SIG 2

Research Article

Using Machine Learning to Explain Paraphasias in Narratives of People With Aphasia

Rosa Zavaleta,[a] Jacob Brue,[b] Sandip Sen,[b] and Laura Wilson[a]

[a] Department of Communication Sciences & Disorders, The University of Tulsa, OK [b] Tandy School of Computer Science, The University of Tulsa, OK

ARTICLE INFO

Article History:
Received December 14, 2023
Revision received May 30, 2024
Accepted November 21, 2024

Editor-in-Chief: Kendrea L. (Focht) Garand
Editor: Dallin John Bailey

https://doi.org/10.1044/2024_PERSP-23-00291

ABSTRACT

Purpose: This study examines how personal, clinical, and word-level features explain paraphasias when using machine learning–based error analysis on the narratives of people with aphasia (PWA).
Method: We used AphasiaBank as the source of narrative transcript data for 236 PWA. We tested machine learning classification algorithms including decision trees and random forests on the utterances of PWA, including nonparaphasic words and intended words when paraphasias were produced. We classified target words as paraphasic or nonparaphasic based on PWA's age; aphasia severity, duration, and type; presence of apraxia or dysarthria; and word-level features including part of speech, word frequency, imageability, syllable count, and location in the utterance. We measured the models' predictive accuracy across classification thresholds on held-out test sets, and we used feature analysis to compare feature importance.
Results: At the word level, our random forest model achieved an area under curve (AUC) of 0.896. We found a sensitivity of 0.821 for semantic paraphasias, 0.764 for phonemic paraphasias, and 0.872 for neologistic paraphasias. The most salient features, in order of importance, were word frequency, imageability, part of speech, age, severity, and syllable count, followed by aphasia duration, location of word, presence of apraxia, type of aphasia (e.g., fluent), and presence of dysarthria. Our random forest model that included information about surrounding words achieved AUC scores ranging from 0.881 to 0.899. Additionally, we developed a model that was trained on surrounding words and their respective features, but not given the actual error word. The best model achieved an AUC of 0.745.
Conclusions: Machine learning can aid in the explanation of paraphasias. In this study, we analyzed word- and person-level features and highlighted the nonrandom nature of paraphasic productions. Furthermore, this lays the groundwork for developing machine learning models with clinical applications at the various stages of treatment of PWA.
Supplemental Material: https://doi.org/10.23641/asha.28474172

Aphasia is an acquired communication disorder that results from damage or loss to the cortical and/or subcortical areas of the brain responsible for language (H. Le & Lui, 2023). Typically, there is damage in the left hemisphere due to its critical role in language (Brady et al., 2012). Specifically, damage to the perisylvian network for language can lead to impairments in expressive and receptive language skills (Grossman & Irwin, 2018). The most common cause for aphasia is a type of cerebrovascular accident called ischemic stroke (Fergadiotis et al., 2019). Aphasia can also be caused by traumatic brain injury, tumors, or progressive disorders such as Alzheimer's and hemorrhagic strokes (Grochmal-Bach et al., 2009). Of people that experience a cerebrovascular accident, about one third will become people with aphasia (PWA; Brady et al., 2012). Overall, this is equivalent to over 1 million people in the United States who

Correspondence to Laura Wilson: Laura-wilson@utulsa.edu. **Publisher Note:** This article is part of the Forum: Neurogenic Communication Disorders: Select Papers From the 2023 ASHA Convention. ***Disclosure:*** *Laura Wilson is a Certified Brain Injury Specialist Trainer through the Academy of Certified Brain Injury Specialists (since 2016). Wilson participated in the 2019 cohort of ASHA's Leadership Development Program and ASHA's Faculty Development Institute in 2021. The other authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

have aphasia and about 180,000 new cases per year (National Institute on Deafness and Other Communication Disorders, 2017).

Characteristics and Classification

The symptoms that PWA experience can vary from mild impairments at the word level to larger language deficits according to lesion site and size. They may experience these impairments in language components such as semantics, phonology, morphology, and/or syntax (H. Le & Lui, 2023). Aphasia subtypes can be used to categorize individuals based on their patterns of impairment across the language modalities. For example, researchers may dichotomize aphasia into nonfluent and fluent aphasia, based on their verbal output (Hallowell, 2017). Nonfluent aphasias (e.g., Broca's aphasia, global aphasia, and transcortical motor aphasia) are characterized by verbal output that is shorter, more effortful, and often agrammatic in nature. Fluent aphasias (e.g., Wernicke's aphasia, transcortical sensory aphasia, anomic aphasia, and conduction aphasia) are characterized by effortless verbal output without hesitations but may include errors or limited meaning. For this study, we analyzed subjects based on their fluent or nonfluent aphasia classification, without reference to subcategories (e.g., Broca's or Wernicke's aphasia).

Paraphasias

Paraphasias are substitution-based, word-level errors that are characteristic of PWA (Dalton et al., 2018). In general, these word-level errors are related to a breakdown in lexical access, related to semantic retrieval, phonological retrieval, or the interaction of the two (Schwartz et al., 2006). Although there are several classification schemes for paraphasias that can aim to categorize types of paraphasias based on the type of error (i.e., phonemic or semantic) or its distance from a target word, a widely used categorization system (that is used by the AphasiaBank) includes *semantic, phonemic*, and *neologistic paraphasias.* Respectively, a person with aphasia may substitute a target word (e.g., *pumpkin*) based on a semantic error (e.g., *apple*), a phonemic error (e.g., the nonreal word "tumpkin" or the real word "bumpkin"), or a phonologically unrelated neologistic error (e.g., *perka*), which can all impact the listener's comprehension of the intended message. Previous studies found evidence of a relationship between external factors, such as the frequency of a word in the English language, and resulting deficits in naming accuracy (Butterworth et al., 1984; Goodglass et al., 1969). Typically, paraphasia studies examine factors such as part of speech, imageability, and frequency. For example, Nickels and Howard (1995) set out to explore the effects that psycholinguistic factors played in word-finding difficulties. Out of the factors they studied, including frequency, word length, and age of acquisition, they found separate correlations between imageability, the ease with which a word can be pictured, and age of acquisition in errors of people with fluent and nonfluent aphasia. These findings were confirmed with later studies that demonstrated PWA perceive words with higher imageability as easier to produce, and across tasks and types, and are more likely to use verbs and nouns with "higher than average" imageability values (H. Bird et al., 2003, p. 6). In our analysis, we build upon these foundational features (part of speech, imageability, and frequency) by incorporating additional features, specifically word length and sentence position. We theorize that word length could serve as an informal measure of conceptual complexity (Lewis & Frank, 2016), while position in a sentence may reveal potential patterns in error production. Our rationale for including these new features is based on the hypothesis that error production is influenced by a simultaneous interaction of various psycholinguistic variables. For example, despite scanty evidence about the effects of psycholinguistic variables, we theorize word length could serve as an informal measure of conceptual complexity (Lewis & Frank, 2016) and position in a sentence as a way of describing potential patterns of error productions.

The evidence described above has largely been generated from studies of language production in confrontation naming tasks. Growing research has investigated the role of paraphasias in narratives (Fromm et al., 2017; MacWhinney et al., 2011). Part of the rationale for using narrative tasks is that they assess real-world performance in a way that isolated confrontation naming tasks cannot by eliciting typically longer and more diverse speech samples.

Although performance on confrontation tasks is related to discourse informativeness—defined as the general extent to which spoken or written discourse provides relevant, useful, and clear information to its audience—on narrative discourse tasks (Fergadiotis et al., 2019), it only accounts for about 62% of the variance in discourse informativeness. This suggests that narrative discourse relies on and can provide different information about linguistic skills.

In addition to an increase in language freedom in participants, narrative tasks allow for context to researchers (Dillow, 2013). In other words, while a confrontation naming task requires an exact response, a narrative task presents as an open-ended question without a required correct response. In this way, participants can speak with the vocabulary, grammar, fluency, and pronunciation that more closely match their typical language while still giving researchers a contextualized response, which allows for more successful identification of target words in the presence of errored productions.

Aphasia Research and Machine Learning

Even with online repositories of language samples (e.g., MacWhinney et al., 2011), the process of analyzing the narratives of PWA can be laborious and time-consuming. As a result, in the past few years, ground-breaking research has been conducted using both transcription and coding-based tools to aid in the process of error analysis as it relates to discourse production. For example, Adams et al. (2017) used natural language models to train a machine learning system to automatically classify errors as phonemic or neologistic based on known target words. They accomplished this by training their system to predict the possible target words that a speaker may be trying to say based on their retelling of the Cinderella story. In a similar manner, D. Le et al. (2018) analyzed the discourse of participants in the AphasiaBank repository and found that 12% of errors in samples fit the criteria for paraphasias. They were then able to automatically classify phonemic and neologistic paraphasias from transcripts with known and unknown target words.

Despite the access to databases with large speech samples of PWA and the recent use of machine learning models to aid automatic diagnosis and discourse analysis in the field of aphasiology (Järvelin & Juhola, 2011; Jothi & Mamatha, 2020), no studies to date have produced a model that can classify all types and evaluate which person-level and psycholinguistic factors influence the production of paraphasias. As a result, the purpose of this study is to address the following research question: How do person-level features (e.g., age) and psycholinguistic features of both target and surrounding words (e.g., word frequency, part of speech) explain paraphasic productions when incorporated into machine learning models? The findings of study help us to better understand the contributors to paraphasic production in naturalistic communication settings, which can support more targeted treatment. Additionally, this study provides insight into the use of machine learning models in the study of aphasia and implications for therapeutic applications.

Method

Source of Data

All data came from AphasiaBank (MacWhinney et al., 2011), a repository commonly used by researchers to study aphasia. It is a large repository of transcribed audio and video recordings of PWA by trained interviewers. All recordings were collected in a span of 9 years by various research groups using a standardized protocol of questions and stimuli. We used subdatabases in English that were collected using the AphasiaBank protocol (MacWhinney et al., 2011, p. 3) and contain a transcription of the interview. Within the interview, participants were asked to do various free speech and semispontaneous tasks, including retelling the story of Cinderella. We focused on the latter by using the subset of each transcript that contained audio and transcribed retellings of the Cinderella story. This type of narrative allows for longer, more natural samples of communication while also providing some context to researchers on the messages being conveyed, which is highly relevant when analyzing errors and determining intended target words (Dillow, 2013). AphasiaBank also contains participants' demographic and clinical features, which we include in the analyses.

Transcripts

AphasiaBank collaborators segment each utterance of a person with aphasia using standard guidelines of syntax, intonation, pauses, and semantics (MacWhinney et al., 2000). This criterion leads to numbered, time-stamped utterances in the Codes for the Human Analysis of Transcripts (CHAT) transcription format. The format includes labels of utterance features (e.g., word repetitions, revisions, fillers, gestures, sound fragments, and unintelligible output). In CHAT, licensed speech-language pathologists marked all errors at the word level with an asterisk [*] and a letter specifying the type of error (phonological, semantic, neologism, morphological, and disfluencies). Further letters indicate the subtype of error, such as [*n:k] for an error categorized as neologistic with known target word, versus [*n:uk] for an error categorized as neologistic with an unknown target word. For the purposes of this study, we consider errors those that were marked as errors followed by the phonological, semantic, and/or neologistic symbols, regardless of the subtype.

Furthermore, for nonwords, including errors where the intended word is not known (and later imputed), the utterances are transcribed using the standardized International Phonetic Alphabet (IPA). In addition, we utilized participants' transcribed narratives in the CHAT format to extract coded features for the machine learning models (e.g., part of speech, position in a sentence, and word length). Transcripts were converted from the CHAT format to an XML (extensible markup language) format. We excluded utterances with untranscribed segments, and we removed word fragments from utterances.

Subjects

All subjects in this study are PWA who are at least 18 years old and resided in the United States at the time of the interview. Participants were excluded if they did not have aphasia as categorized by the Western Aphasia Battery–Revised (WAB-R), if they spoke any language other than English, and if they did not know the Cinderella story. These criteria resulted in narratives from 236 subjects being included in the analyses.

Measures

Independent Variables

We gathered information on word-level features to help with the prediction and identification of each type of error. Word-level features included imageability, frequency, word length, part of speech, and position in a sentence for each word. Specifically, we assign features to nonparaphasic words and the target words of paraphasias. In combination, these features were chosen because they have demonstrated a relationship with word productions in PWA (Howard et al., 2008). For example, the frequency of a word is a strong predictor for how fast a person with aphasia can utter the word in a naming task.

We utilized the Medical Research Council Psycholinguistic Database (Coltheart, 1981) to assign values for psycholinguistic features of the words in the included utterances. The MRC database was developed out of a need to provide researchers with word-level characteristics for psycholinguistic experiments and contains smaller databases for specific features. For example, we used the Coltheart (1981) database to assign imageability to the included utterances. Imageability ratings in the subdatabase were derived from previous sets of norms (Gilhooly & Logie, 1980; Paivio et al., 1968; Toglia & Battig, 1978) and yield subjective ratings between the values of 100 and 700 (from least to most imageable). Frequency values were derived from the Python library wordfreq, a tool that unifies several open web-based sources of English text to provide word usage frequency (Speer, 2022). Word length is an objective measure that was derived from each word's syllable count, as obtained by converting grapheme length to syllable count using the g2pE module in Python (Park & Kim, 2019). The part of speech and absolute position within the sentence are both derived from the AphasiaBank transcripts by the coding provided and our protocol, respectively.

Content words are those with substantive meaning relative to the grammatical structure of the sentence. Typically, they include nouns, verbs, adjectives, and adverbs. For this study, we included errors regardless of part of speech. Although paraphasia analyses are historically done with content words, errors at the narrative level can include functors as well (e.g., pronouns; Arslan et al., 2021). Thus, in the present study, we do not restrict analysis to content words only, as that would limit our examination of the person-level and word-level factors that contribute to paraphasic productions.

At the person level, we incorporated clinical features, including type (i.e., fluent or nonfluent) and severity of aphasia (as measured by the WAB-R Aphasia Quotient), time postonset, and years of therapy prior to the day of testing. To account for language production differences by age (Marini et al., 2005) and because age is related to likelihood of aphasia poststroke, as well as type of aphasia (Ellis & Urban, 2016), we also included age as a variable in the model. A summary of demographic and clinical information about the subjects included in our sample is included in the results (see Tables 1 and 2).

Dependent Variables

The dependent variable of interest was the presence of a paraphasic error. Specifically, we looked at the success with which our model was able to determine the presence or absence of a paraphasic error based on the word-level and person-level features described above.

Procedure

We first parsed the data from the CHAT transcripts for each participant. AphasiaBank provided the Chatter program (MacWhinney et al., 2011) needed to convert the data into an XML format. This separated the data into tags, which could then be used to separate the intended narrative from the rest of the data. We included all the utterances during the retelling of the Cinderella story. We filtered utterances containing untranscribed speech, leaving the final corpus of text. For word errors that had a likely intended target word attached, the intended target word replaced the word error in the analysis. This allowed us to incorporate the features of the intended word (e.g., frequency of the intended word) to help determine the role of that feature in the likelihood of error production.

Imputation Strategy

Not every word that contains an error includes the intended target. For example, neologistic errors that were

Table 1. Demographic information of people with aphasia (*N* = 235).

Information	***M* in years (*SD*)**		
Age (at time of test)[a]	61.7 years (4.0)		
Education	15.5 years (2.8)		
		%	***n***
Sex	Female	41.28	97
	Male	58.72	138
Race	White	83.40	196
	African American	13.20	31
	Other/not reported	3.40	8

[a] Age is the only demographic variable included in the models. Other demographic variables are provided for the purpose of describing the sample.

Table 2. Diagnosis-related information of people with aphasia (*N* = 235).

Information	***M* (*SD*)**		
Severity	69.1 WAB score (17.9)		
Time postonset	5.4 years (5.0)		
Treatment duration	3.3 years (4.0)		
		%	*n*
Etiology	Stroke	90.8	213
	Other	3.8	9
	Unknown/NA	5.5	13
Comorbidity	Apraxia	31.9	75
	Dysarthria	8.5	20
Types of aphasia	Broca	34	79
	Anomic	26	60
	Conductive	12	298
	Wernicke	6	15
	Global	2	6
	Other/NA	20	46
Fluency	Fluent	44	104
	Nonfluent	45	105
	Unknown	11	26

Note. WAB = Western Aphasia Battery; NA = not applicable.

not recognized by the researchers were then left in an IPA format. To capture these words in our data, we developed a protocol for manually determining the most likely intended target word based on the combination of audio, visual, and transcription data provided. We first identified utterances that contained at least five words (excluding fragments and fillers) and included the error within that total count. We then found the phonetic transcript for the utterance and listened to and watched the three lines of utterances that preceded and followed the utterance containing the error. Two researchers (R.Z. and L.W.) then independently recorded their first and second impressions of the target word for the neologistic error and the corresponding parts of speech. Initial interrater reliability was at 55%, rising to 98% consensus with discussion.

In addition to the missing replacement words, some of the words are missing a part of speech label. For the 14.6% of words where the part of speech was missing, we replaced values using a trained part of speech tagger. We used the openly available Natural Language Toolkit Python library (nltk), which provides a part of speech tagging model trained on large bodies of labeled English sentences to predict the part of speech of words in the context of a sentence (S. Bird et al., 2009). The predictions are 69.1% accurate when predicting the part of speech of the labeled words from the data set. The predicted part of speech tags was used only for words missing a part of speech in the transcript.

For the 29.2% of words where the imageability was not available in the Medical Research Council Psycholinguistic Database (Coltheart, 1981), a linear regression imputation model was trained to predict the imageability of the word. First, each word with an assigned imageability in the database is translated to a vector using a word2vec model implemented in GenSim. This model is one that was previously trained on the Google News data set, a popular standard for word2vec models in the field of natural language processing (Mikolov et al., 2013). The linear regression model is then trained on each word's word vectors to predict that word's imageability. This approach is based on existing literature that shows that imageability can be explained in part by these word embeddings (e.g., Ljubešić et al., 2018; Matsuhira et al., 2020). The models on average achieve an R^2 score of .630 on training samples and .580 on unseen test samples after 100 rounds of fivefold cross validation.

There were 1.5% of words that had no associated word vector. These words were imputed with the mean imageability After the imputation was complete, the data set includes 56,419 words across 7,719 utterances.

Supervised Machine Learning Classification Models

Following imputation, we utilized a variety of supervised machine learning classification models. Supervised machine learning models are trained using data samples that include input features and an output label. A classification model specifically assigns a class label to the data (e.g., a certain word is an adjective). Our models were designed to determine the likelihood that an intended target word results in a paraphasia using person-level and word-level features. Specifically, our person-level features included age, aphasia severity (very severe, severe, moderate, mild), aphasia duration, aphasia type (fluent or nonfluent), and presence of apraxia or dysarthria. Our word-level features included part of speech, word frequency, imageability, syllable count, and location within the utterance. For our word-level models, the word-level features of only the target word were included. For contextual models, the same word-level features of the surrounding words were included in the model as well.

We focused on models that balance predictive power with interpretability. A model with high predictive power can represent more complex mathematical relationships between input features and predicted class. An interpretable model is one that produces an understandable connection between the features and the output. For example, a deep learning model has high predictive power, but it is currently very difficult to produce meaningful or clinically relevant interpretations of these models. A logistic regression algorithm is strongly interpretable, but it cannot represent feature interactions or nonlinear relationships, so it has weak predictive power.

One of the more interpretable classifier models is called a decision tree (Charbuty & Abdulazeez, 2021). A decision tree model consists of a set of decision nodes and a set of prediction nodes. Each decision node contains a simple comparison function of one of the features. For instance, if one of the features is "age," a node may contain the comparison "age > 50." The nodes are connected such that each outcome of the comparison of a node leads to another node. Decision nodes at the end of the tree lead to prediction nodes, which contain the model's prediction for the input. The first decision node in a tree is the root of the decision tree. When the input features are provided to the model for prediction, the root decision node comparison is checked, and then a path is taken down the tree to a prediction node.

We use a binary decision tree so that each decision node has exactly two paths out, and we use a limit to the maximum depth of the tree to a path length of eight decision nodes, which improves the generalization of the model. The tree can be interpreted by observing the decision nodes as a sample takes a path through the tree. A measure of the importance of each feature can be calculated using Gini importance, an algorithm that performs an average of the decrease in uncertainty that results from samples passing through a decision node weighted by how many samples pass through the node, attributing each score to the feature being compared at the node.

Decision trees are highly interpretable models, but there are many techniques that generalize to stronger predictive models. Random forests are an ensemble method, meaning they are composed of smaller models. Random forests train multiple individual decision tree models and use the plurality vote of the decision trees for a prediction (Ho, 1995). To promote a diverse view of the set of features, the trees are constrained to use a randomly selected subset of the features (we used a subset with size equal to the square root of the total feature set). To promote a diverse view of the data set, the random forest utilizes bagging during training, a process that samples the data set into multiple randomized overlapping subsets to improve the ensemble (Brieman, 2001). The combination of decision trees results in a model that generalizes better to new samples. However, the model can no longer be interpreted by simply observing a single path through a tree. It is still possible to estimate the importance of a feature using the average Gini index across all of the decision trees.

Training

In all models, 80% of the data were used for training purposes, meaning the decision trees were given all information available for those utterances as well as which instances were errors. The final set of features provided includes personal, diagnostic, and word-level features. The result is then a model that automatically determines the likelihood of a word being a paraphasia in each utterance of the narratives of PWA. The remaining 20% of the data were used for testing the model. All of the input features were provided just like in the training set, but now the trained classifier model uses the information it has learned about the relationship between input features and output classification to determine whether a word in the new data set is likely to be in error.

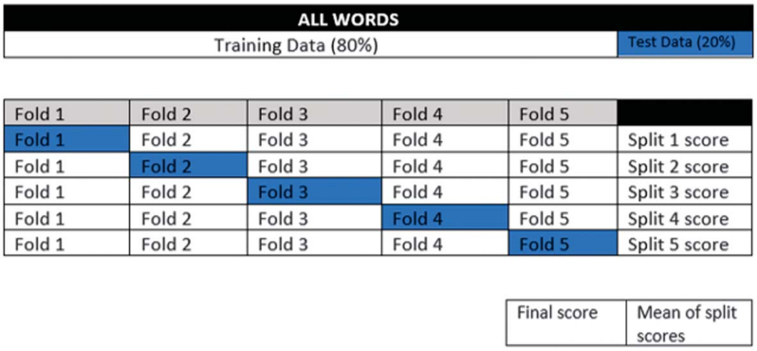
We utilized a fivefold analysis, meaning that the training and testing process used five randomized folds, or subsets, of the data, with a different testing data set each time (see Figure 1 for an example of the cross validation of this study). This yielded random forest models with a different training and testing set each. This process was repeated for 100 different randomized folds for a total of 500 models.

Results

Participant Data

Of the individual transcripts of Cinderella retellings provided by the AphasiaBank, the transcripts of 236

Figure 1. Fivefold cross-validation.



ALL WORDS	
Training Data (80%)	Test Data (20%)

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Split 1 score
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Split 2 score
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Split 3 score
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Split 4 score
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Split 5 score

Final score	Mean of split scores

PWA met eligibility criteria. However, one individual's transcripts included only unusable utterances, so the final number of PWA included in the analyses was 235. Of these individuals, 97 were women and 138 were men, with an average age of 61.7 years (*SD* = 12.4 years). The majority of participants acquired aphasia from a stroke.

In these analyses, we used the diagnostic labels provided in AphasiaBank to classify aphasia type. We included 104 individuals with fluent aphasia, 105 with nonfluent aphasia, and 26 with no classification and/or an unknown classification (see Table 1 for demographic information and Table 2 for diagnosis related information of the PWA).

Word-Level Models

The data yielded 56,419 words, of which 2,037 (3.6%) were paraphasic errors. This included 964 semantic paraphasias, 657 phonemic paraphasias, and 416 neologistic paraphasias. Of the errors without replacements, 1,627 replacements came from AphasiaBank, while the rest were manually imputed. This included a total of 410 errors determined through the manual imputation process, of which 248 were semantic errors and 162 were neologistic errors. Table 3 describes the features of the words included in the data set.

We use a receiver operating characteristic (ROC) curve to test the effectiveness of our decision tree classifiers and random forest models. Specifically, we measured effectiveness as an average of the area under the curve (AUC) of the ROC curve of the 500 models. An AUC of 0.5 suggests a classifier no better than random chance, whereas an AUC of 1.0 suggests a perfect classifier (i.e., 100% sensitivity and 100% specificity in determining whether a word is an error). For the target word level, our decision tree model achieved an AUC of 0.860. Our random forest model, with a maximum depth of 8, achieved the highest AUC of 0.896 at the word level (see Figure 2). We repeated the analysis with a random forest model of unlimited depth with little change in performance (AUC of 0.894; see Table 4).

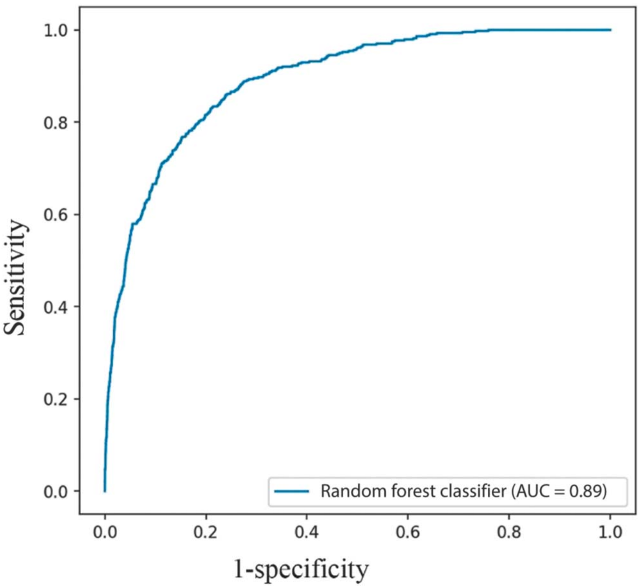
When using the random forest model with maximum depth of 8 on test data, we selected an error classification threshold that balances sensitivity with specificity. This model achieved an average specificity of 0.802 and sensitivity of 0.813. When assessed based on error type, we find that the sensitivity for semantic paraphasias was 0.821, the sensitivity for phonemic paraphasias was 0.764, and the sensitivity for neologistic paraphasias was 0.872.

From our best model (i.e., the random forest model with a maximum depth of 8), we examined the most salient features from person-level (demographic or diagnostic features) and/or word level (e.g., imageability) features that influence paraphasic errors. We used mean decrease Gini impurity to measure feature importance. This feature importance method attributes the decrease in the Gini impurity index caused by each node to that node's feature, taking a weighted average based on how much training data were sorted using that path through the tree. The results for each feature are normalized to sum to 1 (Scornet, 2023). The most salient features, in order of importance, were word frequency, imageability, part of speech, age, severity, and syllable count, followed by aphasia duration, location of word, status of apraxia, type of aphasia (e.g., fluent), and status of dysarthria (see Figure 3). Of those, a separate multiple logistic regression on all data to determine directionality of the variables found that longer syllable count, greater imageability, certain parts of speech (i.e., nouns, pronouns, adjectives, determiners, verbs, prepositions, adverbs), and position later in the utterance were related to greater likelihood of error, while higher frequency and other parts of speech (i.e., infinitives, onomatopoeias, interjections, complementizers, auxiliary verbs, and conjunctions) were related to lower likelihood of error. As part of the same multiple logistic regression model, the person-level features that were related to increased likelihood of error include more severe aphasias, greater age, nonfluent aphasias, and presence of apraxia and/or dysarthria, while longer duration of aphasia reduced the likelihood of error.

Table 3. Features of words included in models (*N* = 56,419).

Word-level feature	Possible values for categorical variables or range for numerical values	Mode for categorical values/average for numerical values
Part of speech	Adjective, adverb, auxiliary verb, complementizer, conjunction, determiner, infinitive, interjection, noun, onomatopoeia, preposition, pronoun, verb	Verb
Location	0–68	5.89
Imageability	195–657.58	337.99
Frequency	0–0.0537	0.0086
Syllable count	0–8	1.22

Figure 2. Receiver operating characteristic curve of the random forest model, maximum depth of 8 (word-level model). AUC = area under curve.



Contextual Models

Our contextual model determines paraphasias based on features of the speaker, the intended word, and the features of the surrounding words. When given information about the target word and the psycholinguistic features of surrounding words (with varying window sizes; see Table 4), AUC ranged from 0.881 to 0.899, which suggests these models are not inherently stronger models than those that include the target word features without word-level features of surrounding words. Notably, the only context-based model that outperformed the word-level model was the random forest model with unlimited depth that included information about the target word, as well as the word before and after that word. This model resulted in an AUC of 0.899 (compared to the target word only model with an AUC of 0.894). As a follow-up analysis, we developed models that could determine the presence of a paraphasia when given the surrounding words and their respective psycholinguistic features, but not given the features of the target word itself. When given the single word before and after the target word, our decision tree model achieved an AUC of 0.687. This rose to an AUC of 0.744 with a random forest model with a maximum depth of 8 and an AUC of 0.737 with a random forest model of an unlimited depth. Models that expanded the window to two words before and after or three words before and after the target word performed similarly (with AUC of 0.743 and 0.745, respectively). The model performance was poorer when given information only about words preceding the target word, with the best model being the random forest model with a maximum depth of 8 that used the words in the −2 and −1 position relative to the target word (i.e., the two words preceding the target word). This model resulted in an AUC of 0.729.

Discussion

In this study, we explored how personal, clinical, and psycholinguistic features could be used to determine if a target word was likely to be produced in error by a person with aphasia during a narrative language task. Our most accurate models (yielding an AUC of 0.897 and 0.899) were those that learned from the target word itself or from the target word plus the features of the single words that surrounded it, respectively. Training on additional information from more distant words in the utterance did not improve the model. Additionally, models without the target word information but with the information about words leading up to the target word yielded a maximum AUC of 0.729. This is not as successful as the models with target word information but does suggest that, even in the absence of a known target word, contextual information can help predict the production of paraphasias. This highlights the possibility of automated technologies that could utilize predictive language capabilities to support targeted, augmented communication in instances of likely paraphasic production.

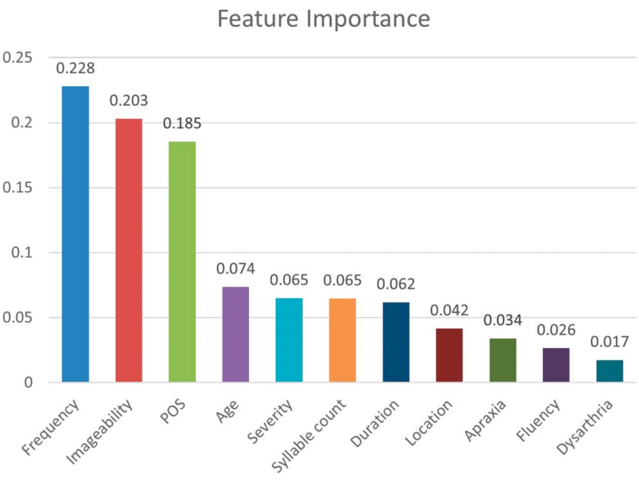
Our decision to introduce word features was influenced by recent findings in aphasiology that provide

Table 4. Area-under-curve values for random forest models with various window sizes.

	Window of words around target word included in analysis						
Model	**0 . . . 0**	**-2 . . . 0**	**-1 . . . 1**	**-4 . . . 0**	**–3 . . . 1**	**-2 . . . 2**	**-3 . . . 3**
Random forest model with maximum depth of 8	0.896	0.885	0.890	0.882	0.887	0.886	0.886
Random forest model with unlimited maximum depth	0.894	0.888	0.899	0.881	0.892	0.893	0.890

Note. The target word is labeled as 0. The word before the target word is –1, and the word after the target word is 1. All windows are inclusive. For example, a window of –3 to 1 would include the three words before the target word, the target word, and the first word following the target word.

Figure 3. Feature importance for the random forest model with a maximum depth of 8. POS = part of speech.



evidence for the connection between word-level features (e.g., imageability) and the likelihood of that word being an error. For example, our finding that part of speech, among other features, is related to error production is similar to the findings of Malyutina and Zelenkov (2019). They found that the structure of verbs, specifically the number of arguments, affected error production (e.g., intransitive verb included in the phrase "he jumps" requires less information than the transitive verb "he bakes," which has a direct object). They also found the effects of this underlying feature were consistent across people with fluent and nonfluent aphasia but differed at the word versus sentence level. This newer research into sentence-level production supported our decision to expand error detection from the word level to the utterance level, as the interaction of word features may affect the likelihood of a word being an error. Like Malyutina and Zelenkov (2019), we found a change in feature importance when expanding from target word models only to those with contextual information from surrounding words, or when using models with only the contextual information. For example, in models relying only on contextual information and not the target word itself, the presence of a determiner before a target word leads to increased likelihood of error in the next word (see Supplemental Material S1). Our automatic detection of the most salient features driving errors in PWA can aid clinicians in selecting effective therapeutic words for improved communication outcomes. For example, word frequency, imageability, and part of speech are the word features that were most salient in predicting whether a word will likely be a paraphasic error. Thus, these features could be systematically manipulated to target words that are most at risk for errored production and most likely to have a functional impact on the production of those with aphasia. Following this line of thought, newer studies in aphasia treatment are exploring the interaction between word-level features and treatment targets (Bailey et al., 2020). Researchers sought out to see the effects of targeting verbs with low concreteness (low imageability), or how easily word can be pictured, and found increased retrieval with targeted verbs that extended to untreated verbs. In this way, they demonstrate how strategically targeting the features that can increase the likelihood of errors can create more effective treatment approaches, thus aiding functional communication.

Applications

The findings of this study could be applied in several unique ways. Our model determined underlying features that are driving the word-level errors common in aphasia. Furthermore, it showed how machine learning can be a tool to sift through the complexities found within natural, disordered speech on a larger scale than is manually feasible during a traditional evaluation or therapy session. The 56,419 words used for this study were within sentences that were often incomplete, unintelligible, or agrammatic. This challenged us to apply statistical tools of natural language processing as preselected features were presented to successfully analyze patterns in disordered speech. Eventually, creating more efficient, real-time models could serve as a diagnostic tool for clinicians to systematically identify error patterns in the speech of PWA at a higher rate and with more nuanced word-level characteristics.

Within existing and capable predictive output systems, such as high-tech augmentative and alternative communication methods, our automatic analysis could eventually be refined and used to create reliable predictions of likely intended words. This predictive system could aid in reducing communication breakdowns between PWA and their communication partners. Over time, the use of machine learning models with accurate medical data, such as speech samples, could lead to a more accurate language recovery profile for a person with aphasia.

Limitations

The primary limitation of this study is driven by the adaptation of techniques based on natural language models to disordered speech. There is limited research in interdisciplinary studies of this nature, which challenged us to rationalize decisions at several stages of this study. For example, the use of decision tree classifiers meant that we needed the intended word and features for all utterances included in the study, which proved difficult with often unintelligible or limited speech. Additionally, we used large databases (Coltheart, 1981) for features that could not be acquired from AphasiaBank alone and found that there was still a large quantity of words without corresponding features. This was a fundamental issue we had to resolve with the use of regression imputation techniques (i.e., using existing data to predict the likely value of the feature for a new word) to successfully advance the study. Overall, missingness values evoked several questions about the best approaches regarding imputation strategies for our purposes.

Furthermore, another limitation of our current system is its reliance on manual transcriptions. For example, our use of the AphasiaBank narratives was made possible by the trained interviewers who recorded and transcribed interviews, along with the added, standardized language tags. Additionally, our manual imputation strategy for missing target words required us to listen to each unintelligible neologism and semantic paraphasia without replacement to create clean utterances with the likely intended words. Clinically, real-time and continuous transcriptions are unrealistic due to time constraints. In research, it can also be time-consuming to create clean transcripts when working with large sets of data. As a result, our study is limited by its replicability in clinical settings. However, the findings of the study can inform clinical practice by focusing on the role that salient features such as frequency and imageability play to help inform interventions, even if the specific transcribing is not feasible for a particular client.

Directions for Future Studies

Our current approach to analysis could be expanded in several ways. First, we could explore additional window sizes (i.e., adjusting the number of words before and after a target word) to determine if window sizes outside of what we tested within the current models can improve model performance. Additionally, we only explored random forest models with either a maximum depth of eight decision nodes or unlimited decision nodes. We could explore additional model options to optimize performance. Finally, we could incorporate sentence-level characteristics (e.g., average word frequency) as a feature to provide additional contextual information.

Future study could approach our broader problem in several different ways. We currently evaluate the likelihood that a target word results in a paraphasia when given a single word and its preselected features as well as surrounding words and their preselected features. We have not explored the impact of providing a machine learning model with longer utterances or entire speech samples. Our current approach is building the groundwork for the expansion of future work in this regard. This would also involve creating imputation strategies for part of speech of word errors without replacements.

Additionally, future work could also consider moving beyond determination of errors and creating a system that can determine the likelihood that a particular error is a semantic, phonemic, or neologistic paraphasia. Previous studies have found that automatic classification is possible for phonemic versus neologistic paraphasias (Adams et al., 2017), but no study to date has attempted to automatically classify all types of paraphasias at the discourse level. The inclusion of additional available data in the AphasiaBank, such as audio and visual recordings, could lead to more nuanced approaches within language models and possible improved outcomes.

The most exciting direction of future work is the creation of a machine learning system that can use clinical, psycholinguistic, and person-level features to predict the most likely word the speaker intended in place of a paraphasia. While our current system relies on manual entries, or a "best guess" approach, a more instantaneous system could potentially aid in communication and/or create more functional screeners for PWA. Lastly, our findings are bound to a highly contextualized narrative task, the Cinderella story. Future work should follow the work of Salem et al. (2023) by applying machine learning principles to novel, decontextualized utterances to increase the generalization and overall clinical applicability of findings.

In conclusion, the information garnered by this study revealed ways that machine learning can determine, with high accuracy, the likelihood of a word produced by PWA being an error. Through use of decision trees, random forest models, and hybrid imputation strategies, our study deepens understanding of the driving features of error productions in PWA and point to potential clinical applications that can lead to more effective interventions, and resulting communication, for PWA. The study also highlights how machine learning techniques can support the development of further models for speech and communication applications.

Author Contributions

Rosa Zavaleta: Conceptualization, Investigation; Writing – original draft, Writing – review & editing, Visualization. **Jacob Brue:** Conceptualization, Methodology, Software, Formal Analysis, Writing – review & editing, Visualization. **Sandip Sen:** Conceptualization, Methodology, Formal Analysis, Resources, Writing – review & editing, Supervision. **Laura Wilson:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision.

Data Availability Statement

The data sets analyzed during the current study are not publicly available, as access is restricted to members of the AphasiaBank consortium group. Those interested in joining the consortium should follow guidelines described at https://aphasia.talkbank.org/. The scripts used for data analyses are available as Supplemental Material S2.

Acknowledgments

Authors Rosa Zavaleta, Jacob Brue, and Laura Wilson received internal funding from The University of Tulsa to travel and present this study at the 2023 ASHA Convention. No other funding was obtained. The AphasiaBank is supported by grant funding from National Institute on Deafness and Other Communication Disorders Grant R01-DC008524 (2022–2027).

References

Adams, J., Bedrick, S., Fergadiotis, G., Gorman, K., & van Santen, J. (2017). Target word prediction and paraphasia classification in spoken discourse. In K. B. Cohen, D. Demner-Fushman, S. Ananiadou, & J. Tsujii (Eds.), *Proceedings of the BioNLP 2017 Workshop* (pp. 1–8). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-2301

Arslan, S., Devers, C., & Ferreiro, S. M. (2021). Pronoun processing in post-stroke aphasia: A meta-analytic review of individual data. *Journal of Neurolinguistics, 59,* Article 101005. https://doi.org/10.1016/j.jneuroling.2021.101005

Bailey, D. J., Nessler, C., Berggren, K. N., & Wambaugh, J. L. (2020). An aphasia treatment for verbs with low concreteness: A pilot study. *American Journal of Speech-Language Pathology, 29*(1), 299–318. https://doi.org/10.1044/2019_AJSLP-18-0257

Bird, H., Howard, D., & Franklin, S. (2003). Verbs and nouns: The importance of being imageable. *Journal of Neurolinguistics, 16*(2–3), 113–149. https://doi.org/10.1016/S0911-6044(02)00016-7

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly.

Brady, M. C., Kelly, H., Godwin, J., & Enderby, P. (2012). Speech and language therapy for aphasia following stroke. *Cochrane Database of Systematic Reviews*. Article CD000425. https://doi.org/10.1002/14651858.CD000425.pub3

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Butterworth, B., Howard, D., & McLoughlin, P. J. (1984). The semantic deficit in aphasia: The relationship between semantic errors in auditory comprehension and picture naming. *Neuropsychologia, 22*(4), 409–426. https://doi.org/10.1016/0028-3932(84)90036-9

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends, 2*(01), 20–28. https://doi.org/10.38094/jastt20165

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Dalton, S. G. H., Shultz, C., Henry, M. L., Hillis, A. E., & Richardson, J. D. (2018). Describing phonological paraphasias in three variants of primary progressive aphasia. *American Journal of Speech-Language Pathology, 27*(1S), 336–349. https://doi.org/10.1044/2017_AJSLP-16-0210

Dillow, E. P. (2013). *Narrative Discourse in aphasia: Main concept and core lexicon analyses of the Cinderella story* [Master's thesis, University of South Carolina]. University Libraries. https://scholarcommons.sc.edu/etd/2623

Ellis, C., & Urban, S. (2016). Age and aphasia: A review of presence, type, recovery and clinical outcomes. *Topics in Stroke Rehabilitation, 23*(6), 430–439. https://doi.org/10.1080/10749357.2016.1150412

Fergadiotis, G., Kapantzoglou, M., Kintz, S., & Wright, H. H. (2019). Modeling confrontation naming and discourse informativeness using structural equation modeling. *Aphasiology, 33*(5), 544–560. https://doi.org/10.1080/02687038.2018.1482404

Fromm, D., Forbes, M., Holland, A., Dalton, S. G., Richardson, J., & MacWhinney, B. (2017). Discourse characteristics in aphasia beyond the Western Aphasia Battery cutoff. *American Journal of Speech-Language Pathology, 26*(3), 762–768. https://doi.org/10.1044/2016_AJSLP-16-0071

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12*(4), 395–427. https://doi.org/10.3758/BF03201693

Goodglass, H., Hyde, M. R., & Blumstein, S. (1969). Frequency, picturability and availability of nouns in aphasia. *Cortex, 5*(2), 104–119. https://doi.org/10.1016/S0010-9452(69)80022-5

Grochmal-Bach, B., Pachalska, M., Markiewicz, K., Tomaszewski, W., Olszewski, H., & Pufal, A. (2009). Rehabilitation of a patient with aphasia due to severe traumatic brain injury. *Medical Science Monitor, 15*(4), CS67–CS76. https://pubmed.ncbi.nlm.nih.gov/19333207/

Grossman, M., & Irwin, D. J. (2018). Primary progressive aphasia and stroke aphasia. *Continuum, 24*(3), 745–767. https://doi.org/10.1212/CON.0000000000000618

Hallowell, B. (2017). *Aphasia and other acquired neurogenic language disorders: A guide for clinical excellence* (2nd ed.). Plural.

Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE. https://ieeexplore.ieee.org/document/598994

Howard, D., Osborne, F., Best, W., Hickin, J., & Herbert, R. (2008). Do picture-naming tests provide a valid assessment of lexical retrieval in conversation in aphasia? *Aphasiology, 22*(2), 184–203. https://doi.org/10.1080/02687030701262613

Järvelin, A., & Juhola, M. (2011). Comparison of machine learning methods for classifying aphasic and non-aphasic speakers. *Computer Methods and Programs in Biomedicine, 104*(3), 349–357. https://doi.org/10.1016/j.cmpb.2011.02.015

Jothi, K. R., & Mamatha, V. L. (2020). A systematic review of machine learning based automatic speech assessment system to evaluate speech impairment. In *3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India* (pp. 175–185). https://doi.org/10.1109/ICISS49785.2020.9315920

Le, D., Licata, K., & Mower Provost, E. (2018). Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication, 100,* 1–12. https://doi.org/10.1016/j.specom.2018.04.001

Le, H., & Liu, M. (2023). *Aphasia*. National Library of Medicine. https://www.ncbi.nlm.nih.gov/books/NBK559315/

Lewis, M. L., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition, 153,* 182–195. https://doi.org/10.1016/j.cognition.2016.04.003

Ljubešić, N., Fišer, D., & Peti-Stantić, A. (2018). Predicting concreteness and imageability of words within and across languages via word embeddings. In I. Augenstein, K. Cao, H. He, F. Hill, S. Gella, J. Kiros, H. Mei, & D. Misra (Eds.), *Proceedings of the Third Workshop on Representation Learning for NLP* (pp. 217–222). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3028

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Lawrence Erlbaum Associates. https://doi.org/10.21415/3mhn-0z89

MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. *Aphasiology, 25*(11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893

Malyutina, S., & Zelenkova, B. (2019). Verb argument structure effects in aphasia are different at single-word versus sentence level. *Aphasiology, 34*(4), 431–457. https://doi.org/10.1080/02687038.2019.1697866

Marini, A., Boewe, A., Caltagirone, C., & Carlomagno, S. (2005). Age-related differences in the production of textual descriptions. *Journal of Psycholinguistic Research, 34*(5), 439–463. https://doi.org/10.1007/S10936-005-6203-z

Matsuhira, C., Kastner, M. A., Ide, I., Kawanishi, Y., Hirayama, T., Doman, K., Deguchi, D., & Murase, H. (2020). Imageability estimation using visual and language features. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (pp. 306–310). https://doi.org/10.1145/3372278.3390731

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://doi.org/10.48550/arXiv.1301.3781

National Institute on Deafness and Other Communication Disorders. (2017). *Voice, speech and language*. Retrieved February 20, 2022, from https://www.nidcd.nih.gov/health/voice-speech-and-language

Nickels, L., & Howard, D. (1995). Aphasic naming: What matters? *Neuropsychologia, 33*(10), 1281–1303. https://doi.org/10.1016/0028-3932(95)00102-9

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*(1, Pt. 2), 1–25. https://doi.org/10.1037/h0025327

Park, J., & Kyubyong, K. (2019). *g2pe*. GitHub. https://github.com/Kyubyong/g2p

Salem, A. C., Gale, R. C., Fleegle, M., Fergadiotis, G., & Bedrick, S. (2023). Automating intended target identification for paraphasias in discourse using a large language model. *Journal of Speech, Language, and Hearing Research, 66*(12), 4949–4966. https://doi.org/10.1044/2023_JSLHR-23-00121

Scornet, E. (2023). Trees, forests, and impurity-based variable importance in regression. *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques, 59*(1), 21–52. https://doi.org/10.1214/21-AIHP1240

Speer, R. (2022). *rspeer/wordfreq: v3.0* (Version 3.0.2). Zenodo. https://doi.org/doidu.org/10.5281/zenodo.7199437

Schwartz, M., Dell, G., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming⋆. *Journal of Memory and Language, 54*(2), 228–264. https://doi.org/10.1016/j.jml.2005.10.001

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Lawrence Erlbaum.