long words, the position of the errors in the words, or the lexical status of the target. But before we could take these results as supporting a resource model of the phonological buffer in language production, we must rule out two more alternatives. The first is the possibility that error segments in longer words have a different profile from those in shorter words. Certain phonemes have more close neighbors than others. For example, /s/ has many neighbors that change only one phonetic feature: /z/, /t/, /ʃ/, and /f/. In contrast, /ʧ/ has only two close neighbors, /ʤ/ and /ʃ/. Therefore, any error with a target of /ʧ/ has a larger minimum ALINE distance than one with a target of /s/. If longer words contain more of these isolated phonemes like /ʧ/, then the ALINE distance may be artificially increased as word length increases. Analysis 3.2. addresses this concern by focusing on three phonemes and tracking their changes as a function of length. The second confound is the possibility that length simply leads to articulatory simplification, which may be correlated with ALINE distance. Analysis 3.3. tests this possibility.

3.2. Length Effects as a Function of Specific Phonemes

To rule out the confound that the length effect may be driven by a different phonological composition of longer vs. shorter words, we selected three phonemes for each of which we had over 100 errors and at least one error across all possible lengths (/t/ 999 instances of which 138 were errors, /k/ 944 instances of which 134 were errors, and /s/ 894 instances of which 153 were errors) and repeated the analyses on them. Figure 4 displays the mean ALINE distance of each phoneme.

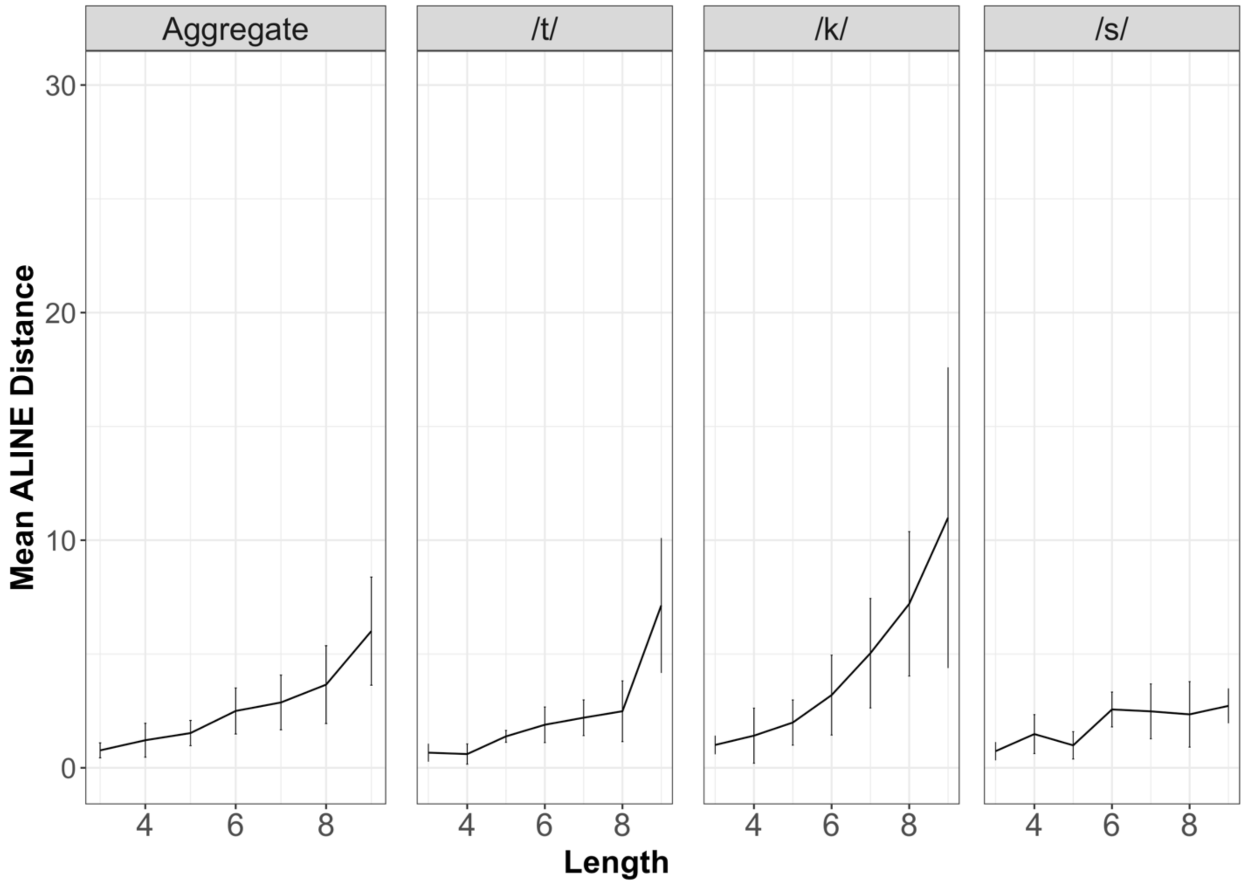

Figure 4. Mean ALINE distance for phonemes /t/, /k/, and /s/ individually and collectively (Aggregate), as a function of word length. Error bars represent standard error.

Table 7 contains the full results of these analyses. There was a significant length effect on the aggregate data (β = 0.230, *t* = 9.68, *p* < .001), which remained significant when all correct phonemes were removed (β = 0.208, *t* = 3.77, *p* < .001). Additional analyses focused on individual phonemes showed that the length effect was significant for each phoneme. This provides evidence against the possibility that the length effect is the result of the type of phonemes changing as word length changes because the same trend is observed when examining the errors of a particular phoneme as a function of length.

Table 7. The effect of length on ALINE distance for individual phonemes.

Phoneme Target	Coefficient	SE	t	p value
Aggregate	0.230	0.023	9.68	< .001
/t/	0.287	0.044	6.59	< .001
/k/	0.266	0.039	6.79	< .001
/s/	0.127	0.037	3.47	< .001

3.3 Errors as a Result of Articulatory Simplification

Figure 5 shows the mean of simplifications for the place and manner of articulation for each participant. A mean greater than zero suggests that the participant simplified the articulation more often than the reverse. One sample t-tests were used to determine significance and the full results can be seen in Table 8. After correcting for 12 comparisons ($\alpha = .004$), P1 showed significant simplification for place of articulation ($t(540) = 7.76$, $p < .001$). However, they also displayed a tendency to produce errors that made the articulation significantly *more* marked in voicing ($t(540) = -7.06$, $p < .001$) and marginally more marked in manner ($t(540) = -2.83$, $p = .005$). P2 also showed significant simplification of the manner of articulation ($t(318) = 5.07$, $p < .001$) as well as the voicing ($t(318) = 3.48$, $p < .001$). None of the other effects were close to significant.

In short, P1 and P2 both showed some evidence of simplification: P1 for place and P2 for manner and voicing. To test whether length caused these simplifications, separate linear mixed effect models with length as a fixed effect were applied to these three simplifications. After correcting for three comparisons ($\alpha = .017$), the effect of length did not reach significance for P1's place simplification ($\beta = 0.21$, $z = 1.97$, $p = .049$), or for P2's voicing simplification ($\beta = 0.21$, $z = 1.68$, $p = .094$). There was a significant effect of length of P2's manner simplification, but the direction of the effect suggested *less* simplification in longer words ($\beta = -0.617$, $z = -4.92$, $p < .001$), which is opposite of what would be expected if simplification drove the length effect. Collectively, these results suggest that, while articulatory simplification may explain some errors in individuals with aphasia, it is not critically linked to length, and can thus not be an alternative explanation for the effect of length on error precision.

Table 8. T-Tests examining simplification for each dimension of articulation, for each participant.

Dimension	Participant	Estimate	t	p value	Degrees of Freedom
Place	P1	0.144	7.763	< .001	540
	P2	0.031	1.232	.219	318
	P3	-0.022	-0.240	.811	44
	P4	0.076	1.487	.14	118
Manner	P1	-0.07	-2.835	.005	540
	P2	0.16	5.071	< .001	318
	P3	0.156	2.006	.051	44
	P4	-0.008	-0.168	.867	118
Voicing	P1	-0.168	-7.063	< .001	540
	P2	0.103	3.481	< .001	318
	P3	0.067	0.621	.538	44
	P4	-0.05	-1	.319	118

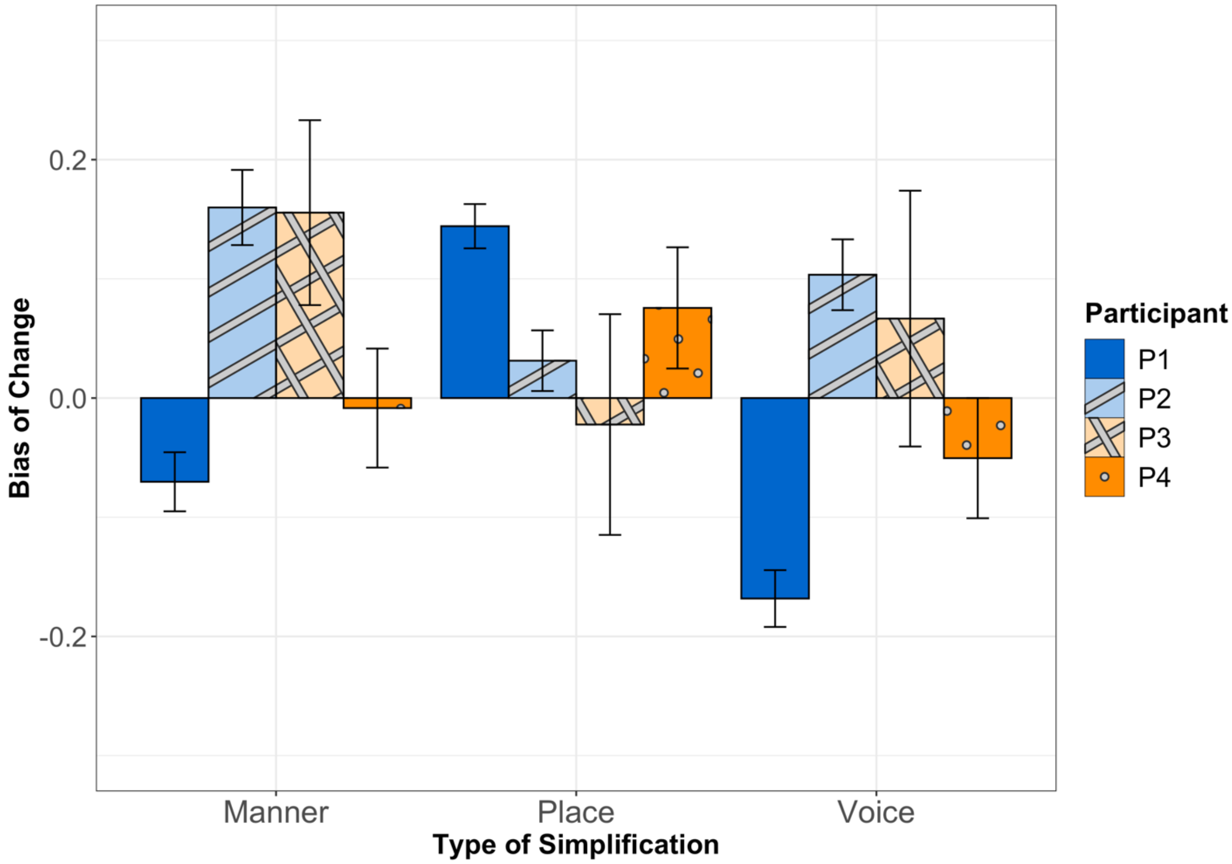

Figure 5: The bias of change toward simpler or less simple articulations for consonant errors, split by manner simplifications, place simplifications, and voice simplifications. Positive means indicate a trend toward simplification, negative means indicate a trend away from simplification. Error bars represent standard error.

4. Discussion

We utilized sublexical speech errors in conjunction with tools from computational linguistics to test the predictions of two dominant theories of working memory resources in language production. The first outcome of this study was a quantitative measure of error "precision", the ALINE distance, indexing the distance between the error phoneme and the target phoneme in the articulatory phonetic space. Using this measure, the second outcome of the study was evidence in support of the resource models of working memory. Specifically, we found that ALINE distance increased with the length of the target word, showing a decrease in error precision as a function of increased working memory, precisely as predicted by resource models of working memory. We further showed that this effect was not accounted for by position, differences in the number of correct segments or the phonological composition of long and short words, or articulatory simplification. Thus, this finding provides the first piece of evidence in support of the resource models of the phonological buffer in language production, and more generally, the first piece of evidence that the scope of such models reaches beyond perception to also include production.

4.1. The importance of extending resource models to production

Evidence for resource models of working memory have been accumulating in recent years in both visual and auditory domains (Bays et al., 2011; Hepner & Nozari, 2019; Joseph et al., 2015; Oberauer & Lin, 2017; van den Berg et al., 2012; Wilken & Ma, 2004). This study adds the first piece of evidence from the production side to these models. This is not a trivial extension because of the different dynamics of perception and production. Specifically, selection demands are much stronger in production because a speaker must ultimately commit to positioning their articulators to produce a certain sound or syllable. This commitment has been a key argument for considering "phonemes" as the critical sound units in spoken production, whereas the psychological reality of phonemes in perception has been seriously questioned (e.g., Samuel, 2020). In line with this claim, Roelofs (1999) found that participants were able to name a set of pictures more quickly when the initial sounds of the items were identical (that is, the same phoneme), compared to when the initial sounds were dissimilar (e.g., /d/ versus /f/). There was, however, no facilitation when the phonemes merely shared many features: /b/ and /p/. He concluded that the facilitation occurred at the level of holistic phonemes that are insensitive to subphonemic information.

In a similar vein, the primacy of phonological representations for production is demonstrated in *accommodation*, or the fact that the allophonic environment often changes to accommodate an error (Fromkin, 1971). For example, if the plural morpheme /s/ erroneously moves to a neighboring word, its voicing feature will accommodate to the voicing of the final phoneme of the new word, such that the utterance remains well-formed by the rules of English (e.g., 'track cows' [træk kæwz] for target 'cow tracks' [kæw træks]; Fromkin, 1971). This pattern suggests that these errors are generated before phonemes are translated into phonetic features. This evidence has been taken to support the non-decompositionality of phonemes as units of production 4/21/2023 9:16:00 AM, which would make error precision undefined and in turn render resource models inapplicable to language production, by definition.

Contrary to the above, recent evidence has shown that subphonemic units play an active role in production and generation of sublexical errors. While such errors may be often missed in transcriptions because they are significantly harder to perceive (Gormley, 2015; Mowrey & MacKay, 1990; Pouplier et al., 2014), acoustic recordings and measurements of articulatory movements have confirmed that gradient, sublexical errors are commonly produced and that those errors show traces of the intended target, suggesting competition between multiple candidate phoneme representations (Goldrick & Blumstein, 2006). Similar shifts in view have occurred regarding accommodation. For example, Pouplier et al. (2014) examined accommodation in simplified consonant clusters. Not only was accommodation inconsistently applied (28% of errors failed to demonstrate accommodation), but the resulting acoustics also displayed a gradient pattern. Gormley (2015) reported similar findings with accommodation of vowel length in the context of incorrect voicing of a following consonant (40.7% of errors were unaccommodated, though 51% of errors were not able to be classified). These findings revived some of the earlier demonstration of non-categorical phonological errors. For example, Mowrey and MacKay (1990) used electromyograph recordings of muscle movement to examine errors, specifically errors made on /s/ in tongue twisters like "She sells seashells…". Of the 48 errors observed, 43 displayed a continuous change across features (primarily the intrusion of the *labial* feature from /ʃ/) while only 5 displayed a categorical phoneme deletion. Thus, gradient errors were not only common but in fact greatly outnumbered the categorical errors.

To summarize, there are both theoretical reasons and empirical support for viewing non-decompositional phonemes as the critical units involved in speech production, which, if true, would make resource models of working memory inapplicable to production. At the same time, there is also evidence that subphonemic units do indeed play an important role in production. Critically, sublexical errors show sensitivity to such subphonemic features. Given this body of evidence, it is perfectly reasonable to assume that sublexical errors can arise at multiple levels. We used the evidence in favor of the role of subphonemic representations to test whether error precision could be measured systematically for sublexical errors and whether it could be used to test the predictions of resource models. Our results confirmed that precision was indeed a useful index and that its variation with working memory load was well-aligned with the predictions of resource models. This finding considerably expands the scope of resource models, beyond generalization over stimulus type, to both perception and production.

4.2. The notion of resource and the biological foundation of resource models

A key concept in resource models of working memory is "resource", but what is a resource? Although the definition seems vague, this resource has a number of defining characteristics: it is limited, divisible, and the degree of its allocation to various items determines the precision with which those items are remembered. Moreover, attention can guide the differential allocation of resources to items. For example, Hepner and Nozari (2019) showed that, in a set of four items, cued items (50% chance of being probed at recall) were remembered more precisely compared to non-cued items within the same set (17% chance of being probed at recall), which were, in turn, remembered with less precision than items in a set with no cueing (25% chance of being probed at recall). In other words, attention guided graded allocation of resources, which manifested in graded precision at the time of recall (see also Wood & Cowan, 1995; Kane et al., 2001, for a close link between attention and working memory).

This characteristic closely links the concept of resource to neural gain, which is often proposed as the neural basis for focused attention (e.g., McAdams & Maunsell, 1999; Eldar et al., 2013). Populations of neurons encode a probability distribution for a given stimulus dimension, which must be retrieved by decoding at the time of recall (Ma et al., 2006; Sanger, 1996; Zemel et al., 1998). The more accurate the result of this decoding process, the greater the precision of the recalled item. The decoding accuracy of an item is itself contingent on the gain with which the item was originally encoded. Increasing gain means increasing neural activity, which consumes nontrivial amounts of energy (Attwell & Laughlin, 2001; Lennie, 2003). It is thus the limits of this energy which limits neural gain for individual items in working memory at the time of encoding. The greater the number of items, the less neural gain, in terms of energy, that can be allocated to each of them, hence the noisier the decoding process and the less precise the recalled item, as predicted by resource models (Bays, 2014, 2015; Ma et al., 2006; Orhan & Ma, 2015; Schneegans & Bays, 2017; van den Berg & Ma, 2018).

Defining resources in a biologically plausible way is more difficult for slot models. One proposal is that the limit in the number of slots comes from locking the pattern of activity for each item to distinct phases of neural oscillations at the service of feature binding (e.g., Lisman & Idiart, 1995). But given the wide range of possible oscillations, there is no clear reason why the number of slots should be as limited as is often assumed by slot models of working memory (Bays, 2015). Additionally, feature binding may be achieved in other ways that do not put clear limitations on the process (e.g., Matthey et al., 2015). In short, resource has a definable biological basis in resource models, but such a basis is not clearly identifiable in slot models, creating a further advantage for resource models, as a general account of working memory, over slot models.

4.3. Consideration of alternative accounts

The current results conclusively rule out the dominant version of slot models. There is, however, a variation of slot models, called the "slot plus averaging" (SA) model (Zhang & Luck, 2008), which must also be considered. SA is similar to slot models in its assumptions that working memory has a fixed number of slots for storing individual items. However, SA makes two additional assumptions. First, a single item can have multiple copies, each stored in one slot, such that multiple slots may store the same item. Second, each copy is stored with some amount of random noise. How "precisely" an item is remembered depends on how many slots store a copy of that item. If an item is stored across many slots, averaging its activation across these slots cancels out random noise and makes its memory representation more precise than when it is stored in fewer slots or a single slot.

The assumptions of the allocation of multiple slots and averaging activation over those slots are closer to the concept of resource division in resource models. Due to this similarity, the two models' predictions are sometimes indistinguishable. Therefore, we will not be able to rule out

SA on the basis of the current data. We will, however, review three lines of criticism against SA in the literature, which makes it a less likely candidate to account for the existing patterns of data. The first criticism, unpacked in the previous section, is that the concept of resource as fixed slots does not have a strong biological support. Thus, in so far as SA utilizes this concept, it is less biologically plausible than the resource models. The second criticism concerns the parameters of the SA model when fitted to the empirical data. These have been shown to lack internal consistency (Bays, 2018) and could be difficult to interpret as they do not necessarily reflect the actual probability of an item being stored in memory (Ma, 2018; see also Schurgin et al., 2018, for a related criticisim). These two characteristics limits the utility of SA. The third criticism is the nature of the evidence often cited in support of SA. In many cases, the support comes from null results (e.g., Todd & Marois, 2004; Vogel & Machizawa, 2004; Xu & Chun, 2006; Zhang & Luck, 2008) or results that are better accounted for by a variant of the resource model, such as the variable-precision resource model, in which the division of resources is subject to stochastic noise (van den Berg et al., 2014).

To summarize, while the pattern of data obtained in the current study may also be compatible with SA, the general body of evidence against the SA model leads us to prefer the interpretation offered by the resource models. Finally, two other influential accounts of capacity limits in working memory are *interference* (Nairne, 1990; Oberauer & Kliegl, 2006; Oberauer & Lin, 2017) and *decay* (Baddeley et al., 1975; Barrouillet et al., 2004; Burgess & Hitch, 1999) models. But these accounts are not direct competitors of resource models. In fact, some of the mechanisms proposed by these models, such as similarity as a source of interference, can be readily accommodated by resource models (see Oberauer et al., 2016, for a review of these models).

4.4. Links to other debates regarding working memory in language processing

Although there is consensus that phonological working memory is necessary for processing language, there is still debate on its nature. This debate reflects a larger debate in the literature regarding the nature of representations involved in working memory. One view, the *embedded processes model*, claims that working memory is essentially the activated portion of long-term memory, and therefore involves the same regions in the brain (Cowan, 1999; Cowan et al., 2021). Another view, *the buffer view*, proposes the existence of specialized temporary stores, also called buffers, for short-term maintenance of information in working memory, which are distinct from long-term representations of the same information (Baddeley et al., 2021; Baddeley & Hitch, 1974; Martin et al., 2021; Purcell et al., 2021; see Jonides et al., 2005 for a review).

Applied to phonological working memory, there is now converging evidence from neuroimaging, brain stimulation and lesion studies, that phonological representations are stored in superior temporal gyrus (STG; see Binder, 2015 and; Nozari, 2021 for reviews). In keeping with the embedded processes model, Acheson et al. (2011) found activation in a similar region for phonological working memory. In contrast, lesion studies have identified a different region, the left supramarginal gyrus (SMG), to be specialized for phonological working memory (Paulesu et al., 2017; Purcell et al., 2021). Using representational similarity analysis, Yue and Martin (2021) successfully decoded phonological representations during a delay period in the left SMG but not the STG. In a more recent study, the same authors reported that decoding was possible in STG, as long as no distractors were present. However, the presence of distractors during the delay period interfered with decoding in left STG. Critically, decoding was still possible in SMG. In a second experiment, the authors showed that transcranial magnetic stimulation to SMG, but not STG, disrupted phonological working memory (Yue & Martin, 2022). These results provide support for the buffer account.

The debate between embedded processes and buffer accounts is parallel to the central debate of the current study regarding the division of resources. Both the embedded processes and buffer accounts acknowledge capacity limitations and require a mechanism to explain how resources are divided. Similarly, both slot and resource models can be situated within these two accounts. Specifically for our current purposes, resource models fit well within both accounts, as they are largely agnostic with regard to whether WM operates over representations in LTM or is a mechanism that has some independence from LTM.

A related debate concerns the notion of domain-generality of working memory. Proponents of the buffer account often maintain domain-specific views, as buffers are specialized for processing distinct kinds of information, such as phonological vs. orthographic (Purcell et al., 2021). Additional evidence for the domain-specificity of working memory comes from the neuropsychological evidence. Individuals with brain damage can be selectively impaired in semantic or phonological working memory (Martin et al., 1994, see 2020, for a review, see also Martin et al., 2021 for complementary results from lesion-symptom mapping analyses). Collectively, this body of evidence points to clear neural and functional dissociation in working memory for various functions, even within the language system. At the same time, there is a notion of domain-generality in terms of the applicability of similar computational principles across domains. The findings of the current study, together with the larger body of evidence in support of resource models, demonstrate that the same general computations are applied across both perception and production to representations from distinct domains, such as visual and auditory processing. Together, these findings support domain-generality at the level of computational principles, and domain-specificity at the level of neural and functional implementations (Freund & Nozari, 2018; Nozari & Novick, 2017).

6. Acknowledgments

We thank Christopher Hepner and Jessa Westheimer for their help with data collection, coding, and calculating ALINE scores. We also thank the SCALE in Baltimore, patients and their families. Finally, we thank Adam Buchwald and one anonymous reviewer for their helpful comments on an earlier draft of this manuscript.

6. References

Acheson, D. J., Hamidi, M., Binder, J. R., & Postle, B. R. (2011). A Common Neural Substrate for Language Production and Verbal Working Memory. *Journal of Cognitive Neuroscience*, *23*(6), 1358–1367. https://doi.org/10.1162/jocn.2010.21519

Atkinson, R. C., & Shiffrin, R. M. (1971). The Control of Short-Term Memory. *Scientific American*, *225*(2), 82–91.

Attwell, D., & Laughlin, S. B. (2001). An Energy Budget for Signaling in the Grey Matter of the Brain. *Journal of Cerebral Blood Flow & Metabolism*, *21*(10), 1133–1145. https://doi.org/10.1097/00004647-200110000-00001

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 575–589. https://doi.org/10.1016/S0022-5371(75)80045-4

Baddeley, A., & Hitch, G. (1974). *Working memory.* (G. H. Bower, Ed.; Vol. 8, pp. 47–89). Academic Press.

Baddeley, A., Hitch, G., & Allen, R. (2021). A multicomponent model of working memory. In *Working memory: State of the science* (pp. 10–43). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0002

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults’ Working Memory Spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Bays, P. M. (2014). Noise in Neural Populations Accounts for Errors in Working Memory. *The Journal of Neuroscience*, *34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P. M. (2015). Spikes not slots: Noise in neural populations limits working memory. *Trends in Cognitive Sciences*, *19*(8), 431–438. https://doi.org/10.1016/j.tics.2015.06.004

Bays, P. M. (2018). Failure of self-consistency in the discrete resource model of visual working memory. *Cognitive Psychology*, *105*, 1–8. https://doi.org/10.1016/j.cogpsych.2018.05.002

Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision*, *11*(10), 6–6. https://doi.org/10.1167/11.10.6

Berent, I. (2017). Is markedness a confused concept? *Cognitive Neuropsychology*, *34*(7–8), 493–499. https://doi.org/10.1080/02643294.2017.1422485

Binder, J. R. (2015). The Wernicke area. *Neurology*, *85*(24), 2170–2175. https://doi.org/10.1212/WNL.0000000000002219

Bisiacchi, P. S., Cipolotti, L., & Denes, G. (1989). Impairment in Processing Meaningless Verbal Material in Several Modalities: The Relationship between Short-term Memory and Phonological Skills. *The Quarterly Journal of Experimental Psychology Section A*, *41*(2), 293–319. https://doi.org/10.1080/14640748908402367

Brooks, P. J., & Macwhinney, B. (2000). Phonological priming in children’s picture naming. *Journal of Child Language*, *27*(2), 335–366. https://doi.org/10.1017/S0305000900004141

Bub, D., Black, S., Howell, J., & Kertesz, A. (1987). Speech output processes and reading. In M. Coltheart, G. Satori, & R. Job (Eds.), *The cognitive neuropsychology of language* (pp. 79–110). Lawrence Erlbaum.

Buchwald, A. (2017). Complexity in articulatory and segmental levels of production. *Cognitive Neuropsychology*, *34*(7–8), 488–492. https://doi.org/10.1080/02643294.2017.1422484

Buchwald, A., & Falconer, C. (2014). Cascading activation from lexical processing to letter-level processing in written word production. *Cognitive Neuropsychology*, *31*(7–8), 606–621. https://doi.org/10.1080/02643294.2014.951609

Buchwald, A., & Miozzo, M. (2011). Finding Levels of Abstraction in Speech Production: Evidence From Sound-Production Impairment. *Psychological Science*, *22*(9), 1113–1119. https://doi.org/10.1177/0956797611417723

Buchwald, A., & Miozzo, M. (2012). Phonological and Motor Errors in Individuals With Acquired Sound Production Impairment. *Journal of Speech, Language, and Hearing Research*, *55*(5). https://doi.org/10.1044/1092-4388(2012/11-0200)

Buchwald, A., & Rapp, B. (2006). Consonants and vowels in orthographic representations. *Cognitive Neuropsychology*, 308–337.

Buchwald, A., & Rapp, B. (2009). Distinctions between orthographic long-term memory and working memory. *Cognitive Neuropsychology*, *26*(8), 724–751. https://doi.org/10.1080/02643291003707332

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*, 551–581. https://doi.org/10.1037/0033-295X.106.3.551

Caramazza, A., Miceli, G., & Villa, G. (1986). The role of the (output) phonological buffer in reading, writing, and repetition. *Cognitive Neuropsychology*, *3*(1), 37–76. https://doi.org/10.1080/02643298608252669

Caramazza, A., Miceli, G., Villa, G., & Romani, C. (1987a). The role of the Graphemic Buffer in spelling: Evidence from a case of acquired dysgraphia. *Cognition*, *26*(1), 59–85. https://doi.org/10.1016/0010-0277(87)90014-x

Caramazza, A., Miceli, G., Villa, G., & Romani, C. (1987b). The role of the Graphemic Buffer in spelling: Evidence from a case of acquired dysgraphia. *Cognition*, *26*(1), 59–85. https://doi.org/10.1016/0010-0277(87)90014-X

Cera, M. L., & Ortiz, K. Z. (2010). Phonological analysis of substitution errors of patients with apraxia of speech. *Dementia & Neuropsychologia*, *4*(1), 58–62. https://doi.org/10.1590/S1980-57642010DN40100010

Cowan, N. (1999). An Embedded-Processes Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory* (1st ed., pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114-185. https://doi.org/10.1017/s0140525x01003922

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2021). An embedded-processes approach to working memory: How is it distinct from other approaches, and to what ends? In *Working memory: State of the science* (pp. 44–84). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0003

Cowan, N., Rouder, J. N., Blume, C. L., & Saults, J. S. (2012). Models of verbal working memory capacity: What does it take to make them work? *Psychological Review*, *119*(3), 480–499. https://doi.org/10.1037/a0027791

Dell, G. S. (1984). Representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 222–233. https://doi.org/10.1037/0278-7393.10.2.222

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Eldar, E., Cohen, J. D., & Niv, Y. (2013). The effects of neural gain on attention and learning. *Nature Neuroscience*, *16*(8), 1146–1153. https://doi.org/10.1038/nn.3428

Freedman, M. L., & Martin, R. C. (2001). Dissociable components of short-term memory and their relation to long-term learning. *Cognitive Neuropsychology*, *18*(3), 193–226. https://doi.org/10.1080/02643290126002

Freund, M., & Nozari, N. (2018). Is adaptive control in language production mediated by learning? *Cognition*, *176*, 107–130. https://doi.org/10.1016/j.cognition.2018.03.009

Fromkin, V. A. (1971). The Non-Anomalous Nature of Anomalous Utterances. *Language*, *47*(1), 27–52. https://doi.org/10.2307/412187

Galluzzi, C., Bureca, I., Guariglia, C., & Romani, C. (2015). Phonological simplifications, apraxia of speech and the interaction between phonological and phonetic processing. *Neuropsychologia*, *71*, 64–83. https://doi.org/10.1016/j.neuropsychologia.2015.03.007

Garrett, M. F. (1975). The Analysis of Sentence Production11The research reported here was supported in part by NIMH grant HD 05/68-02,03,04. Some of the work was done in collaboration with S. R. Shattuck and her contribution is gratefully acknowledged. This paper was written while the author was a Senior Research Fellow of the Australian-American Educational Foundation (Fulbright-Hayes program) at Monash University, Melbourne. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 9, pp. 133–177). Academic Press. https://doi.org/10.1016/S0079-7421(08)60270-4

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*(2), 103–127. https://doi.org/10.1080/09658219408258940

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243. https://doi.org/10.1016/s1364-6613(00)01662-4

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory*, *6*, 225–255. https://doi.org/10.1080/741942359

Goldrick, M. (2017). Encoding of distributional regularities independent of markedness: Evidence from unimpaired speakers. *Cognitive Neuropsychology*, *34*(7–8), 476–481. https://doi.org/10.1080/02643294.2017.1421149

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*(6), 649–683. https://doi.org/10.1080/01690960500181332

Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition*, *102*(2), 219–260. https://doi.org/10.1016/j.cognition.2005.12.010

Gormley, A. (2015). Acoustic Evidence for Phonologically Mismatched Speech Errors. *Journal of Psycholinguistic Research*, *44*(2), 105–117. https://doi.org/10.1007/s10936-013-9282-2

Hepner, C. R., & Nozari, N. (2019). Resource allocation in phonological working memory: Same or different principles from vision? *Journal of Memory and Language*, *106*, 172–188. https://doi.org/10.1016/j.jml.2019.03.003

Huff, P. A. (2010). *PyAline: Automatically Growing Language Family Trees Using the ALINE Distance*. 53.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, *10*(11), 480–486. https://doi.org/10.1016/j.tics.2006.09.002

Jonides, J., Lacey, S. C., & Nee, D. E. (2005). Processes of Working Memory in Mind and Brain. *Current Directions in Psychological Science*, *14*, 2–5. https://doi.org/10.1111/j.0963-7214.2005.00323.x

Joseph, S., Iverson, P., Manohar, S., Fox, Z., Scott, S. K., & Husain, M. (2015). Precision of working memory for speech sounds. *Quarterly Journal of Experimental Psychology*, *68*(10), 2022–2040. https://doi.org/10.1080/17470218.2014.1002799

Kane, M., Bleckley, M., Conway, A., & Engle, R. (2001). A controlled-attention view of working memory. *Journal of Experimental Psychology. General*, *130*, 169–183. https://doi.org/10.1037//0096-3445.130.2.169

Klich, R. J., Ireland, J. V., & Weidner, W. E. (1979). Articulatory and Phonological Aspects of Consonant Substitutions in Apraxia of Speech. *Cortex*, *15*(3), 451–470. https://doi.org/10.1016/S0010-9452(79)80070-2

Kondrak, G. (2000). A new algorithm for the alignment of phonetic sequences. *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 288–295. http://dl.acm.org/citation.cfm?id=974305.974343

Kurowski, K., & Blumstein, S. E. (2016). Phonetic basis of phonemic paraphasias in aphasia: Evidence for cascading activation. *Cortex*, *75*, 193–203. https://doi.org/10.1016/j.cortex.2015.12.005

Lennie, P. (2003). The Cost of Cortical Computation. *Current Biology*, *13*(6), 493–497. https://doi.org/10.1016/S0960-9822(03)00135-0

Levelt, W. J. M. (1989). *Speaking: From intention to articulation* (pp. xiv, 566). The MIT Press.

Lisman, J. E., & Idiart, M. A. P. (1995). Storage of 7 ± 2 Short-Term Memories in Oscillatory Subcycles. *Science*, *267*(5203), 1512–1515. https://doi.org/10.1126/science.7878473

Ma, W. J. (2018). Problematic usage of the Zhang and Luck mixture model. *BioRxiv*, 268961.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), Article 11. https://doi.org/10.1038/nn1790

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia*, *8*(3), 323–350. https://doi.org/10.1016/0028-3932(70)90078-3

Marquardt, T. P., Reinhart, J. B., & Peterson, H. A. (1979). Markedness analysis of phonemic substitution errors in apraxia of speech. *Journal of Communication Disorders*, *12*(6), 481–494. https://doi.org/10.1016/0021-9924(79)90011-X

Martin, R. C., Ding, J., Hamilton, A. C., & Schnur, T. T. (2021). Working Memory Capacities Neurally Dissociate: Evidence from Acute Stroke. *Cerebral Cortex Communications*, *2*(2), tgab005. https://doi.org/10.1093/texcom/tgab005

Martin, R. C., Rapp, B., & Purcell, J. (2020). Domain-Specific Working Memory: Perspectives from Cognitive Neuropsychology. *Working Memory: State of the Science*, 235–287. https://doi.org/10.1093/oso/9780198842286.003.0009

Martin, R. C., Shelton, J. R., & Yaffee, L. S. (1994). Language processing and working memory: Neuropsychological evidence for separate phonological and semantic capacities. *Journal of Memory and Language*, *33*, 83–111. https://doi.org/10.1006/jmla.1994.1005

Matthey, L., Bays, P. M., & Dayan, P. (2015). A Probabilistic Palimpsest Model of Visual Short-term Memory. *PLOS Computational Biology*, *11*(1), e1004003. https://doi.org/10.1371/journal.pcbi.1004003

McAdams, C. J., & Maunsell, J. H. R. (1999). Effects of Attention on Orientation-Tuning Functions of Single Neurons in Macaque Cortical Area V4. *The Journal of Neuroscience*, *19*(1), 431–441. https://doi.org/10.1523/JNEUROSCI.19-01-00431.1999

McCarthy, R., & Warrington, E. K. (1984). A Two-Route Model of Speech Production: Evidence from Aphasia. *Brain*, *107*(2), 463–485. https://doi.org/10.1093/brain/107.2.463

Miller, G. (1956). The Magical Number Seven, plus or minus 2: Some Limits on our Capacity for Processing Information. *Psychological Review*, *63*(2), 81–97.

Mowrey, R. A., & MacKay, I. R. A. (1990). Phonological primitives: Electromyographic speech error evidence. *The Journal of the Acoustical Society of America*, *88*(3), 1299–1312. https://doi.org/10.1121/1.399706

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269. https://doi.org/10.3758/BF03213879

Nozari, N. (2019). The dual origin of semantic errors in access deficit: Activation vs. inhibition deficit. *Cognitive Neuropsychology*, *36*(1–2), 31–53. https://doi.org/10.1080/02643294.2019.1587397

Nozari, N. (2021). The neural basis of word production. In A. Papafragou, J. C. Trueswell, & L. R. Gleitman (Eds.), *The Oxford Handbook of the Mental Lexicon* (pp. 535–558). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198845003.013.12

Nozari, N., & Dell, G. S. (2009). More on lexical bias: How efficient can a “lexical editor” be? *Journal of Memory and Language*, *60*(2), 291–307. https://doi.org/10.1016/j.jml.2008.09.006

Nozari, N., & Dell, G. S. (2013). How damaged brains repeat words: A computational approach. *Brain and Language*, *126*(3), 327–337. https://doi.org/10.1016/j.bandl.2013.07.005

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language*, *63*(4), 541–559. https://doi.org/10.1016/j.jml.2010.08.001

Nozari, N., & Novick, J. (2017). Monitoring and Control in Language Production. *Current Directions in Psychological Science*, *26*(5), 403–410. https://doi.org/10.1177/0963721417702419

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

Oberauer, K., & Kliegl, R. (2006). A formal model of capacity limits in working memory. *Journal of Memory and Language*, *55*(4), 601–626. https://doi.org/10.1016/j.jml.2006.08.009

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Orhan, A. E., & Ma, W. J. (2015). Neural Population Coding of Multiple Stimuli. *Journal of Neuroscience*, *35*(9), 3825–3841. https://doi.org/10.1523/JNEUROSCI.4097-14.2015

Paulesu, E., Shallice, T., Danelli, L., Sberna, M., Frackowiak, R. S. J., & Frith, C. D. (2017). Anatomical Modularity of Verbal Working Memory? Functional Anatomical Evidence from a Famous Patient with Short-Term Memory Deficits. *Frontiers in Human Neuroscience*, *11*. https://www.frontiersin.org/articles/10.3389/fnhum.2017.00231

Pinet, S., & Nozari, N. (2018). “Twisting fingers”: The case for interactivity in typed language production. *Psychonomic Bulletin & Review*, *25*(4), 1449–1457. https://doi.org/10.3758/s13423-018-1452-7

Pouplier, M., Marin, S., & Kochetov, A. (2017). The difficulty of articulatory complexity. *Cognitive Neuropsychology*, *34*(7–8), 472–475. https://doi.org/10.1080/02643294.2017.1419947

Pouplier, M., Marin, S., & Waltl, S. (2014). Voice Onset Time in Consonant Cluster Errors: Can Phonetic Accommodation Differentiate Cognitive From Motor Errors? *Journal of Speech, Language, and Hearing Research*, *57*(5), 1577–1588. https://doi.org/10.1044/2014_JSLHR-S-12-0412

Purcell, J., Rapp, B., & Martin, R. C. (2021). Distinct Neural Substrates Support Phonological and Orthographic Working Memory: Implications for Theories of Working Memory. *Frontiers in Neurology*, *12*, 681141. https://doi.org/10.3389/fneur.2021.681141

Rapp, B., & Dufor, O. (2011). *The Neurotopography of Written Word Production: An fMRI Investigation of the Distribution of Sensitivity to Length and Frequency*.

Roach, A., Schwartz, M., Martin, N., Grewal, R., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clin. Aphasiol.*, *24*.

Roelofs, A. (1999). Phonological segments and features as planning units in speech production. *Language and Cognitive Processes*, *14*(2), 173–200. https://doi.org/10.1080/016909699386338

Romani, C., Galuzzi, C., Guariglia, C., & Goslin, J. (2017). Comparing phoneme frequency, age of acquisition, and loss in aphasia: Implications for phonological universals. *Cognitive Neuropsychology*, *34*(7–8), 449–471. https://doi.org/10.1080/02643294.2017.1369942

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070. https://doi.org/10.1016/j.jml.2019.104070

Sanger, T. D. (1996). Probability density estimation for the interpretation of neural population codes. *Journal of Neurophysiology*, *76*(4), 2790–2793. https://doi.org/10.1152/jn.1996.76.4.2790

Schneegans, S., & Bays, P. M. (2017). Restoration of fMRI Decodability Does Not Imply Latent Working Memory States. *Journal of Cognitive Neuroscience*, *29*(12), 1977–1994. https://doi.org/10.1162/jocn_a_01180

Schnur, T. T. (2014). The persistence of cumulative semantic interference during naming. *Journal of Memory and Language*, *75*, 27–44. https://doi.org/10.1016/j.jml.2014.04.006

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2018). Psychophysical Scaling Reveals a Unified Theory of Visual Memory Strength. *BioRxiv*, 325472.

Shallice, T., Rumiati, R. I., & Zadini, A. (2000). The Selective Impairment of the Phonological Output Buffer. *Cognitive Neuropsychology*, *17*(6), 517–546. https://doi.org/10.1080/02643290050110638

Shattuck-Hufnagel, S., & Klatt, D. H. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, *18*(1), 41–55. https://doi.org/10.1016/S0022-5371(79)90554-1

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature*, *428*(6984), Article 6984. https://doi.org/10.1038/nature02466 van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial Comparison of Working Memory Models. *Psychological Review*, *121*(1), 124–149. https://doi.org/10.1037/a0035234

van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, *7*, e34963. https://doi.org/10.7554/eLife.34963

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), Article 6984. https://doi.org/10.1038/nature02447

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 11. https://doi.org/10.1167/4.12.11

Wood, N. L., & Cowan, N. (1995). The Cocktail Party Phenomenon Revisited: Attention and Memory in the Classic Selective Listening Procedure of Cherry (1953). *Journal of Experimental Psychology: General*, *124*(3), 243.

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), Article 7080. https://doi.org/10.1038/nature04262

Yue, Q., & Martin, R. C. (2021). Maintaining verbal short-term memory representations in non-perceptual parietal regions. *Cortex*, *138*, 72–89. https://doi.org/10.1016/j.cortex.2021.01.020

Yue, Q., & Martin, R. C. (2022). Phonological working memory representations in the left inferior parietal lobe in the face of distraction and neural stimulation. *Frontiers in Human Neuroscience*, *16*. https://doi.org/10.3389/fnhum.2022.890483

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic Interpretation of Population Codes. *Neural Computation*, *10*(2), 403–430.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), Article 7192. https://doi.org/10.1038/nature06860

Ziegler, W. (2017). Complexity of articulation planning in apraxia of speech: The limits of phoneme-based approaches. *Cognitive Neuropsychology*, *34*(7–8), 482–487. https://doi.org/10.1080/02643294.2017.1421148

Appendix 1

As described in Nozari (2019), Philadelphia Naming Test (PNT) and Philadelphia Repetition Test (PRT) errors were coded as belonging to the following types:

- *Semantic: response bears a taxonomic or thematic relation to the target, e.g., cat → dog.*
- *Phonological: target shares at least one phoneme in the same syllabic position as in the target word, or two or more phonemes in any position. These errors can be words or non-words, eg., cup → cap, cup → cip.*
- *Mixed: response bears both a semantic and a phonological relation to the target, e.g., plumber → plunger.*
- *Compound: An error that contains only some of the correct morphemes., e.g., fireman → fire.*
- *Unrelated: Word errors that do not meet the criteria for phonologically or semantically similar words.*
- *No Response: includes omissions and utterances that did not give a clear response to the picture.*

Appendix 2

Percentage of Philadelphia Naming Test (PNT) errors for all participants, split by error type.

	Semantic	Phonological	Mixed	Compound	Unrelated	No Response
P1	6%	80%	1%	1%	10%	3%
P2	41%	27%	1%	1%	12%	18%
P3	36%	22%	6%	4%	10%	22%
P4	13%	39%	4%	6%	7%	30%

Appendix 3

Full set of 444 words used in the picture naming task.

hand	brush	stem	helicopter	donkey	bongos	garage
cat	candy	bathtub	steak	sled	pan	remote
tree	microphone	bull	nose	radio	jellyfish	can opener
door	mussel	waffle	blowfish	avocado	meatloaf	tongue
wrench	broom	T-shirt	shirt	urinal	blimp	wasp
grapes	cauliflower	flask	razor	lion	pliers	train
fly	finger	otter	funnel	bagpipe	lobster	cherries
pen	spatula	mop	fork	strawberry	pinky	glue
cellphone	hyena	drum	chalk	nightstand	monitor	artichoke
sundae	spoon	blueberries	fireplace	frog	bed	grasshopper

tie	mailbox	lamp	hamburger	fern	rhino	bottle
goat	keyboard	goose	cactus	eraser	lime	robe
croissant	glove	record	raccoon	boot	hairdryer	organ
curtains	binder	sock	raft	shark	omelette	tiger
violin	dolphin	tomato	accordion	foot	snake	bedroom
plum	plunger	stork	shorts	skateboard	slipper	bagel
bear	measuring cup	truck	refrigerator	bacon	xylophone	colander
thermos	donut	cheek	kangaroo	grapefruit	stove	octopus
ear	sheep	chocolate	ceiling	toucan	chopsticks	shoulder
brownie	axe	worm	sandpaper	ipad	moth	apple
duck	bush	stool	faucet	oreo	measuring tape	tape
perfume	blackberry	grass	dog	highlighter	paper	banjo
protractor	sweater	grater	cookie	saw	recliner	whale
gate	owl	pancake	cantaloupe	slug	seahorse	olive
harmonica	radish	scorpion	boat	hat	thumb	jeans
airplane	window	date	bee	shelf	eggplant	burrito
scale	knife	napkin	ankle	raisins	skis	tongs
plate	chameleon	screw	record player	turtle	butterfly	cockroach
iPod	wagon	calendar	microwave	cup	trumpet	canoe
spider	flashlight	fox	harp	pizza	orange	highchair
peeler	pretzel	elbow	jacket	tweezers	sushi	deodorant
fig	bra	suit	fish	chicken	desk	bat
stairs	ant	asparagus	poptart	laptop	lizard	bamboo
bib	router	cardinal	motorcycle	knee	nailclipper	trombone
toilet paper	peach	roof	nail	thumbtack	paperclip	cupcake
rooster	grill	banana	kitchen	monkey	toes	flamingo
clarinet	ostrich	unicycle	taco	garlic	wolf	scarf
screwdriver	cucumber	popcorn	eel	rolling pin	potato	broccoli
heel	qtip	marshmallow	ginger	straw	apricot	crib
crackers	bathroom	neck	tambourine	caterpillar	gong	blender
branch	camel	toothbrush	watermelon	underwear	porcupine	chest
ottoman	hoe	seal	dove	rake	dustpan	centipede
nutcracker	charger	cd	ladle	fence	bowl	pear
van	wrist	thermometer	belt	toothpaste	pencil	jar
raincoat	mouse	guitar	popsicle	swan	spaghetti	jukebox
walrus	clipboard	rabbit	calf	notebook	deer	alligator
smore	pants	jet	shin	toaster	cake	kayak
car	pig	controller	shovel	crowbar	pelican	vest
stingray	submarine	stapler	pantry	ladybug	oven	

camera	comb	gorilla	bus	bread	couch	
flute	egg	chair	hippo	raspberry	roots	
teapot	panda	sleeve	lasagna	crayon	pigeon	
rice	dresser	mushroom	corn	mattress	toilet	
flower	pitchfork	salad	scissors	penguin	pot	
apron	soup	snail	zebra	towel	hotdog	
hammer	skunk	saxophone	dress	suspenders	turkey	
starfish	mango	floss	shampoo	onion	acorn	
table	arm	envelope	closet	mosquito	drawer	
marker	eagle	squirrel	cow	tv	lips	
lemon	hashbrowns	coconut	speaker	jeep	dragonfly	
cotton candy	level	tick	whisk	moose	parrot	
chimney	hummingbird	deck	shoe	celery	shoelace	
sandal	sandwich	soap	giraffe	ladder	compass	
peacock	ruler	jack	pineapple	leaf	cabinets	
back	bicycle	elephant	crab	corkscrew	piano	
dishwasher	pie	muffin	carrot	attic	shower	

Appendix 4

Constant values for ALINE distance calculations. Binary values (0, 1) are assigned for nasal, retroflex, syllabic, long, lateral, aspirated, round, and voice features. Operation costs are applied during the alignment processed and are not incorporated into the distances.

Place		***Manner***		***Height***		***Back***	
Bilabial	1	Stop	1	High	1	Front	1
Labiodental	0.95	Affricate	0.9	Near-high	0.83	Near-front	0.75
Dental	0.9	Fricative	0.8	High-mid	0.67	Central	0.5
Alveolar	0.85	Approximant	0.6	Mid	0.5	Near-back	0.25
Retroflex	0.8			Low-mid	0.33	Back	0
Palato-alveolar	0.75			Near-low	0.17		
Palatal	0.7			Low	0		
Velar	0.6						
Uvular	0.5						
Pharyngeal	0.3						
Glottal	0.1						

Salience			
Syllabic	5	High	5
Place	40	Back	5
Voice	10	Manner	50
Nasal	10	Retroflex	10
Lateral	10	Long	1
Aspirated	5	Round	5

Alignment Costs	
Cost_Skip	5
Cost_Substitution	45
Cost_Expansion	45
Cost_Vowel	5

Appendix 5

Full results from second-pass analyses examining the effect of length on ALINE distance. Position and the interaction between position and length were included as a control variable.

	Factor	*Coefficient*	*SE*	*t*	*p value*
Aggregate	*Intercept*	9.07E-03	1.71E-01	0.053	0.961
	Length	1.81E-01	1.42E-02	12.757	< 2e-16
	Position	-1.51E-02	1.03E-02	-1.466	0.143
	Length x Position	4.73E-02	1.05E-02	4.506	0.00000667
P1	*Intercept*	-2.85E-02	2.65E-02	-1.075	0.283
	Length	2.52E-01	2.65E-02	9.528	< 2.00E-16
	Position	-6.46E-03	2.21E-02	-0.293	0.77
	Length x Position	1.08E-01	2.33E-02	4.613	0.00000421
P2	*Intercept*	-7.22E-03	2.63E-02	-0.275	0.7838
	Length	2.11E-01	2.64E-02	7.996	1.05E-14
	Position	-3.63E-02	1.92E-02	-1.889	0.059
	Length x Position	4.79E-02	1.96E-02	2.442	0.0146
P3	*Intercept*	3.93E-02	3.42E-02	1.149	0.2511
	Length	8.19E-02	3.45E-02	2.373	0.0181
	Position	6.25E-03	2.06E-02	0.304	0.761
	Length x Position	1.40E-02	2.10E-02	0.668	0.5043
P4	*Intercept*	3.11E-02	2.60E-02	1.196	0.232
	Length	1.36E-01	2.61E-02	5.212	2.92E-07
	Position	-7.99E-03	2.15E-02	-0.372	0.71
	Length x Position	-2.07E-03	2.12E-02	-0.097	0.922