Supplemental Material S1. MAPPD-12K dataset extraction, preparation, and formatting.

The following outlines our process for extracting and preparing MAPPD-12K, the dataset of target-response pair produced on the Philadelphia Naming Test (PNT; Roach et al., 1996) that was used in the current study. All data are freely available as part of the Moss Aphasia Psycholinguistic Project Database (MAPPD; Mirman et al., 2010). A full list of instructions for extracting and reading MAPPD files is available to account subscribers at https://mappd.org.

Dataset Extraction

All PNT administrations from every participant in the MAPPD were downloaded on July 2, 2019. Items selected in MAPPD interface were: (a) test type: PNT; (b) demographic and clinical information: select all; (c) target word: select "test word"; and (d) response info: select "phonetic response," "regular response," "conventional response code." Control participants, defined by MAPPD as participants whose identifier began with C, were removed. All second to eighth PNT administrations, as applicable, were removed. This resulted in 51,800 PNT target-response pairs from 296 individuals with aphasia.

Dataset Preparation

Legacy IPA font used for the response phonemic transcriptions was converted to Unicode using a house algorithm, and modifiers (i.e., portions of multiword responses encased in parentheses, as defined in the instruction file included on http://mappd.org) were removed. These edited response phonemic transcriptions were listed under a new column titled *New_phonetic_response*.

Then, select subsets of codes were changed for the current study. All changes, along with the unchanged codes, were concatenated and listed in a separate, additional column titled *Conventional_response_with_recodes*.

Apraxia Lenient Coding

Per the PNT scoring guidelines (available at https://mrri.org/philadelphia-naming-test/), correct responses produced by individuals with concomitant apraxia of speech may be allotted one phonemic substitution, addition, or omission; however, this coding criteria is optional and not uniformly applied across participants. As such, we rescored all cases (n = 922) according to the conventional PNT coding guidelines. Recoding was completed independently by two trained annotators (M. C. and a graduate research assistant). Any instance of coding disagreement was reviewed and resolved via joint consensus with input from a third annotator (G. F.).

Two-step Coding

As outlined in the PNT scoring guidelines, any given conventional code can be further refined in the form of a two-step code, reflecting the locus of breakdown at both the lexical-semantic and phonological encoding stages of spoken word production (see Dell et al., 1997, for a description of the theory). In cases where the two-step code was used (n = 335), we converted these to their conventional code analogue.

Incomplete or Missing Codes

A small minority of responses contained either incomplete two-step codes or missing codes (n = 13). These cases were (re-)coded according to the same procedure as outlined for the aforementioned apraxia lenient coding.

Dataset Subsetting

A subset of single-word cardinal paraphasias, as defined in the body of the paper, from the 51,800-response dataset was created, as ParAlg's machinery is not yet optimized to predict multiword or non-paraphasic responses. This was done by removing the following: (a) responses with the codes Adm E, B, C, D, MO, NR, PP, PP-F, PP-N, and Prima (n = 39,461) listed under the *Conventional_response_with_recodes* column; (b) responses with multiple words or attempts as words, operationalized as a space between two character strings in the phonemic transcription (i.e., *New_phonetic_response* column, n = 180) or orthographic transcription (i.e., *Regular_response_with_fixes* column, as described below; n = 62 not excluded in the previous step); (c) responses coded as nonlexical that were missing a phonemic transcription (n = 2), operationalized as a blank cell in the *New_phonetic_response* column and a code of AN or N in the *Conventional_response_with_recodes* column; and (d) responses coded as nonlexical whose phonemic transcriptions contained non-English phonemes (n = 1), operationalized in the same manner as the previous step. This yielded 12,008 target-response pairs from 296 individuals with aphasia.

Orthographic Transcription Editing

All target-response pairs from 12,008 dataset were divided into four subsets: (a) responses with a lexical code (i.e., S, M, F, O) that contained an orthographic transcription (n = 6490), (b) responses with a lexical code that were missing an orthographic transcription (n = 32), (c) responses with a nonlexical code (i.e., AN, N) with an orthographic transcription (n = 698), and (d) responses with a nonlexical code without an orthographic transcription (n = 4,788).

Responses from subset (d) were unchanged, and responses from subset (c) were uniformly stripped of their orthographic transcription. Responses from subset (b) were independently reviewed by three annotators (two undergraduate research assistants and one graduate research assistant), who generated an orthographic transcription from the available phonemic transcription. Any discrepancies in transcription among the three annotators were resolved by two additional annotators (M. C. and G. F.).

Subset (a) was independently reviewed by four annotators (three undergraduate research assistants and one graduate research assistant), who flagged all cases where a one-to-one correspondence between the phonemic and orthographic transcription, operationalized as a match between the two transcriptions in the Merriam Webster online dictionary (https://www.merriam-webster.com/), was absent. A fifth annotator (M. C.) reviewed all of the flagged cases and generated a finalized list (n = 646). This finalized list was then reviewed by two annotators (M. C. and G. F.), who jointly determined whether cases were retained versus edited. Orthography was retained for all responses with the exception of two categories: (a) responses where the orthographic transcription and the target were the same yet the human annotator assigned a lexical code and the phonemic transcription shared a one-to-one correspondence with an entry in the Merriam Webster online dictionary, and (b) responses that contained misspellings or erroneous punctuation in the orthographic transcription. In the former category, if there was more than one possible orthographic representation (i.e., homophone) for a given response, the word with the highest frequency count in the SUBTLEXus database (Brysbaert & New, 2009) was selected. The word was considered to be a noun unless the code assigned was F, in which case other acceptable word classes (i.e., verb, adjective, adverb) were also considered. If there was a clear mismatch among the orthographic transcription, phonemic

transcription, and assigned code; and the annotators were unable to judge whether the original orthography was appropriate, the code of -777 (n = 9) was assigned.

Upon completion of the orthographic transcription changes, all were merged with the pre-existing orthography from MAPPD and listed under a new column titled *Regular_response_with_fixes*. Responses with the code -777 were subsequently removed, yielding 11,999 target-response pairs from 296 individuals with aphasia.

Dataset Formatting for ParAlg

For the purposes of the current study, all demographic or clinical information was also removed with the exception of the following: participant identifier (i.e., *Anonymous_Subject_ID* column), aphasia subtype (i.e., *Diagnosis* column), and months post-onset (i.e., *Months_post_onset* column). Finally, columns containing unedited response orthographic transcriptions (i.e., *Regular_response*), the unedited response phonemic transcriptions (i.e., *Phonetic_response*), and paraphasia codes (i.e., *Conventional_response_code*) were removed. This final change reflects our research group's decision to use our optimization efforts (e.g., edited orthographic transcriptions) to test the agreement of ParAlg against human annotators.

Column headers were renamed and reordered to match the formatting requirements of ParAlg at the time of the current study. Moreover, two additional columns were added. Changes included the following and are listed from left-to-right order of the data file (see Supplemental Material S2): (a) a column listing a numerical identifier for each target-response pair was added and titled *productionID*; (b) the *Anonymous_Subject_ID* column was moved and renamed *ID*; (c) the *Diagnosis* column was moved and renamed *Dx*; (d) the *Months_post_onset* column was moved and renamed *Months.Post.Onset*; (e) the column containing the orthographic target transcriptions was moved, and the column header was changed from *Test_word* to *Target*; (f) a blank column titled *Production* was added, which is later populated with the preprocessed phonemic response transcription (see Appendix A for more details regarding this process); (g) the column containing edited orthographic response transcriptions was moved, and the column header was changed from *Regular_response_with_fixes* to *Production.Orthographic*; (h) the column containing edited paraphasia codes (i.e., *Conventional _response_with_recodes*) was moved and renamed to *Code*; and (i) the column containing edited response phonemic transcriptions (i.e., *New_phonetic_response*) was moved and renamed *Production.Unicode*.

Duplicate Target-response Pair Tagging

In an effort to reduce coding burden for the item-level discrepancy analysis of the current study, duplicate target-response pairs within MAPPD-12K were manually identified and tagged by the first author (M. C.). Duplicates were defined as target-response pairs that were identical with regard to target orthographic transcription, response orthographic transcription, response phonemic transcription, and human-annotated paraphasia code. Differences in capitalization and the presence of extraneous punctuation, aside from the presence of diacritics, were not considered. When a duplicate was identified, the first-appearing pair in terms of *productionID* was treated as the unique pair and all subsequent pairs were tagged with the label "duplicate" under an additional column labeled *Duplicates* that was added as the right-most column to MAPPD-12K. Non-duplicate pairs were left blank. Within the 11,999-pair dataset, 2,719 duplicates were identified.

References

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

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

Mirman, D., Strauss, T. J., Brecher, A., Walker, G. M., Sobel, P., Dell, G. S., & Schwartz, M. F. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology, 27*(6), 495–504. https://doi.org/10.1080/02643294.2011.574112

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology, 24,* 121–133.