CORTEX 185 (2025) 184–210

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Single Case Report

Disassembling an experimental artifact in aphasia: Why phonemic errors with words and semantic errors with numbers?

Ismael Gutiérrez-Cordero [a,b,c,*] *and Javier García-Orza* [b,c,d]

[a] *Cognitive Neurology and Aphasia Unit, Centro de Investigaciones Médico-Sanitarias (CIMES), Universidad de Málaga, Spain*
[b] *Numerical Cognition Lab, Universidad de Málaga, Spain*
[c] *Department of Basic Psychology, Universidad de Málaga, Spain*
[d] *Instituto de Investigación Biomédica de Málaga (IBIMA), Spain*

ARTICLE INFO

Article history:
Received 29 August 2024
Revised 12 February 2025
Accepted 18 February 2025
Action editor Holly Robson
Published online 27 February 2025

Keywords:
Word production
Numbers
Conduction aphasia
Phonemic errors
Semantic errors

ABSTRACT

There is broad consensus as to the significance of speech errors in aphasia. The analysis of errors is understood to provide clear clues for clinical diagnosis, the identification of those cognitive-linguistic processes affected, and the corresponding impaired cerebral structures. However, Stimulus Type Effect on Phonological and Semantic errors (STEPS), a phenomenon in which a person with aphasia produces more phonological errors with words (e.g., "tamle" for "table") but more semantic errors with number words (e.g., "thirteen" for "forty-two"), casts doubt on this consensus view. In this paper two studies are described, in which we explore whether STEPS is in fact a result of the lack of rigorous control over the materials compared (words versus numbers) and the evaluation conditions. Two persons, one with a reproduction conduction aphasia and the other with a repetition conduction aphasia, participated in the studies. Study 1 explored the role of memory load in the emergence of STEPS by eliciting the repetition of pairs of semantically-unrelated words. In Studies 2a and 2b, our participants were asked to produce sequences of high- and low-frequency words from one semantic category (colors), and this was compared to the performance in multi-digit number production tasks. The results showed that sequences of high-frequency colors, like multi-digit numbers, were produced mainly with semantic errors, whereas sequences of low-frequency colors showed a mixed pattern with many phonemic and semantic errors. It seems that the production of semantic errors and the absence of phonemic errors in multi-digit numbers that give rise to STEPS is an experimental artifact caused by the combination of several factors: the use of semantically-related high-frequency words, produced cyclically under high-memory-demand conditions. These findings contribute substantially to the current discussion of

* Corresponding author. Department of Basic Psychology, School of Psychology and Speech Therapy, Universidad de Málaga, Ampliación de Teatinos, c/ Doctor Ortiz Ramos, 12, 29010, Málaga, Spain.
E-mail address: igtezcordero@uma.es (I. Gutiérrez-Cordero).
https://doi.org/10.1016/j.cortex.2025.02.005

language production models and allow for a deeper understanding of the neurocognitive processes that underly speech errors in aphasia.

1. Introduction

Speech errors in aphasia, also called paraphasias, have been used widely as a classification tool for clinical diagnosis (e.g., Goodglass & Wingfield, 1997). From a neurocognitive approach, these paraphasias have also been used as markers of the processes within the linguistic system that are affected (e.g., Caramazza & Hillis, 1990; Hillis, 2001; Ramoo et al., 2021) as well as of the impaired brain areas (e.g., Berthier et al., 2018; Fridriksson et al., 2009; Hickok & Poeppel, 2007; McKinnon et al., 2018; Mirman et al., 2015; Stark et al., 2019). From this lesion-process-symptom mapping approach, it is largely accepted that lexical errors (e.g., semantic paraphasias: "story" instead of "tale") have their origin in semantic and/or lexical processes with damage located to the left anterior temporal lobe and other ventral areas, whereas non-lexical errors (e.g., phonemic errors: "lale" instead of "tale") arise from difficulties in post-lexical processes such as phonological encoding or the storing of phoneme sequences following damage to left inferior parietal lobe and other dorsal areas (e.g., Cloutman et al., 2009; Hillis, 2001; McKinnon et al., 2018; Mirman et al., 2015; Ramoo et al., 2021; Schwartz & Dell, 2016, pp. 701–715; Stark et al., 2019).

However, a variety of data challenges this consensus (e.g., Gold & Kerstez, 2001; Martin et al., 1996). Here we will focus on a phenomenon that Dotan and Friedmann (2015) have called the *Stimulus Type Effect on Phonological and Semantic* errors (STEPS). This frequent effect, first noted by Geschwind (1965), and subsequently by many others (e.g., Cohen et al., 1997; Dotan & Friedmann, 2015; García-Orza et al., 2020; Messina et al., 2009; Ochtrup et al., 2013) refers to the presence, in persons with aphasia (PWA), of predominantly phonemic errors in tasks that involve the production of words (e.g., "lale" for "tale") and pseudowords (e.g., "pesal" for "pepal"), and predominantly semantic errors in producing number words (e.g., "thirteen" for "forty-two"). One part of STEPS – the commission of phonemic errors in word production – is predicted under the lesion-process-symptom mapping approach, in that these patients usually have impairments in left inferior parietal areas and surrounding areas that affect phonological encoding causing phonemic errors; but the commission of semantic errors and the absence of phonemic errors when producing number words cannot thus be explained. Although some interactive models may explain the coexistence of these errors, the specificity of the errors, semantic for numbers, phonemic for all other words, can hardly be explained and so, the consensus that has emerged over a number of years in research and clinical practice regarding the significance of errors in aphasia can in fact be challenged.

Additionally, STEPS has also received attention in recent years in that it has been considered evidence of different production processes for numbers and words (e.g., Bencini et al., 2011; Cohen et al., 1997; Dotan & Friedmann, 2015, 2019; Mesina et al., 2009). As indicated by Dotan and Friedmann, "certain types of aphasia cause phoneme substitutions and omissions in words but not in numbers … This dissociation indicates that words and number words are handled by different sub-processes within the verbal production system" (Dotan & Friedmann, 2019, p. 181).

The aim of the present research is to demonstrate that STEPS is an experimental artifact caused by subtle differences in task demands, and thus, that: (a) a particular PWA may show both phonemic and semantic errors depending on testing conditions; (b) there is no need, on the evidence of STEPS, to claim the existence of different processes for words and numbers, at least at the phonological encoding level; and (c) the lesion-process-symptom mapping consensus remains adequate, although it should be approached in a more flexible way.

In what follows, we will first briefly describe the experimental evidence supporting STEPS, and the account suggested by Cohen et al. (1997), extended by Dotan and Friedmann (2015), to explain these errors. We will then present an alternative position, inspired by previous work (see García-Orza et al., 2020; Ochtrup et al., 2013), that claims that STEPS is caused by insufficient control of the experimental conditions when comparing word to multi-digit production. Empirical evidence will then be reported from the study of two patients who showed STEPS in a previous study (García-Orza et al., 2020); when these PWA were evaluated with non-numerical words in similar circumstances to those used when evaluating numbers, they exhibited predominantly semantic errors depending on the lexical frequency of the words. This supports the claim that STEPS is an experimental artifact. Finally, we provide an account of the STEPS pattern within interactive models of language processing and discuss the implication of our results.

1.1. *Semantic and phonemic errors depending on the type of word*

Several patients have been described in the literature who (with non-numerical words) mainly showed errors in the selection of phonemes while with numerical words they engaged in errors consisting of the substitution of a given number word for another (i.e., syntactic errors: 306 as "thirty six", and semantic [lexical] errors: 306 as "three hundred and four") (e.g., Cohen et al., 1997; Dotan & Friedmann, 2015; García-Orza et al., 2020; Geschwind, 1965; Messina et al., 2009; Ochtrup et al., 2013). Cohen et al. (1997) also found a prevalence of semantic (substitution) errors in letter naming in some patients, and this was subsequently confirmed by Dotan

and Friedmann (2015), who also extended the phenomenon to function words and derivational morphemes.

As noted above, these errors are difficult to explain using classical models of language production because phonemic errors, since they originate in the phonological assembly stage, should also affect the production of all types of words, including numbers, function words, and derivational morphemes. On the other hand, when we focus on number production, semantic errors are unexpected, in that no semantic deficits are typically observed in these patients (i.e., they understand numbers and are able to decide which is the larger number in numerical comparison tasks). Based on Cohen et al. (1997), Dotan and Friedmann proposed the *Building Blocks Hypothesis* (BBH) to account for STEPS. They assume that STEPS originates in an impairment in the phonological output buffer (POB), that is, the deficit associated with conduction aphasia of the reproduction variety (e.g., Shallice et al., 2000; Shallice & Warrington, 1977). It is usually said that the POB is the stage at which phonemes are assembled to create the phonological form of words, and is the place where these phonological representations are subsequently stored until articulatory plans are ready to be executed (e.g., Hillis, 2001). Thus, impairments in selecting or storing phonemes would cause deletions, additions and substitutions of phonemes both in word and nonword production, leading to phonemic errors. The BBH posits that numbers, letters, function words and morphemes would work as phonemes, that is, as basic units for building more complex phonological sequences (i.e., respectively, multi-digits, acronyms, and words). These are represented as pre-assembled, atomic, units in dedicated stores in the POB, and when this system is damaged, the result is phonemic errors in words, but complete substitution of these atomic units in producing numbers, letters, or morphemes; that is, semantic and/or syntactic errors arise with numbers, substitutions with letters, and morphemic errors with morphemes (Dotan & Friedmann, 2015).

Despite the fact that the BBH can explain STEPS, García-Orza et al. (2020) recently cast doubt on this by testing one of the main assumptions of the model: the idea that STEPS is caused by impairment in the POB. To this end they explored two patients, one with an impairment in the POB, and one with an impairment in the phonological input buffer (PIB). According to the BBH, STEPS should have been observed in the POB patient but not in the PIB patient (although, see Fischer-Baum et al., 2018). However, the PIB patient in García-Orza et al. (2020) also clearly showed STEPS in repetition tasks. The authors described the methodological problems regarding the procedures followed to collect the data, these being common to previous studies, and provided an alternative explanation in which STEPS was interpreted as an experimental artifact (see also Ochtrup et al., 2013). In the following section we detail and extend these concerns, but for the sake of clarity, we will restrict our arguments to those studies with non-number words and number words, setting aside those studies with letters, morphological affixes and function words, despite the fact that these studies may suffer from the same experimental issues (see General Discussion, where we will also comment on these latter stimulus types).

1.2. *Methodological concerns in observing STEPS*

Research on language processing has noted the significance of variables like familiarity, lexical frequency, length, etc. in understanding word production and in properly assessing patients with impairments in this skill (e.g., Nickels, 1997). To do this it is also fundamental to consider the concrete nature of the tasks used and the way stimuli are presented. For example, asking a participant to repeat a single word is not the same as asking them to repeat a sequence of words (e.g., Gold & Kerstez, 2001); similarly, asking them to name sequences of pictures from the same semantic category is not the same as using pictures from different categories (Damian et al., 2001; Harvey et al., 2019; Howard et al., 2006). Without a precise control of factors like these, it is difficult to explain word production and its breakdown in aphasia.

As pointed out by García-Orza et al. (2020), a detailed analysis of the studies reporting STEPS shows subtle differences in the way words and numbers are tested. These differences, taken individually, can probably not explain the presence of more phonemic errors with words and more semantic errors with numbers, but it is possible that some of these differences can explain parts of STEPS (e.g., the absence of phonemic errors with numbers), and a combination of these differences may account for the whole effect. Although not all these studies reported their methodological details, it is clear that, when assessing STEPS, words and numbers were not in general treated similarly in terms of several factors which we will group into two dimensions: the nature of the stimuli and the context of the evaluation.

1.2.1. *Differences in the nature of the stimuli*

Most of the cases described as evidence for STEPS do not include errors in the production of single digits; that is, errors in these stimuli are occasional. In fact, semantic errors with numbers arise more clearly when patients are asked to produce multi-digit numbers (e.g., "twenty-one", "three hundred sixty-one") (e.g., see Table 1 in Dotan & Friedmann, 2015; see also García-Orza, et al., 2020; Geschwind, 1965; Ochtrup et al., 2013). Hence, to arrive at the claim that STEPS is in operation, an unbalanced comparison is made: single words are compared to sequences of number words. Interestingly, comparing words to multi-digit words involves significant differences in three aspects that are well-known for affecting word production: length, morphological structure, and frequency.

Regarding length, in most of the studies showing STEPS it seems that multi-digit words were compared to words without controlling for this factor. For instance, Dotan and Friedmann (2015) (see García-Orza et al., 2020 for a similar problem in a study in Spanish) asked their Hebrew-speaking patients to read, name and repeat single words with 2–12 phonemes (1–4 syllables) and compared this with the naming of multi-digits with 1–5 digits, and the repetition of multi-digits with 1–3 digits.[1] Although the list of stimuli used in the experiment is not available, a single

[1] We will use the term "naming of multi-digits" when participants are asked to transcode from Arabic numbers to an oral response, and "reading of multi-digits" when participants are asked to transcode from orthographic number words to an oral response.

example is enough to deduce that differences in length between words and multi-digit words were not controlled for. A three-digit number like 347 has 4 words, 9 syllables and 20 phonemes in Hebrew, this being far more than the longest word they employed (4 syllables, 12 phonemes). The same number-word ("three-hundreds and forty-seven") has 5 words, 8 syllables and 22 phonemes in English, and 4 words 8 syllables and 25 phonemes in García-Orza et al.'s (2020) study using Spanish (*trescientos cuarenta y siete*). As a consequence of the recursive structure of the numerical system, it seems that it is difficult in most languages to match these multi-digit words with any single word in terms of length, because most multi-digit numbers involve several words, forming a sequence, whereas most common names involve a single word. To our knowledge the only study that has explicitly compared words to numbers matched for phoneme length (and frequency) is Bachoud-Lévi and Dupoux (2003). They used a small number of items (22) although they did not include details of the stimuli; however, from the examples provided it can be inferred that most of the digits they used were single-word numbers like "forty". Interestingly, they did find phonemic errors in words but no types of errors in the production of numbers. This pattern, that has been called a case of partial STEPS (see also Bencini et al., 2011), suggests that the amount of information the PWA must produce is relevant to the presence of semantic errors. In most of the studies showing a (complete) STEPS, PWA were asked to produce multi-digit number words that are longer and hence more demanding in terms of memory resources than single words. This is interesting, because interactive models of language processing have pointed out that the phonological load imposed by the production of sequences of words in patients that suffer from a decay of phonological traces may give rise to changes in the nature of errors (Dell, 1986; Dell & O'Seaghdha, 1992; Dell et al., 1997; Gold & Kertesz, 2001; Martin et al., 1996; Martin et al., 1994). In line with this, Martin and Saffran (1992) postulated the *Continuum Hypothesis*, by which phonemic, formal, and semantic errors would lie on a continuum of severity determined by the phonological load of the task. Applying this hypothesis to multi-digits, it is possible that the memory load that involves producing sequences of number words may turn the phonemic errors into semantic ones.

In terms of frequency, by consulting lexical frequency dictionaries we can see that number words are very high-frequency stimuli. Evidence on the role of frequency at the lexical as well as at the POB level has been given in various studies (Kittredge et al., 2008; Shallice et al., 2000). For instance, Shallice et al. (2000) found that low-frequency words were more prone to phonemic errors than high-frequency words in a POB patient. Despite such evidence, the explicit control of this variable in studies of STEPS is not described in many published reports. Again, some studies did control for this factor but failed to find a complete STEPS effect, in that no types of errors were found in single-number words (e.g., Bachoud-Lévi & Dupoux, 2003; but see also Cohen et al., 1997). More importantly, calculating the frequency of multi-digit numbers involves three different problems. First, most databases do not include Arabic numbers in their counts, and given that in some languages multi-digit numbers above one hundred are written using Arabic numbers it is difficult to establish their actual frequency. Second, the frequency of number words presented in multi-digit words can be calculated in two ways, by considering either the whole sequence or the frequency of each lexeme, taking into account the compositionality of multi-digit words. For instance, consider the case of a multi-digit word like "three-hundred". If the logarithm of the frequency of the whole sequence is considered, it will be rather small (in Spanish, in which it is written as a single word [*trescientos*], it is .65 according to the EsPal database; Duchon et al., 2013); however, if the multi-digit word is considered as a compound word, and thus involves the selection of the lexemes "three" and "hundred", then the frequency of the number words that make up this multi-digit would be far higher (in Spanish, the logarithm of the frequency for the word *tres* ["three"] is 2.82, and for word *cientos* ["hundreds"] is 1.49). Research in word recognition has shown that lexical decision times for compound words are affected by the frequency of the constituent lexemes and the whole word frequency (see, e.g., Armando et al., 2023; Duñabeitia et al., 2007). Third, using lexical frequency instead of lemma frequency for multidigit numbers is probably ill-advised due to their compositional nature. For instance, in the EsPal database (Duchon et al., 2013) the lexical frequency per million of the Spanish word *tres* is 667.27 whereas the frequency of the lemma per million (that is, the number of times that a word appears as a single word but also as part of another number like "twenty-three") is almost seven times higher, 4241.7. In conclusion, it is not straightforward to control for the frequency of multi-digit numbers and it seems that, in STEPS research, even when frequency has been taken into account, multi-digit numbers have been compared to words with lower frequency.

Importantly, an additional consequence of this componential view of number words is also very pertinent here; multi-digit number words are formed by the combination of number words (i.e., numerical lexemes) that are usually short. Since phonemic errors increase with length in speakers with conduction aphasia (see, e.g., Caramazza et al., 1986; García-Orza & León-Carrión, 2005; Shallice et al., 2000), it might be that the short length of the words that comprise multi-digit words also protect them from phonemic errors.[2]

In summary, when we consider the properties of multi-digit number words it is possible that the absence of phonemic errors with these stimuli in patients showing STEPS could reflect their high frequency, the short length of their elements (words), or both. Regarding the presence of semantic errors with numbers, it seems that the processing load of multi-digit numbers (usually composed of several number words) may increase semantic errors. Additional factors related with the emergence of semantic errors in the production of multi-digit numbers are discussed in the next section, in which we address differences observed in the way researchers usually

[2] An additional factor that is not addressed in the current research and may favor phonemic errors in common words compared to number words, is the difference at sublexical level (in the phonemic structure, probabilistic phonotactics and/or phonetic complexity) between these stimuli. Number words seem simpler at phonemic and phonetic complexity and higher in phonotactic probability than most common words, and these have been shown to increase the production of phonemic errors (e.g., Goldrick & Larson, 2008; Shallice et al., 2000). Although, as will be evident in our results, other factors seem to play a more relevant role in explaining the STEPS, it is clear that the role of these sublexical factors needs to be explored in future studies.

test the oral production of number and non-number words in cases showing STEPS.

1.2.2. *Differences in the evaluation context*

Performance in number production is usually assessed using semantically-homogeneous blocks, that is, all stimuli in the list belong to the category of numbers. On the contrary, non-number word production is assessed using semantically-heterogeneous lists in which words from different categories such as animals, fruits, clothes, furniture, etc. are mixed (see, e.g., Bachoud-Lévi & Dupoux, 2003; Bencini et al., 2011; Delazer & Bartha, 2001; Marangolo et al., 2004, 2005; Ochtrup et al., 2013; Rodriguez & Laganaro, 2008). As observed in continuous naming paradigms, homogeneous lists favor semantic interference in naming, a well-known phenomenon (e.g., Biegler et al., 2008; Damian et al., 2001; Harvey et al., 2019; Howard et al., 2006; Kroll & Stewart, 1994). Consequently, the usual way of presenting the stimuli (numbers in a homogeneous context, words in a semantically-heterogeneous context) may favor the presence of more semantic errors in number words.

In addition to this, the evaluation conditions in homogenous blocks also resemble a cyclic naming paradigm (e.g., Belke et al., 2005; Navarrete, Del Prato, Peressotti, & Mahon, 2014; Oppenheim et al., 2010; Schnur et al., 2006), in which participants are presented repeatedly with a finite set of pictures randomly ordered into lists that can be either semantically related (homogeneous condition) or not related (heterogeneous condition). The wildly-reported finding here is that naming times are longer in the homogeneous than in the heterogeneous condition. Although considerable debate has arisen as to the nature of this effect, the usual account for the phenomenon is that it is the result of an increase in the competition among the representations of the pictures that belong to the same semantic category (see Belke, 2017, for a review). In the STEPS literature, when faced with the production of multi-digit number words, PWA were asked to repeatedly produce a finite set of number words (e.g., “one”, “two”, “hundred”, “thousand”) that belong to the same category and hence which have overlapping representations. This may favor competition between their semantic representations, leading to semantic errors in number naming. Interestingly, this interference effect has not been described in the literature with healthy people in repetition and reading, whereas it can be easily explained in the case of many PWA. Since they suffer from damage to the sublexical route for reading or repetition (usually patients with conduction aphasia), they rely mainly on their semantic route, the one that is used in picture naming and where the interference effect is originated.

To sum up, when assessing STEPS, there are differences related to the presentation conditions (homogeneous versus heterogeneous lists, and repeated versus non-repeated lexemes) as well as others related mainly to the characteristics of the stimuli to be compared (single names are compared to longer, and more demanding, sequences of multi-digit words that are essentially compound words composed of very short and high-frequency lexemes). It seems to us that, on the one hand, the presence of semantic errors with numbers that characterizes STEPS may be favored by the presentation of numbers in homogeneous lists, repeating the same lexemes in cycles and under higher memory load conditions. On the other hand, the nature of the numerical system that creates multi-digits by combining short and very high-frequency numerical lexemes may protect them from phonemic errors.

1.3. *The present research*

The STEPS effect can be considered as comprising two elements: (1) the existence of phonemic errors in words but their absence in multi-digit numbers, and (2) the presence of semantic errors in multi-digit number production but their absence in word production. In the following studies we will explore: first, whether differences in memory load can explain the predominantly semantic nature of errors in the production of number word sequences and the predominantly phonemic errors in isolated words (Study 1); second, whether it is a combination of memory load, semantic blocking, cyclic naming and use of high frequency stimuli that causes STEPS; to this end, we will compare the production of multi-digit numbers to the production of sequences of high-frequency color words (Study 2a); finally, we will test the relevance of frequency in the emergence of phonemic errors by using low-frequency color words (Study 2b). Before reporting the studies, we will describe the participants and some procedural details.

2. General methods

All the materials used in the studies, their corresponding datasets, and codes employed to analyze and visualize the data, are available at our Open Science Framework (OSF) repository (https://osf.io/5v7nr/).

2.1. *Participants*

Two female patients participated in these studies (a more detailed description can be found in García-Orza et al., 2020). The studies reported in the current paper were carried out between January 2018 and June of 2019. During this time their neuropsychological condition remained stable. The research was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans. The study was approved by the Local Ethics Committee of the Universidad de Málaga. Informed consent was obtained from both participants.

Table 1 – Results on the WAB for patients DNR and ML (from García-Orza et al., 2020).

Area	DNR	ML
Spontaneous language	14/20	16/20
Auditory-verbal comprehension	7.55/10	6.25/10
Repetition	4.2/10	4.6/10
Naming	6/10	7.3/10
Total score	31.75/50	34.35/50
Aphasia Quotient (AQ)	63.5/100	68.7/100

2.1.1. *Patient DNR*

DNR is a right-handed female patient who suffered a middle cerebral artery infarction in 2014. This caused an extensive lesion in left temporal-parietal areas that affected the angular and supramarginal gyri, and also portions of the arcuate fasciculus (see Fig. 1). DNR, who was born in 1943, learned to read and write in a school for adults. For most of her life she worked as a homemaker and seamstress.

When assessed for this research she was 75. She was stabilized and was receiving speech therapy sessions twice a week and took *Memantine* as pharmacological treatment. DNR's results on the Western Aphasia Battery (WAB; Kertesz, 1982) (AQ = 63.5) fits a classification of conduction aphasia (see Table 1), specifically, a conduction aphasia of the reproduction variety (e.g., see Shallice et al., 2000). Phonemic errors in repetition, naming and reading were her most common errors, with occasional instances of *conduite d'approche* (see Gutiérrez-Cordero et al., 2025, for a definition). Some word-finding difficulties that may be considered normal for her age were also noted. Her production was moderately non-fluent; she was able to name 4 animals in a minute and her sentences were agrammatical. Her comprehension of words was almost perfect, but she exhibited more difficulties in the comprehension of more complex utterances, such as sentences which increased in syntactic complexity.

This general pattern was confirmed with specific tests from the Spanish *Batería para la Evaluación de los Trastornos Afásicos* (BETA) (Cuetos & González-Nosti, 2009). Spared word comprehension was confirmed by phoneme discrimination, spoken word-picture matching, and synonym judgement tasks. Difficulties were observed in picture naming, repetition, and reading tasks, in which most of her errors were phonemic. Reading tasks suggested a phonological dyslexia, i.e., an impairment in the graphene to phoneme converter. The assessment of her auditory-verbal short-term memory showed better performance in recognition than in recall tasks. So, when a spoken output was demanded, a memory span of 2 was found that increased to 3 in the recognition task with words and to 5 for the recognition task with digits. Additionally, in a phoneme discrimination task with a delay of 5 sec, DNR's score was slightly smaller than the one observed in the same task without delay, and within the range of normality for that task according to the BETA. All these results point to the preservation of her phonological input buffer (PIB), but an impairment in the phonological output buffer (POB).

2.1.2. *Patient ML*

ML is a right-handed female patient who suffered a stroke in 2017 that affected anterior portions of the left middle cerebral artery. It caused brain damage in the superior temporal and middle temporal gyri, the inferior parietal cortex (i.e., angular and supramarginal gyri), the middle and inferior frontal gyri, the ventral premotor cortex, the insula, as well as underlying white matter tracts, including segments of the arcuate fasciculus (see Fig. 2). ML, who was born in 1949, had two years of formal schooling in which she learned to read and write words and numbers. She worked as a cleaner for most of her life. She had been retired for two months when she suffered the stroke. Three years after the stroke, when she was 68, she was assessed for this study and was diagnosed with conduction aphasia but with some characteristics of Wernicke's aphasia according to the WAB (AQ = 68.7) (see Table 1).

ML exhibited fluent spontaneous language almost without paraphasias. In conversation, as well as in picture naming, she sometimes showed minor difficulties in word retrieval. Her sentence production was grammatical, although she sometimes lost track of the conversation. She had almost no problems with yes/no questions, but struggled with more complex questions and showed more problems in executing sequential orders, completing sentences presented orally, and in auditory recognition.

In repetition tasks she produced mainly phonemic errors, with some occasional formal and semantic paraphasias, and perseveration. On the contrary, she was good in naming and reading tasks, and thus it can be considered that she suffers from a conduction aphasia of the repetition variety (e.g., Martin et al., 1999; Sidiropoulos et al., 2008), that is, an impairment in the phonological input buffer (PIB).

ML's diagnosis was confirmed with additional tests using the BETA. At the word level, a near-normal performance was observed in phoneme discrimination, auditory lexical decision and auditory word-picture matching tasks, and synonym judgement tasks, which demand spared auditory recognition

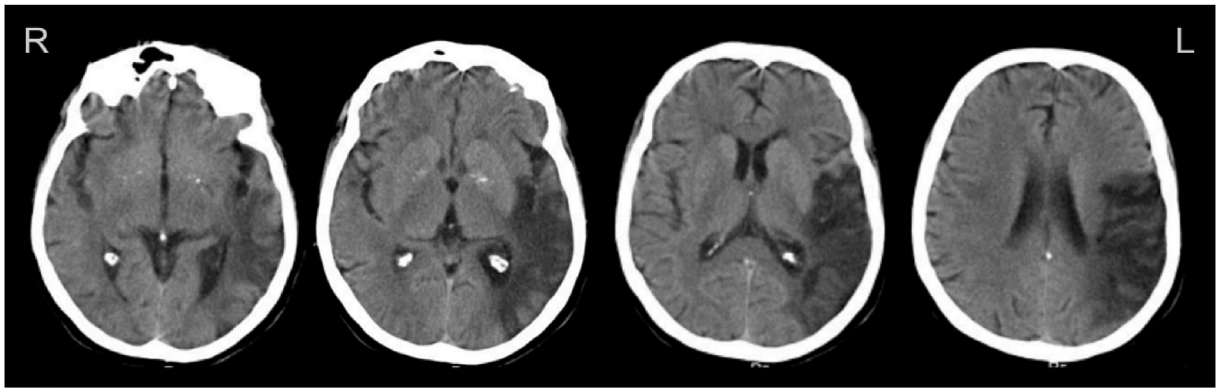

Fig. 1 – CT scan of DNR (reproduced with the permission of Elsevier).

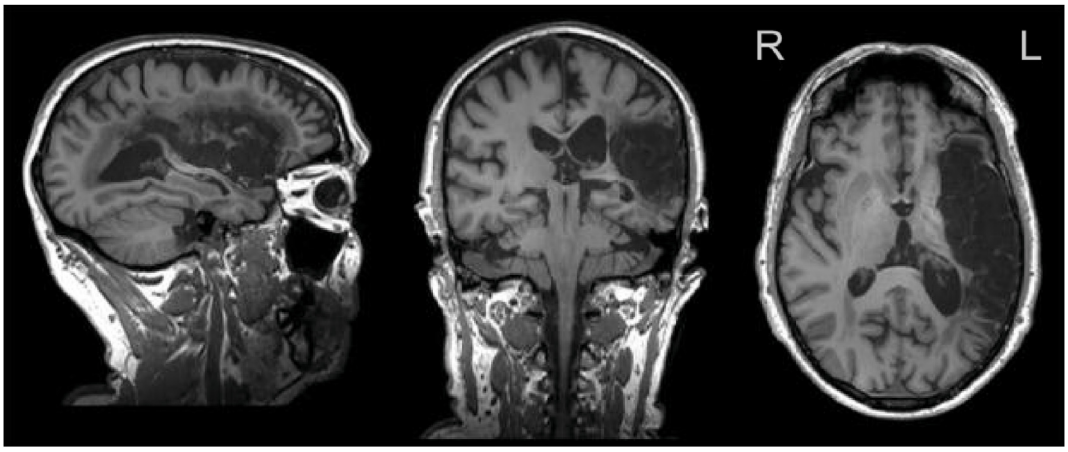

Fig. 2 – MRI of ML showing damage in left temporal areas (reproduced with the permission of Elsevier).

skills, including an unimpaired semantic system. Although she occasionally experienced difficulties with picture naming, which we suspect were influenced by her limited sociocultural exposure, her overall performance in this test was within normal limits; word reading was also normal. On the contrary, many difficulties were observed in word and pseudoword repetition. She had also difficulties in reading pseudowords, suggesting a graphene to phoneme impairment. In line with her problems with repetition, the assessment of her auditory-verbal short-term memory showed a forward digit-span of 2 items and a backward digit-span of 1, these scores being pathological for her age group, indicating a deficit in her verbal short-term memory. Additionally, her score in recognition tasks was also impaired. In a phoneme discrimination task, her score fell from 28/32 in a non-delayed condition to a 16/32 (i.e., a score no different from chance) in a 5-s delayed condition. All these results point to the preservation of her POB, but impairment in the PIB.

2.2. *Inclusion criteria*

Two main characteristics made these patients particularly relevant to our study. First, both had a phonological deficit, but with different characteristics. ML had a PIB impairment, whereas DNR had a POB impairment, and we note that the STEPS have been considered to be caused by a deficit that is restricted to the POB (Dotan & Friedmann, 2015). The two patients fulfilled the features that distinguish these two impairments according to Gvion and Friedmann (2012): (a) in word production, PIB patients typically show difficulties limited to repetition tasks, whereas POB patients show similar difficulties in repetition, reading and naming; (b) patients with impairment in the PIB show difficulties in STM tasks that demand both word recall and word recognition, whereas POB patients show difficulties mainly in word recall.

Second, both participants had previously shown a STEPS pattern (see García-Orza et al., 2020). Using the usual procedure in STEPS studies (i.e., without a proper control of the variables described above), they were presented with 60 two-digit numbers (larger than 30), 20 three-digit numbers, and 20 four-digit numbers and 60 common nouns from several different semantic categories and asked to repeat, name and read them aloud. Both participants committed significantly more phonemic errors than semantic errors in word repetition (DNR: 12 *vs* 0; ML: 7 *vs* 1) and more semantic errors than phonemic in multi-digit number repetition (DNR: 24 *vs* 2; ML: 34 *vs* 2). Additionally, as expected, due to the locus of their impairments, DNR also showed STEPS in naming (22 phonemic *vs* 4 semantic errors in picture naming, whereas 35 semantic *vs* 0 phonemic errors in naming Arabic multi-digits) and reading (14 phonemic errors *vs* 1 semantic error with noun words, but 47 semantic *vs* 2 phonemic errors in reading number words). By contrast, ML showed partial STEPS in naming (she made similar number of semantic and phonemic errors in word naming, 3 *vs* 5, and more semantic than phonemic errors in multi-digit naming, 11 *vs* 1), with the reverse pattern in reading (4 phonemic errors *vs* 0 semantic errors in word reading, and the same number of semantic and phonemic errors, 4, in multi-digit word reading). However, both in reading and naming tasks, as expected due to her impairment, ML's errors were not as numerous as in repetition, so the results of these tasks should be taken with caution.

Interestingly, in both PWAs we discarded the possibility of an impairment in the number semantic system, the one devoted to understanding quantities, which might have explained the presence of semantic errors in number production tasks. The performance in number comparison tasks with single digits and multi-digit numbers was perfect (10/10). Additionally, ML performed perfectly in a comparison task with the same numbers presented as words, whereas DNR, who has reading difficulties, showed more challenges in this task (5/10), but was able to correctly answer 14 questions related to numerical knowledge when giving a written response (e.g., "How long is the first half of a football match?"). Neither had issues naming single digits or in a comparison task with dots. These findings indicate that their numerical semantic knowledge was preserved.

2.3. *Error classification*

Speech errors were classified using a classical taxonomy (e.g., Dotan & Friedmann, 2015) with some small differences (see below): *semantic errors* were real words semantically related to the target word; *phonemic errors* were nonwords that contained at least 50% of the phonemes of the target word; *neologisms*

were nonwords that contained less than 50% of the phonemes of the target word; *formal errors* were real words phonologically related to the target word; *unrelated errors* were real words not related semantically or phonologically to the target word; *no attempt* to produce a response was considered as a nonresponse; *lexical errors* in numbers were real number words different from the number word corresponding to the target; *syntactic errors* in numbers were real number words with a different place-value structure to that in the target (e.g., 47 for 407); both lexical and syntactic errors with numbers were considered *semantic errors* following the criteria used by Dotan and Friedmann (2015); *counting errors* were identified when the PWA was involved in the strategic use of counting sequences with the aim of producing the target number (e.g., 10, 20, 30 … 35 for 35); *omissions* were considered when a full word, within a sequence, was omitted (e.g., 45 for 345 or "red - green" for "red - yellow - green"). Finally, we adopted a simple view in our identification of perseverations, and considered only those within the multi-digit or the color sequences (e.g., red-green-red for red-green-yellow), and not perseverations between sequences, since using a small, closed set of semantically-related stimuli makes it difficult to distinguish between semantic errors and perseverations.

Following García-Orza et al. (2020), errors that shared characteristics with two different types of errors were classified as both (e.g., saying "thirtrin" instead of "thirty-three" implies the production of both a semantic and a phonemic error) but, in each sequence, each type of error was counted once (e.g., in 234 for 278 there are two semantic substitutions but we counted this as one semantic error). However, we distinguished between semantic errors and omissions in numbers, since the latter might have their origins in a post-lexical level; this is in contrast to Dotan and Friedmann (2015), who considered them as one type of error, they also distinguished between phonemic errors and formal errors inasmuch as the latter might have their origin in a lexical level (e.g., Gold & Kertesz, 2001; Martin & Saffran, 1992). Note that these criteria were adopted in order to be more conservative in our analyses.

3. Study 1: increased memory load hypothesis

A fact that has been consistently ignored in the literature is that STEPS is found when comparing single-words to multi-digit numbers, instead of using one-digit numbers. In this last type of stimuli, the number of errors is typically small or even absent (e.g., Bachoud-Levy & Dupoux, 2003; Delazer & Bartha, 2001; Dotan & Friedmann, 2015; García-Orza et al., 2020; Girelli & Delazer, 1999; Marangolo et al., 2004). By their nature, multi-digit numbers, composed of a combination of several lexemes (and morphemes), involve the production of a longer sequence of elements (e.g., "forty-two", "two-hundred and forty-one") that must be maintained in memory while producing the initial part of the sequence, as opposed to single words that entail a single lexeme (e.g., "tale") and hence require less memory load. In this study we explore whether this difference in the amount of information may account for the STEPS effect or, at least, for part of it, namely, the presence of semantic errors when producing multi-digit numbers.

Our hypothesis is driven by the fact that interactive models of language processing have emphasized that memory load can induce a change in the nature of errors made during speech production (Dell, 1986; Dell & O'Seaghdha, 1992; Dell et al., 1997; Martin et al., 1994). In line with this, Martin and Saffran (1992) postulated the *Continuum Hypothesis*, in which phonemic, formal, and semantic errors lie on a continuum of severity determined by the auditory-verbal short-term memory (AV-STM) demand of the task. On this account it is assumed that, especially in repetition tasks, the input spreads activation across a nodal network from phonological to lexical and semantic levels (Martin et al., 1994, 1996; Martin & Saffran, 1992). Information in the phonological nodes is the most vulnerable to decay due to its temporary nature, whereas lexical and semantic levels take advantage of being long-term representations. Consequently, the language system would rely on the lexical/semantic nodes when the phonological nodes are not available due to the decay of phonological information caused by a phonological impairment, an increase in the AV-STM task demand, or both. The prediction of this account is that memory load, operationalized as demanding the production of word pairs instead of a single word, in PWA with phonological deficits would lead to: (a) a worse general performance, and (b) the emergence of more semantic and formal errors in contrast to phonemic errors, particularly in the second word of a sequence, this is, when phonological resources are overloaded but the semantic nodes are still activated (Martin et al., 1996). Gold and Kertesz (2001), who asked a person with conduction aphasia to repeat single-word and two-word sequences, confirmed the idea of errors changing across this continuum in relation to the amount of memory load in AV-STM. Phonemic errors were predominant in conditions of low memory load (single words and in the first item of two-word sequences), whereas semantic errors increased in conditions of high memory load, that is, in the second item (Gold & Kertesz, 2001). Using a different way of increasing memory load, a 5-s delay in a repetition task, Jefferies et al. (2006) also found an increase in semantic errors in a group of brain-damage persons with phonological disfunction.

Taking all of this into consideration, we consider here whether the production of semantic errors in STEPS is caused by differences in AV-STM load during speech production. While non-numerical words are repeated in isolation and are produced with more phonemic errors, multi-digit numbers, which are in essence composed of multiple lexemes, form longer sequences, and which are thus more demanding stimuli, are associated with the emergence of more semantic errors (García-Orza et al., 2020). We test this hypothesis by comparing PWA's performance in a single word repetition task and a paired-word repetition task. According to our hypothesis, we expect the two participants to produce: (a) more errors in paired words than in single words, and (b) a change in the proportion of phonemic/semantic errors, from more phonemic errors when repeating single words to more formal and semantic errors when repeating paired words, especially in the second word of the pair. On the contrary, according to the BBH, and given that we used non-number words, we

Table 2 – Characteristics of the words in Study 1 for both single and paired conditions.

	Single words		Paired words	
	M (*SD*)	Range	*M* (*SD*)	Range
Frequency	84.83 (107.29)	.50–591.35		
Imageability	4.63 (1.58)	2.06–7.00		
Number of phonemes	5.82 (1.21)	3–10	11.64 (1.75)	7–16
Number of syllables	2.50 (.57)	2–4	5.00 (.78)	4–7
Number of letters	6.01 (1.10)	4–10	12.01 (1.55)	9–16

Note. Mean frequency (per million) and imageability is the same for single and paired words.

expected more phonemic than semantic errors without notable effects of memory load on the type of errors, apart from an increase in the overall number of errors.

3.1. *Materials*

An initial list of 160 words was selected from the EsPal database (Duchon et al., 2013). These words were nouns of various semantic categories, such as agriculture and fisheries, everyday objects, science and academy, places, and attitudes. Similar proportions of high and low frequency and high and low imageability words were selected, although since these words differed in number of phonemes, we did not analyze the role of these variables. From the initial 160 single words, two lists were derived: (a) a list of single words was semi-randomly created in an attempt to avoid the contiguous presentation of words of the same semantic category, and (b) a list of word pairs, in which the word pairs were created semi-randomly toward avoiding semantic relationships between the words in each pair. The linguistic properties of the selected stimuli are presented in Table 2.

3.2. *Procedure*

Words from both single words and paired words lists were read aloud by the experimenter, leaving 1 sec between the words in the case of the paired words. Participants were encouraged to repeat the word or the word pair immediately after. The list of word pairs was presented first, and after a rest and other tasks, the single word list was presented, all in a single session.

3.3. *Specific data analyses*

All statistical analyses in this study were conducted in R (version 4.1.2 – "Bird Hippie"; R Core Team, 2021).[3] The STEPS was considered by comparing phonemic versus semantic errors across presentation conditions (isolated versus paired words) using Wilcoxon's signed-rank tests. Differences in correct responses and error types between conditions and positions (position 1 *vs* 2) in the paired-word condition were assessed using Pearson's chi-square tests (with Yates' correction when needed).

[3] A complete list of the software used is available at our OSF repository (https://osf.io/5v7nr/).

Table 3 – Patients' performance and errors in the repetition task with single words and with word pairs in Study 1.

Patient	Repetition task	Correct Responses	Errors						
			Total	Semantic	Phonemic	Formal	Neologism	Perseveration	Unrelated	Omission
DNR	Single (n = 160)	128	32	0	**24**	4	0	2	0	2
	Paired (n = 80)	41	47	0	**23**	9	0	5	1	9
	Position 1 (n = 80)	57	23	0	**12**	2	0	3	1	5
	Position 2 (n = 80)	56	24	0	**11**	7	0	2	0	4
ML	Single (n = 160)	120	40	0	**30**	6	2	2	0	0
	Paired (n = 80)	16	86	2	**37**	12	6	22	1	7
	Position 1 (n = 80)	56	24	0	**20**	2	1	0	0	1
	Position 2 (n = 80)	18	62	2	**17**	10	5	22	1	6

Note. More than one error may occur in each item. Errors in the paired condition correspond to the sum of errors in each word at each position. Number of semantic and phonemic errors are in bold when the comparison between them showed significant differences.

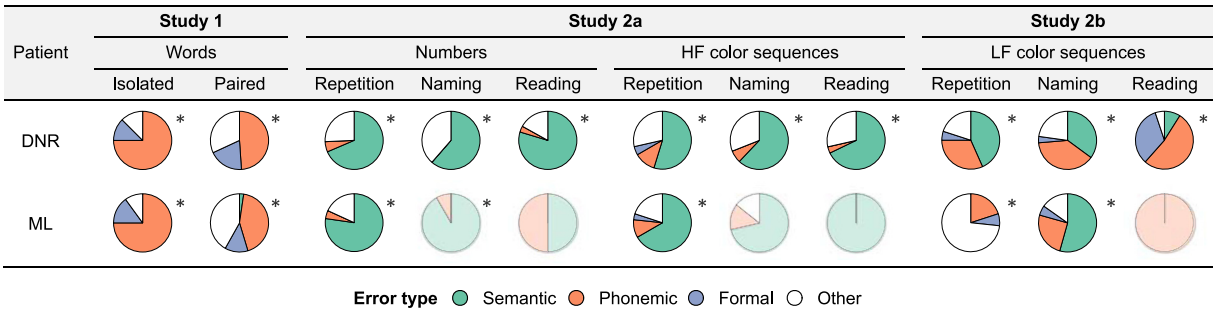

Fig. 3 – Proportion of errors across repetition task conditions (single versus paired words) for patients DNR and ML in Study 1, as well as in reading, naming, and repetition tasks involving numbers (García-Orza et al., 2020) and high-frequency (HF) and low-frequency (LF) color sequences in Studies 2a and 2b. *Note.* A full pie chart represents the total number of semantic, phonemic, formal, and other types of errors. * indicates significant differences ($p < .05$) between semantic and phonemic errors. Less intense colors in the pie charts indicate good performance in the tasks (i.e., few errors, ≤12). ML was only assessed for 20/60 sequences with LF colors (the ones belonging to the length 2 condition) in the repetition task (Study 2b).

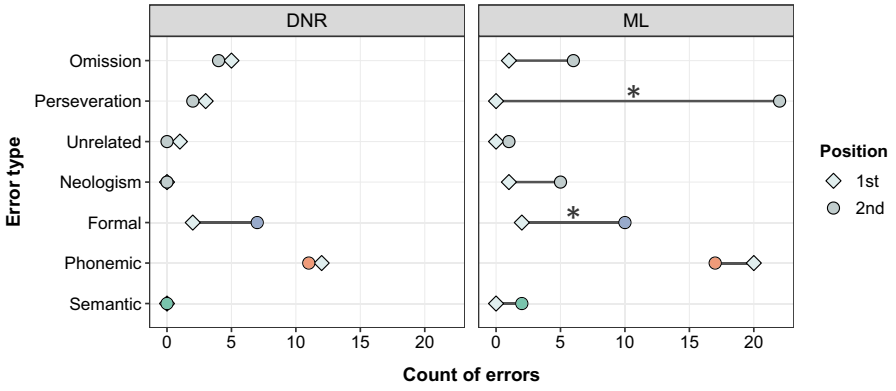

Fig. 4 – Errors per type and position during paired-word repetition of patients DNR and ML. *Note.* * indicates (corrected) significant differences ($p < .05$) in the number of errors between positions 1 and 2.

3.4. *Results*

DNR and ML performance are presented in Table 3. In Fig. 3 we offer a graphic representation of the proportion of errors of both participants in the single and paired-word repetition task; in Fig. 4 we reported the differences in the number of types of errors in words at positions 1 and 2 during the paired-word repetition.

3.4.1. *Patient DNR*

DNR correctly repeated 128 out of 160 (80%) single words, whereas the repetition of paired words resulted in more difficulties: she correctly produced 113 out of 160 (71%) (41 correctly repeated pairs of words out of 80, 51%). More interestingly, in both conditions she produced predominantly phonemic errors (24/160 in single words and 23/160 in paired words), but no semantic errors, and hence the difference between these types of errors was significant both in the single word condition ($Z = 4.90$, $p < .001$) and in the paired-word condition ($Z = 4.80$, $p < .001$).

The comparisons of the overall number of errors between conditions showed fewer errors in the single word condition, but this difference was only marginally significant [$\chi^2(1) = 3.78$, $p = .051$]. When we compared the number of phonemic errors (24 *vs* 23), no differences were observed [$\chi^2(1) = .02$, $p = .87$], and the same was found with the semantic errors (0 *vs* 0). A non-significant increase, from 4 to 9, in formal errors was found [$\chi^2(1) = 1.28$, $p = .26$], and regarding the remaining errors, again non-significant increases were found from the single to the paired condition in omissions (from 2 to 9, $p = .066$), perseverations (from 2 to 5, $p = .44$), or unrelated errors (from 0 to 1, $p > .99$).

Thus, the pattern with pairs of words was roughly similar to that with single words, and as usual within STEPS, phonemic errors were the most frequent type of error.

3.4.1.1. Serial position effect. In the repetition of word pairs, DNR's responses were equally accurate at position 1 (71.25%) and position 2 (68.75%) [$\chi^2(1) = .03$, $p = .86$]. When we compared

the number of each type of errors in each position (see Fig. 3), only a non-significant increase in the number of formal errors at position 2 was observed (from 2 to 7) [$\chi^2(1) = 1.88$, $p = .17$]. The rest of errors varied minimally (all $ps > .82$).

3.4.2. *Patient ML*

In the repetition of single words ML was much more accurate, 120 words out of 160 (75%), than in the paired word, in which she only correctly produced 74 words out of 160 (46.25%), (only 16 correct pairs out of 80, 20%). She made 40 errors with single words, most of these phonemic (30), and no semantic errors, this difference being significant ($Z = 5.47$, $p < .001$). In paired-word repetition there were again more phonemic (37) than semantic (2) errors ($Z = 5.60$, $p < .001$).

The number of errors in the repetition task with word pairs was about twice than with single words (86 and 40 errors, respectively) [$\chi^2(1) = 27.70$, $p < .001$]. This increase in the number of errors was observed in almost all types of errors: it was slight in semantic, neologisms, and unrelated errors; clear but non-significant, in the case of phonemic (from 30 to 37) [$\chi^2(1) = .93$, $p = .33$] and formal errors (from 6 to 12) [$\chi^2(1) = 2.12$, $p = .15$]; but significant for perseverations (from 2 to 22) [$\chi^2(1) = 18.02$, $p < .001$] and omissions (from 0 to 7) [$\chi^2(1) = 4.25$, $p = .039$].

3.4.2.1. Serial position effect. In the repetition of word pairs ML showed better performance with words at position 1, with 70% success (56/80), than with words at position 2, with only 22.5% success (18/80) [$\chi^2(1) = 36.30$, $p < .001$]. This indicated that ML had lost the recency effect, as expected given her considerable AV-STM impairment (Majerus et al., 2015; Martin & Saffran, 1997; Salis et al., 2015). In terms of the different types of errors, increases were found for all of these apart from phonemic errors, which showed a non-significant decrease (from 20 to 17). Within the increases from position 1 to 2, the greatest were for perseverations (from 0 to 22) [$\chi^2(1) = 25.51$, $p < .001$], and formal errors (from 2 to 10) [$\chi^2(1) = 4.41$, $p < .036$]. Finally, for semantic errors, unrelated errors, omissions and neologisms there were non-significant increases, all $ps > .21$. The change in the proportion of errors is presented in Fig. 4.

3.5. *Discussion*

In this study we explored whether the production of semantic errors in the STEPS arises from the fact that word production is assessed using single words whereas multi-digit numbers are generally assessed using strings of various words. In line with the Continuum Hypothesis (Martin et al., 1994, 1996; Martin & Saffran, 1992), we hypothesized that the increasing load involved in PWAs repeating pairs of words, instead of single words, would lead to an increase of formal and semantic errors.

Our results showed, not surprisingly, that DNR and ML produced more errors when repeating paired words in contrast to single words. This increase seems to be caused by the increase in the amount of information to be processed and produced—the AV-STM demand—during the repetition of paired words. Such an effect was clearly stronger in the case of ML, who has a PIB impairment, and was only marginal for DNR.

In terms of the type of speech errors, it was found that: (a) phonemic errors were consistently the more numerous errors across conditions in both the single and the paired conditions; (b) semantic errors were almost absent in the case of ML (with none for DNR) and were restricted to the second word of the pair (echoing the performance of patient MMB described by Gold & Kertesz, 2001); (c) formal errors increased from the single to the paired words in both patients, whereas phonemic errors remained the same for DNR and increased slightly for ML; additionally, formal errors increased clearly in words repeated in the second position compared to the first position, and although this difference (from 2 to 7) was not significant in the case of DNR, it was (from 2 to 10) in the case of ML. So, the increase in formal errors occurred under the more demanding conditions, at the second word of the pair, and more clearly in the PWA who showed the more reduced STM span, ML. This may suggest that a stronger manipulation of the load may lead to a stronger change of tendency that should extend to an increase in semantic errors.

So, the overall pattern of errors in this study, at least regarding an increase in formal errors, partially supports the Continuum Hypothesis (Gold & Kerstez, 2001; Martin & Saffran, 1992; Martin et al., 1994, 1996). However, against our predictions about the origin of STEPS, no evidence of load causing a change from phonemic to semantic errors was found.

4. Study 2a: resemblance of colors and numbers regarding STEPS

The results of the first study showed that the increase in memory load alone is not able to explain the semantic errors in numbers that characterize STEPS. One possibility is that the manipulation we introduced in the study was not strong enough to reveal these effects. For instance, the load in two-word sequences is still far less than the load in three- or four-digits numbers (i.e., 3–5 words). Additionally, number words have higher frequency and imageability than those words in Study 1, and this, according to Gold and Kertesz (2001), might increase the change from phonemic to formal or semantic errors, in that there is more reliance on the lexical route in stronger representations. Besides, the words high frequency may reduce the production of phonemic errors (e.g., Shallice et al., 2000). Finally, in contrast to Study 1, when producing lists of multi-digit numbers, participants are asked repeatedly produce the same limited number of words from the same semantic category, and this, as noted in the Introduction, resembles the cyclic naming paradigm that causes interference effects in naming due to the over-activation of those semantic representations (e.g., Belke, 2017; Oppenheim et al., 2010), thus adding another factor which potentially increases semantic errors over phonological ones when producing numbers.

To explore this, in a second study we take words that resemble numbers in their lexical-semantic characteristics and use them to build word sequences that are similar to the multi-digit number words. We expect that by putting semantically-related words, with high frequency and imageability, in sequences of 2–4 words (to increase load), and by

Table 4 – Linguistic properties of multi-digit numbers and color sequences (all $n = 60$) assessed in Study 2a and Study 2b.

	Study 2a				Study 2b	
	Numbers		High-frequency (HF) colors		Low-frequency (LF) colors	
	M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range
Lexical frequency (log count per million)	2.08 (.54)	.39–2.92	1.62 (.22)	1.12–2.12	.22 (.14)	.027–.62
Lemma frequency (log count per million)	4.66 (.64)	2.47–5.69	4.41 (.27)	3.85–5.01	2.35 (.32)	1.44–3.06
Number of phonemes	21.25 (7.42)	10–34	14.65 (4.19)	8–24	14.43 (4.30)	6–25
Number of letters	21.68 (7.47)	11–34	15.07 (4.38)	8–25	14.88 (4.18)	8–25
Number of syllables	8.22 (2.53)	4–13	5.92 (1.77)	3–10	5.90 (1.74)	3–10

Note. The properties for multi-digit numbers were computed based on all their constituent words, in the same way as for the color sequences.

presenting them repeatedly (to increase interference), we will be able to induce the production of semantic errors in our subjects. To this end, we selected highly familiar colors, since (like numbers) these also constitute a semantically-homogeneous group with high lexical frequency, high imageability, short names, and common syllabic combinations. Additionally, we assess DNR's and ML's performance in color sequences in the three tasks in which STEPS has already been reported (reading, repetition, and naming tasks), then we compare this to the performance in (parallel sequences of) multi-digit numbers described in García-Orza et al. (2020).[4]

According to the BBH, since color words are content words and thus are built by combining individual phonemes extracted from the POB, more phonemic than semantic errors are expected in these sequences. On the contrary, the emergence of semantic errors in the production of high-frequency (HF) color sequences (the error pattern characteristic of numbers in STEPS) would support the claim that this phenomenon arises as a consequence of the experimental conditions and characteristics of the stimuli used. Specifically, in the case of DNR, who suffers from a POB impairment, we expect more semantic than phonemic errors in the three tasks. On the contrary, given that ML is only impaired in her PIB, we expect more semantic errors in repetition and a low number of errors in the reading and naming tasks, as she is supported by the visual and permanent presentation of the information in these tasks. These were in fact the results obtained with multi-digit numbers in García-Orza et al. (2020); our focus here, then, is to see whether we can replicate this with non-number words, in this case, sequences of HF-color words.

4.1. *Materials*

We designed a list of color sequences that was equivalent to the list of multi-digit numbers used in a previous study with these participants (see García-Orza et al., 2020). That list was composed of 20 two-digit numbers larger than 30, 20 three-digit numbers, and 20 four-digit numbers. To create the parallel color sequences from this number list, each number from 0 to 9 was associated with a HF-color.

These color words had a similar number of syllables, phonemes and letters as single number words (all *ps* > .18). However, although we used the most frequent colors, according to EsPal (Duchon et al., 2013) color words were still far less frequent than single numbers in both lexical and lemma frequency [regarding the logarithm of their count per million, $t(15.83) = 3.81$, $p = .001$ and $t(12.46) = 2.18$, $p = .048$, respectively].

To create the color sequences, we replaced each number in the multi-digit list with its corresponding color (e.g., for the number *trescientos sesenta y cuatro* ["three hundred sixty-four", 364], *rojo - verde - gris* ["red - green - grey"] was the corresponding color sequence). Although the number of elements was matched in the color and number sequences, multi-digit numbers were linguistically both more frequent and larger than color sequences (note that the words "hundreds" and "thousands" had no equivalents in our sequences of colors), and this was also true for the mean log lemma frequency per multi-digit/sequence according to Welch *t*-tests (all *ps* < .009). The linguistic properties of the lists are reported in Table 4.

4.2. *Procedure*

Previous studies have shown STEPS in repetition, naming and reading (e.g., Dotan & Friedmann, 2015). To confirm our hypothesis, we evaluated performance using HF-color sequences across the three tasks, following the same order as in García-Orza et al. (2020). The order of presentation for the color sequences was randomized for each task.

4.2.1. *Repetition task*

The experimenter read each stimulus once, but it could be repeated if patients asked. Patients were encouraged to repeat the stimulus immediately after the experimenter read it.

4.2.2. *Reading task*

Color sequences (separated by dashes, e.g., "green - blue") were presented on a laptop with a 15.6″ monitor using PowerPoint. Each full sequence was presented in a single presentation slide, words were situated in the center of the screen using Calibri 60-point black font over a white background. Patients were asked to read the words on the screen as soon as they were presented. Each sequence remained visible on the screen until a response was made.

[4] Neuropsychological conditions of the patients did not vary between the administration, approximately 6 months, of the tasks used in this and the following study in terms of WAB measures.

4.2.3. *Naming task*

Sequences of color patches were presented in the same way as in the reading task, i.e., horizontally aligned over a white background on a single slide. Each color patch was a 5.47 cm × 6.01 cm rectangle. Patients were asked to name the sequence of color patches from left to right. The sequence of color patches remained on the screen until the response.

Additionally, before presenting the sequences, to ensure both PWA were able to correctly name the color from the color patches, they were asked to name them. DNR committed 4 phonemic errors with HF colors, but no other errors were observed. ML made 2 phonemic errors and 1 semantic error when producing single HF colors. This was slightly worse than their performance with single-digits, which they produced without errors (see García-Orza et al., 2020, Footnote 4).

4.3. *Specific data analyses*

We compared semantic versus phonemic errors across tasks and sequence lengths for HF-color sequences using Wilcoxon's tests (the corresponding analysis for multi-digit numbers is reported in García-Orza et al., 2020, thus is only discussed here). General performance and errors with HF colors were compared to numbers in García-Orza et al. (2020) using Pearson's chi-square tests (with Yates' correction when needed). Differences between tasks or sequence lengths (both three-level factors) for HF colors and numbers were analyzed using Friedman's chi-square test, followed by Bonferroni-corrected Wilcoxon tests when significant effects were found.

4.4. *Results*

The results of the performance of both participants are presented in Table 5 and Fig. 3, with Fig. 5 providing a graphical representation of performance regarding length. Importantly, we scored the correctness of each complete sequence, but to compute errors we considered each color word of each sequence; thus, more than one type of error may be found in a single sequence and even in a single word. Since formal errors were usually absent, no analyses were run on this type of error unless they were clearly numerous.

For clarity, we have streamlined the results section to emphasize the key aspects of the STEPS phenomenon, presenting the main findings for each patient. Readers seeking a more detailed report can refer to the Supplementary Materials (also available on our OSF: https://osf.io/5v7nr/). For the sake of simplicity, we have used the symbol Δ in some instances to denote the difference in the number of errors (or correct responses) between conditions.

4.4.1. *Patient DNR*

4.4.1.1. General exploration of STEPS. DNR displayed the same pattern of errors with HF-color sequences as she did with multi-digit numbers (as found by García-Orza et al., 2020), producing consistently more semantic than phonemic errors across tasks: repetition (23 semantic *vs* 5 phonemic errors), naming (36 semantic *vs* 4 phonemic errors), and reading (19 semantic *vs* 1 phonemic error) (all $Z > 3.52$, $ps < .001$) (see Fig. 3 for a comparison).

4.4.1.2. Multi-digits numbers versus HF colors. The performance of DNR in repetition was similar for multi-digit numbers and HF colors (Δ = 1.2% accuracy) [$\chi^2(1) = .14$, $p = .71$], but was worse for numbers in naming (Δ = 18.33% accuracy) and reading ($\Delta = 53.33\%$ accuracy) [both $\chi^2(1) \geq 8.29$, $ps < .001$]. The number of semantic errors did not differ between numbers and colors in either the repetition or naming tasks (Δ was 1 error in both cases) [$\chi^2(1) = .03$, $p = .85$], but significantly more instances were found with numbers in reading ($\Delta = 28$ errors) [$\chi^2(1) = 26.40$, $p < .001$]. The number of phonemic errors did not differ between colors and numbers across tasks (all $\Delta s \leq 4$ errors) [all $\chi^2(1) \leq 2.64$, $ps \geq .13$] (see also Table 5 for a reference).

4.4.1.3. Length effects. A length effect was observed in DNR's global performance, with poorer performance on longer stimuli across all tasks and stimulus types [for HF colors: all $\chi^2(2) \geq 7.63$, $ps \leq .022$; for numbers: $\chi^2(2) = 2$, $p = .049$, in naming, and $\chi^2(2) = 22.84$, $p < .001$, in repetition], except for numbers in reading [$\chi^2(2) = 4$, $p = .14$]; this general decrease varied slightly across tasks and stimulus types. For semantic errors, she also exhibited a length effect, with errors increasing as stimulus length grew, in all the tasks involving HF colors [repetition: $\chi^2(2) = 6.5$, $p = .039$; naming: $\chi^2(2) = 11.57$, $p = .003$; reading: $\chi^2(2) = 6.17$, $p = .046$]. When using multi-digit numbers, the length effect was observed in repetition [$\chi^2(2) = 10.94$, $p = .004$], but it was absent in reading and naming [$\chi^2(2) = .89$, $p = .64$ and $\chi^2(2) = 5.29$, $p = .071$, respectively]. Phonemic errors were scarce (five or fewer in total), but when they could be analyzed, no length effect was observed for either numbers or HF colors [all $\chi^2(2) \leq 2$, $ps \geq .14$] (see Fig. 5).

We also contrasted the number of semantic and phonemic errors for each stimulus length. In length 2 sequences, there was a consistent tendency to produce more semantic than phonemic errors in all cases; however, the differences were scarce in all tasks with HF-colors and for multi-digit repetition (all $\Delta s \leq 5$ errors) (all $Z \leq 2.24$, $ps \geq .063$), and were significant only for number naming ($\Delta = 12$ errors) ($Z = 3.46$, $p < .001$) and number reading ($\Delta = 14$ errors) ($Z = 3.77$, $p < .001$). When using stimuli of lengths 3 and 4, DNR made significantly more semantic than phonemic errors in all tasks with both types of stimuli (all $\Delta s \geq 6$ errors) (all $Z \geq 2.45$, $ps \leq .031$) and these differences generally increased, except in number naming and reading, where the trend remained stable at these lengths (see Fig. 5).

To sum up, DNR committed significantly more semantic than phonemic errors in the three tasks with HF-color sequences, resembling the pattern found with multi-digit numbers. The manipulation of length reduced the number of correct responses and increased the semantic errors in the three tasks with HF-color sequences. This same pattern was observed in the repetition task with numbers, and a non-significant trend in the naming and reading tasks. More interestingly, the presence of significantly more semantic than phonemic errors in the three tasks was observed from length three-color sequences onward, whereas it was present from two-digit numbers in naming and reading and from three-digit numbers in repetition. This means that three-word

Table 5 – Patients' performance and errors in the repetition, naming, and reading tasks with multi-digit numbers, HF-color sequences (Study 2a) and LF-color sequences (Study 2b) ($n = 60$ items per task and stimuli type).

Patient	Task	Stimulus type	Correct Sequences	Errors								
				Total	Semantic	Phonemic	Formal	Neologism	Perseveration	Unrelated	Omission	Addition
DNR	Repetition	Numbers	25	35	**24**	**2**	0	0	0	0	9	0
		HF colors	27	42	**23**	**5**	2	0	1	0	9	2
		LF colors	23	60	26	19	3	2	0	0	10	0
	Naming	Numbers	3	57	**35**	**0**	0	0	0	0	22	0
		HF colors	14	58	**36**	**4**	0	0	0	0	18	0
		LF colors	16	57	20	22	2	4	0	0	9	0
	Reading	Numbers	2	59	**47**	**2**	0	0	0	1	9	0
		HF colors	34	28	**19**	**1**	0	0	0	0	8	0
		LF colors	14	73	**7**	**41**	26	4	0	0	0	0
ML	Repetition	Numbers	17	44	**34**	**2**	0	0	0	2	6	0
		HF colors	18	60	**40**	**6**	2	0	0	0	9	3
		LF colors[a]	2	29	**0**	**6**	2	11	7	0	4	0
	Naming	Numbers	48	12	**11**	**1**	0	0	0	0	0	0
		HF colors	53	7	5	1	0	0	0	0	0	1
		LF colors	2	91	**50**	**23**	5	5	2	5	2	0
	Reading	Numbers	53	8	4	4	0	0	0	0	0	0
		HF colors	58	2	2	0	0	0	0	0	0	0
		LF colors	57	3	0	3	0	0	0	0	0	0

Note. More than one error may occur in each sequence and item. Correct sequences indicate the number of whole sequences produced correctly from a total of 60. Errors are computed for each word within a sequence, so the maximum number of errors is 180. Data on multi-digit numbers are drawn from García-Orza et al. (2020). Number of semantic and phonemic errors are in bold when the comparison between them showed significant differences (i.e., *ps* < .05).

[a] ML was only assessed for 20 sequences out of 60 (the ones belonging to the length 2 condition).

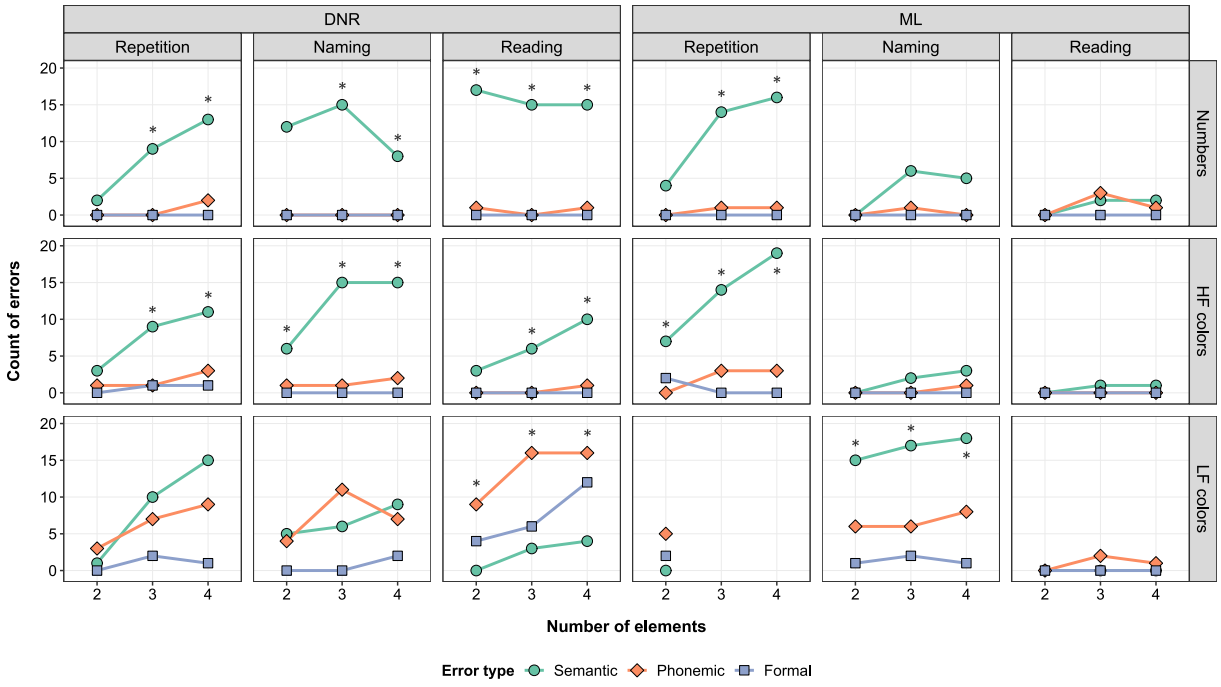

Fig. 5 – Formal, phonemic, and semantic errors made by DNR and ML in producing multi-digit numbers, high-frequency (HF) color word sequences, and low-frequency (LF) color word sequences of varying lengths, grouped by task. *Note.* * indicates (corrected) significant differences ($p < .05$) between semantic and phonemic errors.

sequences seem to constitute the load needed to make the effect appear in DNR.

4.4.2. *Patient ML*

4.4.2.1. General exploration of STEPS. The error pattern we found in ML's production of HF colors was similar to the one previously reported with multi-digit numbers by García-Orza et al. (2020). More semantic than phonemic errors were found in the repetition task (40 semantic *vs* 6 phonemic) ($Z = 5.67$, $p < .001$), and although a similar trend was observed in naming (5 semantic *vs* 1 phonemic) and reading (2 semantic *vs* 0 phonemic) (both $Z < 1.63$, $ps \geq .22$), as expected given the locus of her impairment, HF colors elicited very few errors in ML to make these differences significant.

4.4.2.2. Multi-digits numbers versus HF colors. ML's performance was slightly better with HF colors than with multi-digit numbers, though the differences were not statistically significant in any task (all $\Delta s \leq 8.83\%$ accuracy) [all $\chi^2(1) \leq 1.92$, $ps \geq .17$]. Both colors and multi-digit numbers elicited similar numbers of semantic errors (all $\Delta s \leq 6$ errors) [all $\chi^2(1) \leq 2.6$, $ps \geq .11$] and the same was found with phonemic errors (all $\Delta s \leq 4$ errors) [all $\chi^2(1) \leq 2.33$, $ps \geq .13$] in each task.

4.4.2.3. Length effects. An effect of length was observed in ML's performance in all number tasks [all $\chi^2(2) \geq 6.33$, $ps \leq .042$] and in the HF color repetition task [$\chi^2(2) = 23.74$, $p < .001$], with accuracy decreasing as stimulus length increased. In the two remaining tasks with HF colors (i.e., naming and reading), no effects of length were observed [both $\chi^2(2) \leq 5.43$, $ps \geq .066$]. Significant increases with length were observed on semantic errors in the repetition tasks for both numbers and colors [both $\chi^2(2) \geq 16.53$, $ps < .001$] and in the number naming task [$\chi^2(2) = 6.2$, $p = .045$]. The naming task for colors and both reading tasks showed no significant length effects [all $\chi^2(2) \leq 2.80$, $ps \geq .25$] (see Fig. 5). Regarding phonemic errors, these were scarce across tasks (all ≤ 6), and when analyzable, no length effect was found, either for numbers or for HF colors [all $\chi^2(2) \leq 3.5$, $ps \geq .17$].

We also compared the number of semantic and phonemic errors for each length condition. Although no significant differences were found in naming and reading tasks due to the low number of errors produced (all $\Delta s \leq 5$ errors) (all $Z < 2.24$, $ps \geq .063$), these differences emerged in repetition, the task in which ML faced more difficulties. Specifically, in HF color repetition, more semantic than phonemic errors were already observed at length 2 ($\Delta = 7$ errors) ($Z = 2.65$, $p = .016$), with this difference increasing with sequence length (for length 3 and 4, the Δs for errors were 11 and 16, respectively) ($Z = 3.05$, $p = .003$ and $Z = 4$, $p < .001$, respectively). A similar pattern was found in the repetition of multi-digit numbers, where, although differences in two-digit numbers were not significant ($\Delta = 4$ errors) ($Z = 2$, $p = .13$), they became significant for three-digit ($\Delta = 13$ errors) ($Z = 3.35$, $p < .001$) and four-digit numbers ($\Delta = 15$ errors) ($Z = 3.87$, $p < .001$).

Overall, ML's performance, as expected given her impairment in the PIB, shows considerably more errors in repetition than in reading and naming tasks, with both multi-digits and HF-color sequences. In terms of stimulus types, her performance was roughly similar. More importantly for the aims of this study, semantic errors were in general more frequent than phonemic ones, with both HF-color sequences and

numbers, and with length having a considerable role in increasing this pattern. These differences were clearly significant in the repetition task but less clear in the reading and naming tasks, for which there were few errors.

4.5. Discussion

In this second study we assessed whether the combination of different factors, namely, using high-frequency words, a higher memory load, and the continuous (cyclical) production of a limited set of semantically-related items, is decisive in the emergence of semantic errors that are observed with numbers in STEPS. The point here is that instead of using numbers we used sequences of HF-color words, that is, stimuli that according to the BBH should behave like non-number words and hence should be produced mainly with phonemic errors.

DNR' results, both with multi-digit and HF- color sequences, showed the same pattern: semantic errors were far more frequent than phonemic errors. This was found in the three tasks used, i.e., repetition, naming and reading, with slight differences in the strength of the effect. The results in the case of ML were also clear. She showed significantly more semantic than phonemic errors in the repetition of HF-color sequences. This pattern was not significant in the naming and reading tasks due to her relatively good performance in these tasks, but it did follow the same trend.

Regarding the study of changes in the error rate with memory load, we found that in general DNR and ML produced more errors as length increased. With some exceptions in specific tasks (e.g., omissions in DNR's naming task and ML's repetition tasks), the increase in errors with length was mainly associated with the increase in semantic errors, with increases in other types of errors, including phonemic ones, few in number. A task analysis showed some differences between stimulus types. In the case of DNR, the number of semantic errors with HF-color sequences increased significantly across lengths in the three tasks, and although the trend was similar for numbers, significant differences only arose in the naming and the repetition tasks. ML produced more semantic errors in longer sequences when repeating HF colors and numbers. Phonemic errors showed a similar trend, but this was not significant.

The findings here constitute evidence on how a pattern of errors (see Study 1) can be modified by a combination of several factors. We have, then, found support for the claim that the semantic errors in the STEPS are not intrinsically linked to numbers; rather, they emerge from the presentation of sequences of high-frequency words that belong to the same semantic category and are presented repeatedly. Multi-digit numbers meet the same requirements, as do sequences of HF-color words, and hence both are characterized by the production of predominantly semantic errors and only small numbers of phonemic errors.

5. Study 2b: the role of word frequency in STEPS

The previous study addressed one part of the STEPS paradox: why participants without semantic or lexical impairments commit semantic errors. Our findings suggest that it is the interaction between the STM memory load, semantic blocking, and recursive use of the same words that leads to the emergence of notable number of semantic errors; we found a similar pattern when HF-color words sequences had to be produced. The other part of the STEPS paradox involves the absence of phonemic errors in producing numbers when phonemic errors are the default errors in these PWA. In the study that we will now describe, we consider the hypothesis that it is high frequency that leads to a reduction in phonemic errors. Specifically, this study will aim to verify the role that word frequency plays in the production of phonemic errors in STEPS.

According to several studies (e.g., Kittredge et al., 2008; Shallice et al., 2000) lexical frequency plays a relevant role in the production of phonemic errors: they appear more in low-frequency than in high-frequency words. Here we have argued that the absence of phonemic errors in the production of multi-digit number words and HF-color words in Study 2a is due to the very high-frequency of these stimuli. In other words, that the absence of phonemic errors in number words is not due to the existence of a different mechanism for these stimuli in the phonological output buffer, as claimed by the BBH, but rather to these words having far higher frequency than the non-number words to which they are usually compared (see Introduction for detailed arguments on this). To test this hypothesis, we compared the pattern of errors found in Study 2a with HF-color sequences to the pattern with low-frequency (LF) color sequences. We expect that phonemic errors, the default errors in this patient, as observed in Study 1, but which almost disappeared during the production of multi-digit numbers and HF-color sequences (Study 2a), will reappear (in Study 2b, here) when the lexical frequency of the stimuli decreases, that is, during the production of LF-color sequences.

5.1. Materials

For this study we created a list of 60 LF-color sequences, as in the previous study: each LF-color was exclusively associated with a HF-color and a digit number. Consequently, each LF-color sequence was equivalent to a given HF-color sequence and a multi-digit number from the previous study (e.g., *lila - malva - beis* ["lilac - mallow - beige"] is the sequence equivalent to 236 and *rojo - verde - gris* ["red - green - grey"]).

The linguistic properties of LF-color sequences are reported in Table 4; these color words had significatively less lexical and lemma frequency than the HF colors [$t(13.96) = 8.16$, $p < .001$ and $t(14.58) = 8.04$, $p < .001$], while they were both matched in length (number of phonemes, letters, and syllables; all $ps > .99$). Additionally, the sequences comprising them also differed in terms of lexical and lemma frequency [$t(97.73) = 41.46$, $p < .001$ and $t(113.74) = 38.29$, $p < .001$], while being statistically similar in length (all $ps > .82$).

5.2. Procedure

The procedure in this study was the same as in Study 2a regarding the color tasks (repetition, naming, and reading) with one exception: in the repetition task ML was only presented with length 2 sequences as she had complained of too

great a load and that it was impossible for her to recall things in longer sequences. As in Study 2a, before running the sequences we asked both participants to name the color patches and read and repeat these color words in isolation, with the aim of testing their knowledge of these colors. Whereas DNR performed without errors, ML made many errors in the naming of LF-colors; for instance, she used "lilac" both for "lilac" and "mallow", and "brown" for both "beige" and "ochre". Although she was corrected in order to avoid errors in the sequence task, we acknowledge that this data from the naming task may show an artificially increased number of semantic errors, and thus should be taken with caution (note that this problem in identifying the color does not apply to the reading and repetition tasks).

5.3. *Specific data analyses*

In this study, we performed the same analyses as those described in Study 2a. However, since numbers and HF colors showed roughly similar patterns, and we were interested in the role of word frequency, we now compared the data obtained in the production tasks with HF-color sequences to the data for the LF-color sequences.

5.4. *Results*

In this section, we present the most notable results of the study; those interested in a more detailed report can refer to the Supplementary Materials (also available at our OSF: https://osf.io/5v7nr/). A comprehensive overview of errors produced in the three tasks by both patients is provided in Table 5. Fig. 3 illustrates the specific error patterns for each task and stimulus type, while Fig. 5 visually represents performance in relation to sequence length.

5.4.1. *Patient DNR*

5.4.1.1. General exploration of STEPS. Non-significant differences in the number of phonemic and semantic errors in the production of LF colors in repetition and naming (both $\Delta s \leq 5$ errors) (both $Z \leq 1.53$, $ps \geq .19$) were found. However, in reading, DNR produced significatively more phonemic errors than semantic ones (41 *vs* 7) ($Z = 5.83$, $ps < .001$) (see Fig. 3).

5.4.1.2. LF colors versus HF colors. DNR showed similar accuracy with HF and LF words in the repetition ($\Delta = 46.67\%$ accuracy) [$\chi^2(1) = .55$, $p = .46$] and naming tasks ($\Delta = 3.3\%$ accuracy) [$\chi^2(1) = .18$, $p = .67$]. In reading, ML's performance was lower for LF colors ($\Delta = 20$; 33.33%) [$\chi^2(1) = 13.89$, $p < .001$]. Regarding semantic errors, these were significantly more frequent for HF colors than for LF colors in naming and reading (both $\Delta s \geq 12$ semantic errors) [both $\chi^2(1) \geq 7.07$, $ps \leq .008$], but not in repetition ($\Delta = 3$ errors) [$\chi^2(1) = .31$, $p = .58$]. Alternatively, more phonemic errors were found with LF colors than with HF colors across all tasks (all $\Delta s > 14$ phonemic errors) [all $\chi^2(1) \geq 10.21$, $ps \leq .001$] (see also Table 5 for a reference).

5.4.1.3. Length effects. DNR's performance decreased with the length of LF-color sequences [all $\chi^2(2) \geq 9.45$, $p \leq .009$]. Although there was a tendency toward an increase in semantic errors with length, the effect was significant in the repetition task [$\chi^2(2) = 16.78$, $p < .001$] but non-significant in the naming and reading tasks [$\chi^2(2) = 3.17$, $p = .21$ and $\chi^2(2) = 3.71$, $p = .16$, respectively]. Similarly, phonemic errors showed a tendency to increase with sequence length in all tasks; however, this effect was significant only in the reading task [$\chi^2(2) = 8.17$, $p = .017$] while no significant effects were observed in the repetition and naming tasks [$\chi^2(2) = 4$, $p = .14$ and $\chi^2(2) = 5.28$, $p = .071$, respectively].

Finally, when the number of semantic and phonemic errors were compared at each sequence length by task, no significant differences were found for any length in repetition (all $\Delta s \leq 4$) (all $Z \leq 1.90$, $ps \geq .11$) or naming (all $\Delta s \leq 5$) (all $Z \leq 1.15$, $ps \geq .39$). However, in the reading of LF-color words, phonemic errors were more frequent than semantic ones at length 2 ($\Delta = 9$) ($Z = 3$, $p = .004$), length 3 ($\Delta = 13$) ($Z = 3.61$, $p < .001$), and length 4 ($\Delta = 12$) ($Z = 3.46$, $p < .001$) (see Fig. 5).

5.4.2. *Patient ML*

5.4.2.1. General exploration of STEPS. In repetition, more phonemic than semantic errors were found (5 *vs* 0) ($Z = 2.45$, $p = .031$). The pattern reversed for naming, with more semantic than phonemic errors being observed (50 *vs* 23) ($Z = 4.56$, $p < .001$). In reading, errors were scarce and the advantage of phonemic errors over semantic errors (3 *vs* 0) did not reach significance ($Z = 1.73$, $p = .25$).

As mentioned above, we firmly believe that most of the semantic errors in naming were due to ML not knowing the correct name of some color patches. Therefore, although we report the results and analysis of the naming task, they should be interpreted with caution, as this factor may have artificially inflated the number of semantic errors.

5.4.2.2. LF colors versus HF colors. In repetition, only LF-color sequences of length 2 were administered due to ML's difficulties in performing the task. We compared the unequal and small sample sizes, such as the 20 items here against the 60 items in the HF colors condition, by using Fisher's exact tests (see Mehrotra et al., 2003). ML produced more correct responses with HF than LF colors, but the difference did not reach the significance level (18/60 *vs* 2/20 correct responses, 30% *vs* 10%; Fisher's exact test, $p = .083$). However, when considering only length 2, ML produced significantly more correct responses with HF colors than LF colors (13 *vs* 2; Fisher's exact test, $p < .001$). ML made more semantic errors with HF-color words than with LF-color words, both overall [40/60 (67%) *vs* 0/20 (0%), respectively; Fisher's exact test, $p < .001$] and when analyzing only length-2 sequences [7 *vs* 0, respectively; Fisher's exact test, $p = .008$]. Conversely, ML made proportionally more phonemic errors with LF-color sequences. Although this difference was not significant when considering all stimuli [HF: 6/60 (10%) *vs* LF: 6/20 (30%); Fisher's exact test, $p = .064$], it became significant when focusing solely on length-2 sequences [HF: 0 *vs* LF: 6; Fisher's exact test, $p = .02$].

In the naming task, where she was tested with all the material, ML showed a very high number of errors with LF colors, this likely due to confusions with the names of these. This led to better performance with HF colors ($\Delta = 85\%$ accuracy) [$\chi^2(1) = 87.31$, $p < .001$]. Semantic errors and phonemic errors were significantly more frequent in LF- than HF-color

sequences [$\Delta = 45$ semantic; $\chi^2(1) = 67.97$, $p < .001$ and $\Delta = 22$ phonemic; $\chi^2(1) = 25.21$, $p < .001$]. Finally, her performance in reading, where she was also tested on all the materials, was almost error-free, and thus no significant differences were found between HF and LF color sequences in the number of correct responses ($\Delta = 1.67\%$) [$\chi^2(1) \approx 0$, $p > .99$], semantic errors ($\Delta = 2$) [$\chi^2(1) = .51$, $p = .48$], or phonemic errors ($\Delta = 3$) [$\chi^2(1) = 1.37$, $p = .24$].

5.4.2.3. Length effects. In repetition, analyses of length effects and comparisons between semantic and phonemic errors were not possible due to the sole assessment of length 2 sequences. In reading, these analyses were not conducted due to ceiling effects. In LF color naming, no length effects were observed in correct responses [$\chi^2(2) = 2$, $p = .37$], nor did the number of semantic and phonemic errors vary substantially with length [$\chi^2(2) = 1.75$, $p = .42$ and $\chi^2(2) = .17$, $p = .92$, respectively].

Finally, when comparing semantic to phonemic errors across different lengths in the naming task, we observed an increasing difference, with more semantic errors as the LF color sequence length increased: length 2 ($\Delta = 8$) ($Z = 2.53$, $p = .021$), length 3 ($\Delta = 9$) ($Z = 2.71$, $p = .012$), and length 4 ($\Delta = 10$) ($Z = 2.67$, $p = .013$).

5.5. *Discussion*

In this study we began with the hypothesis postulated in Study 2a: that the interaction of the memory demands and semantic blocking is responsible for the emergence of semantic errors, whereas the high frequency of number (and color) words has a protective effect against the production of phonemic errors. We tested whether a reduction in the lexical frequency of the semantically-related stimuli inserted in sequences is linked to a re-emergence of the phonemic errors that were absent when producing HF-color sequences and multi-digit numbers.

The results showed that DNR produced a similar number of phonemic and semantic errors in both naming and repetition tasks; she eventually came to produce more phonemic than semantic errors during the reading task. ML, as expected, due to her impairment, had a good overall performance in reading. She made many semantic errors in naming that may have been caused by her ignorance of the names of some LF-colors, but, interestingly, phonemic errors were also abundant. In repetition, despite her difficulties, neologism and phonemic errors were within the most frequent error types, whereas no semantic errors were made.

The main finding of this study is that phonemic errors, almost absent in the production tasks with HF-color sequences and multi-digit numbers, re-emerge while producing LF-color sequences, these results being consistent when comparing both types of color sequences. DNR and ML produced significantly more phonemic errors with LF-colors than with HF colors in all tasks except the reading task of ML, in which she had almost perfect accuracy. However, semantic errors were observed in the tasks, which suggests that the mechanisms that cause them are still at play. So, it seems that the HF of number words is what avoids the presence of phonemic errors in patients with phonological impairments. The uneven comparison between non-number and number words in terms of frequency seems to be one of the factors behind STEPS here.

6. General Discussion

In this research the causes of STEPS were explored. We have argued that the specificity of this phenomenon, that is, more phonemic errors in words but more semantic errors when producing number words, casts doubt on the lesion-process-symptom mapping approach, for several reasons. From a cognitive perspective, the categorial specificity of both types of errors is unexpected; on one hand, given that phonemic errors have a postlexical origin, they should appear in the production of any type of word, both non-number and number words; on the other hand, in that there is no evidence of semantic impairments in the quantity system of most of these patients or in their lexical representations, the production of semantic errors with number words is unexpected. From a neuroanatomical perspective, it is also difficult to explain why lesions mainly affecting dorsal areas such as the STG or the SMG, which it has been suggested are causally involved in the production of phonemic errors (e.g., McKinnon et al., 2018; Mirman et al., 2015; Ramoo et al., 2021; Schwartz & Dell, 2016, pp. 701–715; Stark et al., 2019), do not induce these errors with numbers words but, rather, lead to semantic errors, which are usually associated with ventral areas.

Despite STEPS being a relatively common behavior in PWA, it was Dotan and Friedmann (2015) who brought this phenomenon into focus, supporting and extending the hypothesis of Cohen et al. (1997). They proposed the BBH, which posits that number words are building blocks of multi-digits in the same way that phonemes are building blocks of words. The model assumes that number words and phonemes are represented in dedicated stores, within the phonological output buffer, to form more complex stimuli (respectively, multi-digits and words), and this might explain how a lesion to this level causes the substitution of number words (apparently lexical or semantic errors) in multi-digits, and the substitution of phonemes (phonemic errors) in words (Dotan & Friedmann, 2015; Cohen et al., 1997). Unfortunately, without further specifications, the BBH encounters difficulties in explaining the presence of STEPS in patients with impairment of the phonological input buffer (i.e., people with conduction aphasia of the repetition variety), as has been described in García-Orza et al. (2020, but see Fischer-Baum et al., 2018). Additionally, by attributing grammatical properties to a phonological device, the phonological output buffer, but giving no role to frequency, the model loses plausibility (see García-Orza et al., 2020; Ochtrup et al., 2013).

In this study we have explored an alternative account: the hypothesis that STEPS is a consequence of the defective experimental control of factors such as memory load, word frequency, semantic context, and task demands when comparing the production of (non-number) words versus number words. To this end, we analyzed the performance of two participants with conduction aphasia, one of the repetition and one of the reproduction variety, both of whom had shown STEPS previously.

The results of these studies are clear. In Study 1 we explored whether memory load affected word repetition by favoring formal and semantic errors. We asked our two participants to repeat non-number word pairs, as a means of creating the higher demand number words make when compared to isolated non-number words. Neither DNR nor ML showed any significant increase in the number of semantic errors or a reduction in phonemic errors (see Fig. 3). Only a slight increase in formal errors was observed (significant for words in position 2 in ML) (see Fig. 4), and although this may provide moderate support for the Continuum Hypothesis, which considers that load will turn phonemic errors into formal and semantic ones (Gold & Kertesz, 2001; Martin et al., 1994, 1996; Martin & Saffran, 1992), our findings also indicate that simply increasing load is not the cause of STEPS effects. It seems that additional factors are needed to produce these.

Studies 2a & 2b provide a clearer view on the mechanisms underlying STEPS. By combining a stronger memory load, more akin to that of multi-digits (sequences of 2–4 words), and using repeatedly HF-words from a single semantic category (colors), a similar pattern to that with numbers appeared in our patients (Study 2a). Additionally, the pattern of semantic and phonemic errors was mixed when sequences of LF-color words were presented instead (Study 2b).

It is also interesting to note that the similarities in the pattern of errors between numbers and HF-color sequences is consistent with the locus of the cognitive impairment of each patient. For patient DNR, impaired in her POB, semantic errors were dominant in the naming, repetition and reading tasks with both numbers and HF colors. In the case of ML, the PIB patient, there were more semantic errors in the repetition task with both stimuli. As expected in a PIB patient, errors were quite low in both naming and, especially, in reading; however, even in the former a trend showing more semantic than phonemic errors was observed, both with HF colors and numbers (the difference was only significant in the naming task with numbers). In general, the pattern with HF-color sequences and multi-digits was similar, and subtle difference between stimuli might have had its origin in the fact that, as shown in both patients, errors with multi-digit numbers tend to be more frequent than those with colors, in that the former sequences were slightly longer than the latter (see materials section in Studies 2a & 2b).

From these results, we argue that the STEPS found in different PWA arises as a consequence of a cumulus of small differences in the conditions employed when evaluating number production in the clinical and experimental setting. We list these here, but see the Introduction for a more detailed analysis of these differences. (a) Number word production is assessed using homogenous lists composed of number words that share semantic meaning, whereas non-number word lists comprise words from different semantic categories (heterogenous). Evidence of semantically-homogenous lists increasing naming times has been reported in several studies using the continuous naming paradigm (e.g., Biegler et al., 2008; Damian et al., 2001; Howard et al., 2006; Kroll & Stewart, 1994). (b) Multi-digits are compound words composed of a limited set of lexemes (e.g., one, two … hundreds, thousands) that are presented repeatedly during the evaluation task. Studies using the cyclic naming paradigm have shown that the continuous repetition of a group of semantically-related items increases activation at the semantic and/or lexical levels, creating more competition and causing a slowdown in naming times and leading to the incorrect selection of semantic competitors (e.g., Belke, 2017; Harvey et al., 2019; Oppenheim et al., 2010; Schnurr et al., 2006). (c) At the same time, since the production of multi-digit words, compared to that of single words, usually involves producing longer phonological sequences, a strong difference in phonological demands is established. Stronger demands in the phonological storage devices may produce a decay of the whole phonological trace, which will prevent this trace from guiding the choice of the correct representation among the lexical competitors, thus favoring lexical and semantic errors (i.e., other number words) (Dell et al., 1997; Gold & Kertesz, 2001; Martin et al., 1994, 1996; Martin & Saffran, 1992). These three differences seem to account for the production of semantic errors in number words, despite no difficulties in number semantics being observed in our patients. A fourth difference may explain the absence of phonological errors in multi-digits production, despite the existence of a phonological deficit in our patients. (d) Multi-digit number production, as seen in our sequences of HF colors, involves selecting, sequencing and storing phonemes that belong to very high-frequency and short lexemes. Research on people with conduction aphasia has shown that phonemic errors increase as frequency decreases, or in other words, that very high lexical frequency protects these people from committing phonemic errors (e.g., Shallice et al., 2000), whereas, as highlighted in Study 2b, reducing this frequency increases phonemic errors.

The present results have clear implications from a number of perspectives. At the cognitive level, it seems that the complexities of the BBH are not needed to explain STEPS effects. From functional anatomic models of language production, which assign phonemic errors to dorsal lesions and semantic errors to ventral damage, the existence of PWA showing speech errors that vary so consistently depending (apparently) on word category requires an explanation. In what follows we expand on these and other issues.

6.1. *Implications for cognitive models of word production*

Classical models of speech production encounter difficulties in explaining STEPS, because phonemic errors should be pervasive and affect the production of all types of words, not respecting any specific categories. Moreover, semantic errors should not arise when there are no difficulties in the number semantic system or at the lexical level. To account for this phenomenon the BBH was proposed, in which numbers are considered to be a special category of stimuli (Cohen et al., 1997; Dotan & Friedmann, 2015) that would be stored as pre-assembled units in dedicated stores within the phonological output buffer. Then, when these stores are damaged, items in each store are substituted by other elements from the same store, so when producing numbers, number words are substituted by another number word, whereas phonemes are substituted by another phonemes.

In order to provide support for the BBH, Dotan and Friedmann (2015) successfully showed that the effect (substitutions but not phonemic errors) occurred also in other stimuli that can be considered parts of words, such as letters, function words and morphological affixes. More interestingly, as the effect was sometimes elusive, they also proposed that the effect would occur only when these stimuli worked as building blocks. So, in the case of numbers words, when they are part of a multi-digit number they will work as building blocks, but not when they are presented as parts of sentences or words in which the number loses (part of) its numerical meaning, as in the name of the Hebrew city "Be'er Sheva", which includes the word "sheva" that means "seven" (Dotan & Friedmann, 2015); similarly, in the case of function words, when they are part of a sentence they will work as building blocks but when they are presented in a list of function words they will lack syntactic context, and hence will not have a building-block function. Results in that study indicated that complete substitutions of number and function words, respectively, were observed in the first cases, thus confirming their hypothesis, and they pointed out the relevance of the role in which a word appears: "when number words and function words appear in the relevant role (number words with numeric meaning and function words with syntactic role), they are produced with semantic rather than phonological errors. Conversely, when the number and function words were deprived of their role by changing the task and context, they were produced with many phonological errors, and without semantic errors, just like content words" (Dotan & Friedmann, 2015, p. 338).

The BBH accounts for most of Dotan and Friedmann's (2015) data. However, as we pointed out in the Introduction, it faces conceptual issues and cannot cope with the presence of STEPS in patients without damage to the POB (García-Orza et al., 2020; Ochtrup et al., 2013). More importantly, the model cannot easily explain the results reported in our present research. It is not initially predicted that HF-color would show the same pattern observed with multi-digit numbers, where there were more semantic than phonemic errors. Color words cannot be considered building blocks as they do not form part of any "productive processes", so it is implausible to think that they are pre-assembled units in a dedicated store for colors in the POB. To overcome this argument, one interesting idea is to consider that any word can be deemed a building block if it is integrated into sequences; but then the idea of "dedicated stores neuroanatomically implemented for different word categories" loses its meaning, and consequently the BBH loses its nature. Likewise, it would be problematic to explain why LF colors, which would also be building blocks when embedded in sequences, showed a mix of semantic and phonemic errors instead of behaving like HF colors. To make things yet more complicated for the model, the explicit denial of a role for frequency at the POB level (see Cohen et al., 1997; Dotan & Friedmann, 2015) limits possible accounts of this difference.

6.2. *Explaining STEPS as an experimental artifact*

If we reject the BBH, how can we explain our participants' behavior in the framework of models of word production? Candidate models need to be flexible enough to account for the coexistence of semantic and phonological errors in word production, and in this context, we think connectionist models provide an adequate framework. For reasons of space, we will limit our explanation to the naming and repetition tasks using the dual route model of repetition proposed by Dell and colleagues (Nozari et al., 2010; Nozari & Dell, 2013; see also Ueno, Saito, Rogers, & Lambon Ralph, 2011, for other computational models implementing two routes), and thus we will not focus on other models which, with some adjustments, might also explain our data.

The model is composed of a lexical route, which would be involved in the naming and repetition of known words, and a non-lexical route, which simply maps input phonology to output phonology; this would be compulsory for the repetition of unknown words and can be also involved in the repetition of words. The lexical route includes three layers: one that stores semantic representations, one that stores word (abstract) representations, and one that stores phonemes (see, e.g., Nozari et al., 2010, for a detailed explanation). When a picture to-be-named is presented to a subject, a set of semantic features are activated and these spread their activation to the corresponding representations at the word level; this activation spreads back to the semantic level and forward to the phoneme level, in which the phonemes corresponding to the most active words should receive more activation. After a number of cycles, the correct word is selected at the word level, and then, finally, after a number of cycles of interaction between the lexical and phoneme level, the corresponding phonemes are selected. Compared to naming, repetition involves additional processes: through the lexical route, input processes will activate the to-be-repeated word and (optionally) its semantic features, then production will follow from the (semantic and) word level to the phonemic (output) level, as in naming; importantly, repetition may also be carried out through the non-lexical route, because once input phonology is processed it is stored temporarily in a buffer (termed by Nozari & Dell, 2013, as "temporary non-lexical node", but usually known as phonological input buffer, PIB) and then mapped into the output phonology (the phoneme level for the model, but also known as phonological output buffer, POB). Something that is not wholly resolved by the model is when one route or another is used. Nozari et al. (2010) suggested that there is variability in healthy persons regarding whether they rely more on one or the other route, but the nature of the stimuli plays a fundamental role, for instance, pseudowords can only be repeated through the non-lexical route. In the case of PWA, the location of brain damage seems to play a role too. For example, it seems that when the non-lexical route is damaged, if it cannot retain the phonological input information for long in the phonological memory due to anormal fast decay, then PWA rely more on the lexical route for repetition. However, when this latter route is the one that is impaired, for instance, when access to meaning is affected, PWA rely more on the non-lexical route (Nozari & Dell, 2013).

Using this framework, which it is sensitive to the four factors we consider to be involved in STEPS, namely, semantic relatedness, cyclic naming, frequency and memory load, we will try to explain the STEPS effect. First, both of our patients showed difficulties in manipulating phonological information, but for different reasons; while ML had difficulties in

temporarily storing input phonological information (PIB impairment), DNR had issues to maintain the output phonological information (POB impairment). For instance, when the repetition of words is demanded, for ML it is hard to maintain the sequence of phonemes in the input buffer, whereas for DNR it is hard to maintain it in the output buffer. These difficulties lead to omissions, substitutions, additions, transpositions etc., as observed when our PWA were asked to produce words (see PWA's clinical evaluation and Study 1). For DNR the impairment in the output phoneme level would also explain the commission of the same errors in naming and reading, whereas the failure to maintain phonological information in the temporary non-lexical node would cause ML to produce errors mainly in the repetition task.[5]

We argue that when presented with sequences of multi-digit numbers and HF colors repeatedly, the dynamics of the system change, as accounts of cyclic naming and continuous naming tasks suggest (e.g., Belke, 2017; Harvey et al., 2019; Oppenheim et al., 2010). The existence of semantic relatedness between the words in sequences produces, in the lexical route, the activation of shared numerical features at the semantic level, which activates semantically-related representations at the word level; moreover, the repeated naming of a small set of number words increases this activation at the semantic and word levels. These two factors then create higher levels of interference between semantically-related representations that temporally give rise to semantic errors in the process of semantics-to-word mapping. The absence of semantic errors in our Study 1, but its very notable presence when using numbers and colors in Study 2a, provides strong support for the role of semantic relatedness. We also consider that semantic errors benefit from the absence of a strong phonological trace, so when phonological traces of the word are effectively maintained (e.g., in repetition or reading), semantic errors would be reduced, because the phoneme information via backward activation will help in the process of word selection by discarding semantic competitors. Despite this, under high memory load conditions, as in multi-digit number words and in longer color-word sequences, the high memory demands on an impaired phonological system reduce the chances of retaining phonological information that may help in the word selection process, leaving the way open for the incorrect selection of semantically-related competitors. The general impact of length on increasing semantic errors, as seen in Study 2a, is in line with this claim.

Whereas this account would explain the presence of semantic errors in multi-digit numbers and HF colors, the absence of phonemic errors in these words remains to be explained. As we saw in the contrast between HF and LF colors in Study 2b, frequency plays a relevant role here. Word-level representations from super high-frequency stimuli, as is the case of numbers, HF-color words and function words, activate more strongly the phoneme level, to the extent that this allows the proper storing of those representations until the articulatory process is executed. However, when frequency is lower, as in LF-color words, phonemes vanish more easily due to phonological impairment (faster decay), and phonemic errors are then observed (see Kittredge et al., 2008). Our manipulation of frequency in color-word sequences (Studies 2a & 2b) provides empirical support for this account.

How is it, then, that other studies have disregarded frequency as playing a relevant role in STEPS? Cohen et al. (1997, Table 10) compared the number of phonemic errors in number words to those in non-number words of the same length that had, at least, the same frequency. They found considerably more phonemic errors in non-number words, with errors being almost absent in number words (see also Bachoud-Lévi & Dupoux, 2003); from this they concluded that frequency was not related to the production of phonemic errors. However, as we noted in the Introduction, frequency is underrepresented in these studies because, among other factors, they tend to consider lexical frequency of the whole word rather than the lemma frequency of each lexeme in the multidigit. The importance of using lemma frequency is underlined by Cohen et al. in their observations of patient behavior: "with numerals comprising several words, the patient resorted to a word-by-word reading strategy" and that when "French compound number words, such as 'dix-neuf' or 'quatre-vingt' were printed with no dash, … the patient treated the two components as independent words" (Cohen et al., 1997, p. 1048). Due to the compositionality of the number system, it seems that lemma counts for each lexeme should be taken into account when exploring the role of frequency in number production.

So, arguments exist to suggest that previous studies underreported the frequency of number words and made unbalanced comparisons with words. By contrast, our findings in Study 2b correctly address this, and strongly confirm the role that frequency plays in the presence/absence of phonemic errors in these patients.

An additional aspect that merits attention in this discussion is the ability of the BBH to explain contextual effects (Dotan & Friedmann, 2015), that is, the finding that the same words may suffer from semantic or phonemic errors depending on the role of the word. We firmly believe that these effects may also be accounted for by differences in evaluation conditions within the framework of connectionist models. Regarding the contextual effects reported with numbers by Dotan and Friedmann (2015), discussed above, they hypothesized that words including numbers, such as "Be'er Sheva", in which these words have less salient numerical meaning, would be produced with more phonemic and less semantic errors than single digits ("seven") or digits presented in sentences with numerical meaning (e.g., "the man ate seven apples"). Their results confirmed this word-role effect; however, we believe that it is in fact simple to explain if we consider that tasks make different demands on the cognitive devices involved in word production. First, names like "Be'er Sheva" have a meaning which, as Dotan and

[5] A small but crucial variation has been included to the original model to explain STEPS, because according to the original interactive dual-route model, an impairment in the temporary non-lexical store should not affect the repetition of words. The model assumes that known words will directly activate the word level without requiring phonemic input storage. Whereas this can be for short words, it seems reasonable to think that this device is needed in the recognition of longer words, multi-digit numbers and sequences of words; in other words, that this device is involved even in comprehension, but depending on the stimuli length. On this view, ML's errors in repeating words can be better explained.

Friedmann (2015) indicate, guide people loosely to a numerical semantic representation, and hence semantic interference between number representations is not an issue under these conditions, and this precludes the appearance of semantic errors. Second, lexical representations of some compound words, such as brand names and idioms (other materials that have been used in these studies) are usually longer, and probably less frequent, than numbers, which makes them more prone to phonemic errors.

Let us now turn to the findings with function words. Dotan and Friedmann (2015) asked participants to read function words embedded in sentences (i.e., with a syntactic and then a building-block function) or presented in a list (i.e., without a building-block context). Their focus was on the comparison of phonological errors (they included phonemic and formal errors), and as predicted by the word-role hypothesis, they found that these errors were significantly greater in number in the list condition. Additionally, they compared the number of errors in function words presented in lists with the number of errors in content words of similar phoneme length, finding similar rates of phonological errors. From our hypothesis concerning the several factors that affect the production of errors, it is difficult to explain, given that the words are the same, why participants committed fewer phonological errors in reading sentences than in reading lists. We acknowledge here that the hypothesis is speculative, and that function words in lists are sometimes interpreted as pseudowords, thus favoring a sublexical reading and the production of phonological errors, whereas this would not happen in the context of sentences, where the reading for meaning activates structural and semantic predictions that facilitate the recovery of the correct word from the lexicon, ensuring a strong activation toward the phonological level. Additionally, Dotan and Friedmann (2015) compared the number of semantic errors (in fact, the substitution of one function word by another) in both tasks and found that they were scarce and comparable in all participants but one, who showed more substitutions when presented in sentences. According to the model, more semantic errors should have been observed in a syntactic context, but the BBH has no explanation for the absence of these errors. Our view, on the contrary, is that the low demand of reading single sentences in a very predictive context for syntactic elements is sufficient to account for this. With these arguments, we are not claiming that function words cannot suffer semantic (substitution) errors, provided they are presented in circumstances with high memory load and in a context that involves the production of the same function words repeatedly. Semantic errors, as in any other word, are possible. Further studies are needed to provide evidence on this.

All in all, our claim that STEPS is an experimental artifact has involved an account of the experimental findings reported in previous studies and our own experimental findings here. It seems that word production is subject to different constraints that may modify the nature of errors as a consequence of the complex dynamics established between the cognitive systems involved in word production: the damage these systems suffer, the psycholinguistic properties of the to-be-produced words, and the evaluation conditions.

6.3. *Implications for anatomic-functional models of word production*

Although not the main focus of the present research, our results provide relevant information on our general understanding of the brain mechanisms that produce speech errors in PWA. According to a generally accepted lesion-process-symptom view, semantic errors are the consequence of lexical and/or semantic impairments caused by damage to ventral areas, whereas phonemic errors are post-lexical in nature and the consequence of damage to dorsal areas (e.g., Berthier et al., 2018; Fridriksson et al., 2009; Hickok & Poeppel, 2007; Hillis, 2001; McKinnon et al., 2018; Mirman et al., 2015; Ramoo et al., 2021). This correspondence between areas and errors, which forms the basis of extant neurofunctional models of word production (e.g., Hickok & Poeppel, 2007; Schwartz, Faseyitan, Kim, & Coslett, 2012; Ueno, Saito, Rogers, & Lambon Ralph, 2011), is supported by several studies with PWA. These studies found that semantic errors in naming correlate with damage to the mid-to-anterior temporal lobe, whereas phonological errors in word production are associated with damage to posterior superior temporal and frontoparietal regions, probably with a special role for the supramarginal gyrus (Buchsbaum et al., 2011; Cloutman et al., 2009; Fridriksson et al., 2016; Mirman et al., 2015; Schwartz, Faseyitan, Kim, & Coslett, 2012; Stark et al., 2019). In relation to the dual stream model (Hickok & Poeppel, 2007), when one looks at the tracts involved in the production of errors, these also link semantic errors to axonal loss in the ventral pathway (inferior longitudinal fasciculus, ILF), whereas phonemic errors correlate with axonal loss in the dorsal pathway (superior longitudinal fasciculus, SLF) (e.g., McKinnon et al., 2018). These studies show that areas involved in these two types of errors do not overlap, and thus support the idea that they are independent. In other words, patients with mainly phonological errors will have impairments in dorsal areas, whereas those with mainly semantic errors will have impairments in ventral areas; finally, a patient who makes semantic and phonemic errors indistinctly in different words is predicted to have damage in both dorsal and ventral areas.

This view, however, is challenged by STEPS because the presence of semantic errors limited to some word categories and phonemic errors limited to others is not predicted. The phonemic errors that DNR and ML usually commit with words, and their clinical and neuroimage evaluations, suggest that both patients have mainly an impairment in the dorsal route. However, to explain their semantic errors they should have additional damage to ventral areas, yet if this were the case, these semantic errors would not be limited to certain categories (numbers, affixes, function words, HF colors) but would appear in any type of word, irrespective of its category. A conceptual deficit limited to number and color processing may explain semantic errors limited to these categories, but our participants' performance, as shown in the number comparison task, and in naming single Arabic digits and colors, leads us to discount this possibility (see also García-Orza et al., 2020). So, for these models it remains to be explained how brain damage causes semantic errors in numbers and HF colors but not in other words.

We firmly believe that considering (a) the nature of the tasks and their demands together with the psycholinguistic properties of the words involved in oral production, and (b) a view of word production in aphasia as the fruit of the joint work of both streams, makes it possible for the existing consensus on lesion-process-symptom mapping to accommodate STEPS. In fact, most anatomic-functional models, although not always explicitly, are sensitive to task demands, and thus they posit differences between tasks (e.g., naming involves more weight in the ventral route, whereas repetition in the dorsal route), and confer a moderated role on word properties such as lexical status and lexical frequency, in predicting the involvement of these routes (e.g., Hickok, 2014; Schwartz, Faseyitan, Kim, & Coslett, 2012).

Moreover, evidence on the interaction between routes has also been described in the literature. For instance, with impairments of the dorsal stream, it has been found that the impaired as well as the preserved streams contributed to production, and this led to the conclusion that even when impaired "the dorsal system is essential for accurate phonological encoding of meaningful speech" (Schwartz, Faseyitan, Kim, & Coslett, 2012, p. 3809). Similarly, it has been suggested that *conduite d'approche* in PWA with damage to the dorsal route benefits from a preserved ventral pathway (see Ueno & Lambon Ralph, 2013). So, as other authors have noted (López-Barroso & de Diego-Balaguer, 2017), probably thanks to the redundancy that exists in the connectivity between language areas, compensatory functions between the dorsal and the ventral stream occur, giving rise to a final production that involves the contribution of both streams. This would be especially evident in STEPS as a whole, but also in the naming of LF colors, where we found many instances of semantic errors combined with phonemic ones (e.g., "mawllo" [*mallow*] for "beige").

As our data show, task demands as well as word properties may change the balance between routes. When the dorsal route is impaired in the auditory posterior part of the route (e.g., as in ML) errors would only arise in the repetition tasks, and would not affect naming or reading. However, when the impairment affects areas like the supramarginal gyrus and postcentral sulcus, this affects the production of words in any task (repetition, reading and naming, as was the case with patient DNR), and the errors would be phonemic in nature (e.g., Schwartz, Faseyitan, Kim, & Coslett, 2012). Specifically, in STEPS during the production of multi-digit numbers, the loading conditions, together with the damage to the dorsal stream that ML and DNR both have, make it impossible to temporarily maintain the phonological properties of the to-be-produced words. Under such circumstances, the preserved ventral stream may take the lead. However, the repeated presentation of multi-digit number words (or other semantically-related categories like colors) generates an overactivation of semantically-related representations that hinders the selection of the correct representation in anterior temporal areas. The selection of the correct lexical representations from the most activated semantically-related candidates does not benefit from phonological clues due to the impairment of the dorsal stream. This impairment affects the input phonological trace that ML has to maintain in the repetition task, and the output DNR needs to employ in repetition, naming and reading, giving rise to different patterns, despite the fact that the two patients have (different) lesions to the dorsal pathway. Finally, once a representation is selected, both the high frequency and short length of the to-be-produced words will allow appropriate selection of the phonological information, and hence a production without phonemic errors, whereas reducing frequency will lead to a mixture of semantic and phonemic errors. This dynamic and plastic trade-off between the dorsal and ventral streams in phonological and semantic tasks occurs not only in damaged brains but also in complex linguistic tasks in healthy subjects, as shown in previous studies (addressing, e.g., backward speech; Torres-Prioris et al., 2020; or new word learning; López-Barroso et al., 2011, 2015).

The account we provide here to explain STEPS can, without substantial modifications, be incorporated into the principal dual route models available in the literature (e.g., Hickock & Poeppel, 2007; Schwartz, Faseyitan, Kim, & Coslett, 2012; Ueno, Saito, Rogers, & Lambon Ralph, 2011). These models should consider that the type of production error in this case is not only a consequence of the main areas impaired, but also depends on task demands and the availability of information. When the non-lexical route (dorsal) is impaired, the ventral is recruited more strongly, and when the ventral stream fails, the non-lexical route is encouraged; but in the end, the outcome of the production process depends on the availability of the phonological or the semantic information that the (altered) streams may provide, and, as this research suggests, this is easily modified by the materials and the evaluation conditions, to the extent that STEPS appears.

7. Conclusions

STEPS is a paradoxical behavior in which PWA makes predominantly phonemic or semantic errors depending, apparently, on word category. Here we have suggested, using experimental data, that STEPS is a consequence of the dynamic interactions between the dorsal and ventral brain structures that support word production, the psycholinguistic properties (frequency) of the to-be-produced stimuli, and the assessment conditions (homogeneous versus heterogeneous; repeated versus non-repeated set of stimuli; single-words versus word-sequences) in persons with phonological deficits. Word production is sensitive to multiple factors, and if we are not aware of this, erroneous conclusions may be drawn. Specifically, STEPS has been used, in our view erroneously, to support the idea of the existence of various peripheral processes for different type of words (numbers, letter names, function words) (e.g., Bachoud-Lévi & Dupoux, 2003; Bencini et al., 2011; Cohen et al., 1997; Dotan & Friedmann, 2015; Fischer-Baum et al., 2018).

Our research with STEPS here should make us aware of the relevance of properly controlling the evaluation conditions and the materials used in clinical experiments. Performance of PWA in classical evaluation tasks, like digit-memory span or spelling tasks, might be affected by some of these factors,

causing worse scores than expected according to people's storing capacity, or their knowledge of letters. Being aware of this will benefit our understanding of the behavior of PWA.

Additionally, our findings, which show an increase in semantic errors with phonological load, align with the view that there is a continuum between moderate phonological impairments in repetition and reading and more severe conditions such as deep dysphasia and deep dyslexia (see Crisp & Lambon Ralph, 2006; Jefferies et al., 2006; Martin & Saffran, 1992; Martin et al., 1994, 1996). Future studies should investigate this population specifically, to determine whether the same mechanisms underlying STEPS are responsible for the semantic errors in more severe (deep) pathologies and the reduction of these errors in less severe ones.

To conclude, we would like to note the value that STEPS may have in the future as a tool to gather evidence on the dynamics of the dorsal and ventral streams, and then in addressing the uncertainty around the interaction between these components in current models of word production.

CRediT authorship contribution statement

Ismael Gutiérrez-Cordero: Writing – review & editing, Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Javier García-Orza:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization.

Data statement

All materials, collected data, and R scripts have been made available to ensure the reproducibility of the analysis and data visualization in the following Open Science Framework repository: https://osf.io/5v7nr/.

Funding

This work was supported by a PhD scholarship provided by the Universidad de Málaga to IGC via the I Plan Propio de Investigación, Transferencia y Divulgación Científica, and a grant from the Junta de Andalucía awarded to JGO (ProyExcel_00744).

Conflict of interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgements

We would like to thank the participants and their relatives for their assistance in this study. We also thank Marcelo Berthier, María José Torres-Prioris and Diana López-Barroso for thoughtful discussions on the theoretical consequences of the data. We also especially thank José Miguel Rodríguez-Santos for his comments and Martina Guandalini for her help in collecting data. Previous versions of this manuscript have benefited from the invaluable feedback of Dror Dotan, Naama Friedmann, and two anonymous reviewers.

Scientific transparency statement

DATA: All raw and processed data supporting this research are publicly available: https://osf.io/5v7nr/

CODE: All analysis code supporting this research is publicly available: https://osf.io/5v7nr/

MATERIALS: Some study materials supporting this research are publicly available, while some are subject to restrictions: https://osf.io/5v7nr/

DESIGN: This article reports, for all studies, how the author(s) determined all sample sizes, all data exclusions, all data inclusion and exclusion criteria, and whether inclusion and exclusion criteria were established prior to data analysis.

PRE-REGISTRATION: No part of the study procedures was pre-registered in a time-stamped, institutional registry prior to the research being conducted. No part of the analysis plans was pre-registered in a time-stamped, institutional registry prior to the research being conducted.

For full details, see the *Scientific Transparency Report* in the supplementary data to the online version of this article.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cortex.2025.02.005.

REFERENCES

Armando, M., Grainger, J., & Dufau, S. (2023). Multi-LEX: A database of multi-word frequencies for French and English. *Behavioral and Research Methods, 55*, 4315–4328. https://doi.org/10.3758/s13428-022-02018-9

Bachoud-Lévi, A. C., & Dupoux, E. (2003). An influence of syntactic and semantic variables on word form retrieval. *Cognitive Neuropsychology, 20*(2), 163–188. https://doi.org/10.1080/02643290242000907

Belke, E. (2017). The role of task-specific response strategies in blocked-cyclic naming. *Frontiers in Psychology, 7*, 1955. https://doi.org/10.3389/fpsyg.2016.01955

Belke, E., Meyer, A. S., & Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology Section A, 58*(4), 667–692. https://doi.org/10.1080/02724980443000142

Bencini, G. M. L., Pozzan, L., Bertella, L., Mori, I., Pignatti, R., Ceriani, F., & Semenza, C. (2011). When two and too don't go together: A selective phonological deficit sparing number words. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 47*(9), 1052–1062. https://doi.org/10.1016/j.cortex.2011.03.013

Berthier, M. L., Torres-Prioris, M. J., López-Barroso, D., Thurnhofer-Hemsi, K., Paredes-Pacheco, J., Roé-Vellvé, N., Alfaro, F., Pertierra, L., & Dávila, G. (2018). Are you a doctor?…

are you a doctor? I'm not a doctor! A reappraisal of mitigated echolalia in aphasia with evaluation of neural correlates and treatment approaches. *Aphasiology, 32*(7), 784–813. https://doi.org/10.1080/02687038.2016.1274875

Biegler, K. A., Crowther, J. E., & Martin, R. C. (2008). Consequences of an inhibition deficit for word production and comprehension: Evidence from the semantic blocking paradigm. *Cognitive Neuropsychology, 25*(4), 493–527. https://doi.org/10.1080/02643290701862316

Buchsbaum, B. R., Baldo, J., Okada, K., Berman, K. F., Dronkers, N., D'Esposito, M., & Hickok, G. (2011). Conduction aphasia, sensory-motor integration, and phonological short-term memory–an aggregate analysis of lesion and fMRI data. *Brain and Language, 119*(3), 119–128. https://doi.org/10.1016/j.bandl.2010.12.001

Caramazza, A., & Hillis, A. E. (1990). Where do semantic errors come from? *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 26*(1), 95–122. https://doi.org/10.1016/s0010-9452(13)80077-9

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

Cloutman, L., Gottesman, R., Chaudhry, P., Davis, C., Kleinman, J. T., Pawlak, M., … Hillis, A. E. (2009). Where (in the brain) do semantic errors come from? *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 45*(5), 641–649. https://doi.org/10.1016/j.cortex.2008.05.013

Cohen, L., Verstichel, P., & Dehaene, S. (1997). Neologistic jargon sparing numbers: A category specific phonological impairment. *Cognitive Neuropsychology, 14*(7), 1029–1061. https://doi.org/10.1080/026432997381349

Crisp, J., & Lambon Ralph, M. A. (2006). Unlocking the nature of the phonological-deep dyslexia continuum: The keys to reading aloud are in phonology and semantics. *Journal of Cognitive Neuroscience, 18*(3), 348–362. https://doi.org/10.1162/089892906775990543

Cuetos, F., & González-Nosti, M. (2009). *BETA: Batería para la Evaluación de los Trastornos Afásicos*. Madrid: EOS.

Damian, M. F., Vigliocco, G., & Levelt, W. J. (2001). Effects of semantic context in the naming of pictures and words. *Cognition, 81*(3), B77–B86. https://doi.org/10.1016/S0010-0277(01)00135-4

Delazer, M., & Bartha, L. (2001). Transcoding and calculation in aphasia. *Aphasiology, 15*(7), 649–679. https://doi.org/10.1080/02687040143000104

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., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104*(1), 123–147. https://doi.org/10.1037/0033-295x.104.1.123

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition, 42*(1–3), 287–314. https://doi.org/10.1016/0010-0277(92)90046-k

Dotan, D., & Friedmann, N. (2019). Separate mechanisms for number reading and word reading: Evidence from selective impairments. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 114*, 176–192. https://doi.org/10.1016/j.cortex.2018.05.010

Dotan, D., & Friedmann, N. (2015). Steps towards understanding the phonological output buffer and its role in the production of numbers, morphemes, and function words. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 63*, 317–351. https://doi.org/10.1016/j.cortex.2014.08.014

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2007). The role of the frequency of constituents in compound words: Evidence from Basque and Spanish. *Psychonomic Bulletin & Review, 14*, 1171–1176. https://doi.org/10.3758/BF03193108

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods, 45*(4), 1246–1258. https://doi.org/10.3758/s13428-013-0326-1

Fischer-Baum, S., Mis, R., & Dial, H. (2018). Word deafness with preserved number word perception. *Cognitive Neuropsychology, 35*(8), 415–429. https://doi.org/10.1080/02643294.2018.1515734

Fridriksson, J., Baker, J. M., & Moser, D. (2009). Cortical mapping of naming errors in aphasia. *Human Brain Mapping, 30*(8), 2487–2498. https://doi.org/10.1002/hbm.20683

Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, & Rorden, C. (2016). Revealing the dual streams of speech processing. *Proceedings of the National Academy of Sciences, 113*(52), 15108–15113. https://doi.org/10.1073/pnas.1614038114

García-Orza, J., Gutiérrez-Cordero, I., & Guandalini, M. (2020). Saying thirteen instead of forty-two but saying lale instead of tale: Is number production special? *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 128*, 281–296. https://doi.org/10.1016/j.cortex.2020.03.020

García-Orza, J., & León-Carrión, J. (2005). Lexical effects in verbal STM: Evidences from a phonological output buffer. *Brain and Language, 95*, 44–45. https://doi.org/10.1016/j.bandl.2005.07.016

Geschwind, N. (1965). Disconnexion syndromes in animals and man (Part II). *Brain: a Journal of Neurology, 88*(3), 585. https://doi.org/10.1093/brain/88.3.585, 585.

Girelli, L., & Delazer, M. (1999). Differential effects of verbal paraphasias on calculation. *Brain and Language, 69*, 361–364.

Gold, B. T., & Kertesz, A. (2001). Phonologically related lexical repetition disorder: A case study. *Brain and Language, 77*, 241–265. https://doi.org/10.1006/brln.2000.2441

Goldrick, M., & Larson, M. (2008). Phonotactic probability influences speech production. *Cognition, 107*(3), 1155–1164. https://doi.org/10.1016/j.cognition.2007.11.009

Goodglass, H., & Wingfield, A. (1997). Word finding deficits in aphasia: Brain-behavior relations and symptomatology. In H. Goodglass (Ed.), *Anomia*. Academic Press.

Gutiérrez-Cordero, I., Torres-Prioris, M. J., & García-Orza, J. (2025). Definition: Conduite d'approche. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior*. https://doi.org/10.1016/j.cortex.2025.02.007

Gvion, A., & Friedmann, N. (2012). Phonological short-term memory in conduction aphasia. *Aphasiology, 26*(3–4), 579–614. https://doi.org/10.1080/02687038.2011.643759

Harvey, D. Y., Traut, H. J., & Middleton, E. L. (2019). Semantic interference in speech error production in a randomized continuous naming task: Evidence from aphasia. *Language, Cognition and Neuroscience, 34*(1), 69–86. https://doi.org/10.1080/23273798.2018.1501500

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language and Cognitive Processes, 29*(1), 2–20. https://doi.org/10.1080/01690965.2013.834370

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews. Neuroscience, 8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hillis, A. E. (2001). The organization of the lexical system. In E. B. Rapp (Ed.), *The handbook of cognitive neuropsychology* (pp. 185–210). Philadelphia: Psychology Press.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition, 100*(3), 464–482. https://doi.org/10.1016/j.cognition.2005.02.006

Jefferies, E., Crisp, J., & Lambon Ralph, M. A. (2006). The impact of phonological or semantic impairment on delayed auditory repetition: Evidence from stroke aphasia and semantic dementia. *Aphasiology, 20*, 963–992. https://doi.org/10.1080/02687030600739398

Kertesz, A. (1982). *Western aphasia Battery (WAB)*. New York: Grune & Stratton.

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology, 25*(4), 463–492. https://doi.org/10.1080/02643290701674851

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

López-Barroso, D., & de Diego-Balaguer, R. (2017). Language learning variability within the dorsal and ventral streams as a cue for compensatory mechanisms in aphasia recovery. *Frontiers in Human Neuroscience, 11*, 476. https://doi.org/10.3389/fnhum.2017.00476

López-Barroso, D., de Diego-Balaguer, R., Cunillera, T., Camara, E., Münte, T. F., & Rodriguez-Fornells, A. (2011). Language learning under working memory constraints correlates with microstructural differences in the ventral language pathway. *Cerebral Cortex, 21*(12), 2742–2750. https://doi.org/10.1093/cercor/bhr064

López-Barroso, D., Ripollés, P., Marco-Pallarés, J., Mohammadi, B., Muente, T. F., Bachoud-Lévi, A. C., & de Diego-Balaguer, R. (2015). Multiple brain networks underpinning word learning from fluent speech revealed by independent component analysis. *Neuroimage, 110*, 182–193. https://doi.org/10.1016/j.neuroimage.2014.12.085

Majerus, S., Attout, L., Artielle, M. A., & Van der Kaa, M. A. (2015). The heterogeneity of verbal short-term memory impairment in aphasia. *Neuropsychologia, 77*, 165–176. https://doi.org/10.1016/j.neuropsychologia.2015.08.010

Marangolo, P., Nasti, M., & Zorzi, M. (2004). Selective impairment for reading numbers and number words: A single case study. *Neuropsychologia, 42*(8), 997–1006. https://doi.org/10.1016/j.neuropsychologia.2004.11.001

Marangolo, P., Piras, F., & Fias, W. (2005). "I can write seven but I can't say it": A case of domain – specific phonological output deficit for numbers. *Neuropsychologia, 43*(8), 1177–1188. https://doi.org/10.1016/j.neuropsychologia.2004.11.001

Martin, R. C., Breedin, S. D., & Damian, M. F. (1999). The relation of phoneme discrimination, lexical access, and short-term memory: A case study and interactive activation account. *Brain and Language, 70*(3), 437–482. https://doi.org/10.1006/brln.1999.2184

Martin, N., Dell, G. S., Saffran, E. M., & Schwartz, M. F. (1994). Origins of paraphasias in deep dysphasia: Testing the consequences of a decay impairment to an interactive spreading activation model of lexical retrieval. *Brain and Language, 47*(4), 609–660. https://doi.org/10.1006/brln.1994.1061

Martin, N., & Saffran, E. M. (1992). A computational account of deep dysphasia: Evidence from a single case study. *Brain and Language, 43*(2), 240–474.

Martin, N., & Saffran, E. M. (1997). Language and auditory-verbal short-term memory impairments: Evidence for common underlying processes. *Cognitive Neuropsychology, 14*(5), 641–682. https://doi.org/10.1080/026432997381402

Martin, N., Saffran, E. M., & Dell, G. S. (1996). Recovery in deep dysphasia: Evidence for a relation between auditory–verbal STM capacity and lexical errors in repetition. *Brain and Language, 52*(1), 83–113. https://doi.org/10.1006/brln.1996.0005

McKinnon, E. T., Fridriksson, J., Basilakos, A., Hickok, G., Hillis, A. E., Spampinato, M. V., … Bonilha, L. (2018). Types of naming errors in chronic post-stroke aphasia are dissociated by dual stream axonal loss. *Scientific Reports, 8*(1), Article 14352. https://doi.org/10.1038/s41598-018-32457-4

Mehrotra, D. V., Chan, I. S., & Berger, R. L. (2003). A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics, 59*(2), 441–450. https://doi.org/10.1111/1541-0420.00051

Messina, G., Denes, G., & Basso, A. (2009). Words and number words transcoding: A retrospective study on 57 aphasic subjects. *Journal of Neurolinguistics, 22*(5), 486–494. https://doi.org/10.1016/j.jneuroling.2009.04.001

Mirman, D., Chen, Q., Zhang, Y., Wang, Z., Faseyitan, O. K., Coslett, H. B., & Schwartz, M. F. (2015). Neural organization of spoken language revealed by lesion–symptom mapping. *Nature Communications, 6*, 6762. https://doi.org/10.1038/ncomms7762

Navarrete, E., Del Prato, P., Peressotti, F., & Mahon, B. Z. (2014). Lexical selection is not by competition: Evidence from the blocked naming paradigm. *Journal of Memory and Language, 76*, 253–272. https://doi.org/10.1016/j.jml.2014.05.003

Nickels, L. (1997). *Spoken word production and its breakdown in aphasia* (1st ed.). Psychology Press. https://doi.org/10.4324/9781315804620

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

Ochtrup, M. T., Rath, D., Klein, E., Krinzinger, H., Willmes, K., & Domahs, F. (2013). Are number words fundamentally different? A qualitative analysis of aphasic errors in word and number word production. *International Journal of Speech & Language Pathology and Audiology, 1*, 12–28. https://doi.org/10.12970/2311-1917.2013.01.01.3

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition, 114*(2), 227–252. https://doi.org/10.1016/j.cognition.2009.09.007

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ramoo, D., Olson, A., & Romani, C. (2021). Repeated attempts, phonetic errors, and syllabifications in a case study: Evidence of impaired transfer from phonology to articulatory planning. *Aphasiology, 35*(4), 485–517. https://doi.org/10.1080/02687038.2021.1881349

Rodriguez, J., & Laganaro, M. (2008). Sparing of country names in the context of phonological impairment. *Neuropsychologia, 46*(7), 2079–2085. https://doi.org/10.1016/j.neuropsychologia.2008.02.009

Salis, C., Kelly, H., & Code, C. (2015). Assessment and treatment of short-term and working memory impairments in stroke aphasia: A practical tutorial. *International Journal of Language & Communication Disorders, 50*(6), 721–736. https://doi.org/10.1111/1460-6984.12172

Schnur, T., Schwartz, M., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language, 54*(2), 199–227. https://doi.org/10.1016/j.jml.2005.10.002

Schwartz, M. F., & Dell, G. S. (2016). Word production from the perspective of speech errors in aphasia. In *Neurobiology of language*. Academic Press. https://doi.org/10.1016/B978-0-12-407794-2.00056-0.

Schwartz, M. F., Faseyitan, O., Kim, J., & Coslett, H. B. (2012). The dorsal stream contribution to phonological retrieval in object naming. *Brain, 135*(12), 3799–3814. https://doi.org/10.1093/brain/aws300

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

Shallice, T., & Warrington, E. K. (1977). Auditory-verbal short-term memory impairment and conduction aphasia. *Brain and Language, 4*(4), 479–491. https://doi.org/10.1016/0093-934X(77)90040-2

Sidiropoulos, K., de Bleser, R., Ackermann, H., & Preilowski, B. (2008). Pre-lexical disorders in repetition conduction aphasia. *Neuropsychologia, 46*(14), 3225–3238. https://doi.org/10.1016/j.neuropsychologia.2008.07.026

Stark, B. C., Basilakos, A., Hickok, G., Rorden, C., Bonilha, L., & Fridriksson, J. (2019). Neural organization of speech production: A lesion-based study of error patterns in connected speech. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 117*, 228–246. https://doi.org/10.1016/j.cortex.2019.02.029

Torres-Prioris, M. J., López-Barroso, D., Càmara, E., Fittipaldi, S., Sedeño, L., Ibáñez, A., … García, A. M. (2020). Neurocognitive signatures of phonemic sequencing in expert backward speakers. *Scientific Reports, 10*(1), Article 10621. https://doi.org/10.1038/s41598-020-67551-z

Ueno, T., & Lambon Ralph, M. A. (2013). The roles of the "ventral" semantic and "dorsal" pathways in conduite d'approche: a neuroanatomically-constrained computational modeling investigation. *Frontiers in Human Neuroscience, 7*, 422. https://doi.org/10.3389/fnhum.2013.00422

Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph, M. A. (2011). Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron, 72*(2), 385–396. https://doi.org/10.1016/j.neuron.2011.09.013