Discourse- and lesion-based aphasia severity estimation using machine learning

Nicholas Riccardi

Satvik Nelakuditi

Chris Rorden

Julius Fridriksson

Rutvik H. Desai

Abstract

Discourse is a fundamentally important aspect of communication, and discourse production provides a wealth of information about linguistic ability. Aphasia commonly affects, in multiple ways, the ability to produce discourse. Comprehensive aphasia assessments such as the Western Aphasia Battery are time- and resource-intensive. We examined whether discourse measures can be used to assess aphasia severity, and whether this can serve as an ecologically valid, less resource-intensive measure. We used lexical features extracted from discourse tasks using three AphasiaBank prompts involving picture description, story narrative, and procedural discourse. These features were used to train a machine learning model to predict the Aphasia Quotient. We also compared and supplemented the model with lesion location information from structural neuroimaging. We found that discourse-based models could estimate aphasia severity well, and that they outperformed models based on lesion features. Addition of lesion features to the discourse features did not improve the performance of the discourse model substantially. Inspection of the most informative discourse features revealed that different prompt types taxed different aspects of language. These findings suggest that discourse can be used to estimate aphasia severity, and provide insight into the linguistic content elicited by different types of discourse prompts.

[orcid=0000-0001-7243-6100]

[orcid=0000-0002-7554-6142]

[orcid=0000-0001-9255-2045]

[orcid=0000-0002-7937-5109]

Key Words: Aphasia, Stroke, Discourse, SVM,

1. Introduction

Brain injury via stroke or neurodegenerative disease can often result in aphasia, defined as impaired language and communication. Aphasia can lead to significant declines in quality of life and well-being (Bullier et al., 2020; Spaccavento et al., 2014), as the ability to communicate effectively is vital for interpersonal relationships, employment, and navigating the world. A major part of this decline can be related to impairments in spoken discourse (Galski, Tompkins, & Johnston, 1998). Spoken discourse provides a wealth of information about linguistic ability that is related to aphasia severity. Hence, evaluation of discourse in persons with aphasia has gained increasing recognition for clinical assessment and treatment (Bryant, Ferguson, & Spencer, 2016a; Stark & Fukuyama, 2021). The majority of current aphasia assessments, such as the Western Aphasia Battery (WAB; (Kertesz, 1982, 2007)) are rigorous but relatively demanding standardized tests that can be burdensome for survivors of stroke, their families, and clinicians. In the United States, it is often difficult for people to even be approved or financially supported for comprehensive baseline language evaluations post-stroke (Walker et al., 2022). Hence, supplementary assessments that are brief but comparable can be valuable. If reliable, such assessments could be used for triage purposes, measuring change in language abilities over time, or for individuals who have limited access to healthcare resources (e.g., rural or impoverished). In this context, discourse analysis is a promising line of research, given the rich set of microstructural (lexical-semantic, syntactic) and macrostructural (cohesion, coherence) elements in discourse.

Compared to a multi-hour standardized test, eliciting discourse is more tractable for a non-specialist, thanks to resources such as AphasiaBank (Fromm, Forbes, Holland, & MacWhinney, 2020; Macwhinney, Fromm, Forbes, & Holland, 2011). Tasks include a description of a sequence of pictures (Broken Window), narrative discourse without visual aids ('tell me the story of Cinderella'), and procedural ('tell me how to make a peanut butter and jelly sandwich). The tasks are brief (< 5 minutes) and data collection could be done remotely via mobile phone applications or wearable monitors. The prompts allow for more continuous and naturalistic output than other language assessments such as confrontation naming, sentence-picture matching, or production of isolated sentences. The variety of prompts (e.g., picture description vs. procedural) also allows for the inspection of relationships between linguistic and more domain-general cognitive processes such as procedural or episodic memory (Stark, 2019). These unique demands mean that discourse samples can measure language loss or recovery in a more naturalistic way than long-form standardized tests (Bryant, Ferguson, & Spencer, 2016b). The most time-intensive aspect of discourse analysis is transcription and coding, requiring 6-12 minutes of time per minute of collected discourse (Boyles, 1998). However, recent advances in computerized transcription and natural language processing are likely to aid automated transcription and coding in the coming years (S. G. Dalton et al., 2022; Jacks, Haley, Bishop, & Harmon, 2019). Some work has been done to develop at-home aphasia screenings such as the mobile aphasia screening test (Choi, Park, Ahn, Son, & Paik, 2015) or others designed for detection of paraphasia (Le, Licata, & Provost, 2017) or primary progressive aphasia (PPA; (Fraser et al., 2014)). However, these have largely been designed with the binary goal of detecting the

presence or absence of aphasia or classifying subtypes of PPA, instead of assessing the entire range of aphasia severity.

There is a growing body of research that uses discourse as an outcome measure of therapy (Bryant et al., 2016b), usually focusing on macro-level measures of fluency or information content. However, there are relatively few studies that address how the multitude of linguistic features elicited by discourse map onto aphasia severity. Some work has been done to investigate higher-level conceptual properties (macrostructure) of spoken discourse, such as main concept production and informativeness metrics. These studies have found that people with aphasia tend to produce less informative speech and that different aphasia subtypes have differing levels of main concept production (S. G. H. Dalton & Richardson, 2019). Other work has focused more on discourse microstructure. For example, Stark (2019) quantitatively established that the different discourse prompt types (e.g., picture description, narrative, and procedural) tend to tax different aspects of the language system in both controls and people with aphasia. For example, narrative discourse was found to elicit the most content-rich speech. Procedural discourse, on the other hand, elicited the lowest syntactic complexity. These findings suggest that using multiple prompt types may be important for discourse-based language assessment (see also Stark and Fukuyama (2021)). However, the vast majority of studies that have used discourse as an outcome measure have used a picture description prompt (Bryant et al., 2016b). Another difficulty of using discourse analysis is the complex, multidimensional, and collinear relationships between microstructural variables and overall language ability. For this reason, dimension reduction, multivariate, or machine learning methods may be well-suited for discourse analysis over traditional univariate methods (Stark & Fukuyama, 2021).

Here, we used picture-based, narrative, and procedural discourse tasks in a group of 71 stroke survivors with available structural neuroimaging scans. Our first aim was to quantitatively establish whether microstructural discourse analysis can accurately capture aphasia severity, as measured by the WAB Aphasia Quotient (AQ). A second aim was to determine if the three prompt types differed in elicitation of discourse features that predict AQ, and which features were the most predictive in each case. Previous research suggests that narrative discourse may be the most promising (Stark, 2019), but picture-based prompts are the most common discourse measure used in clinical studies (Bryant et al., 2016b). Finally, we examined how the inclusion of lesion features impacts model performance. We used Support Vector Regression (SVR) to predict AQ and assess feature importance.

2. Materials and Methods

2.1 Participants

Speech recordings were obtained from 71 unilateral left-hemisphere chronic (>12 months post-stroke, mean = 60 months, range = 12 – 237) stroke survivors by the Center for Study of Aphasia Recovery (C-STAR), as part of a multi-day data collection

battery (see Spell et al. (2020)). Participants were a mean of 61.7 years old (range = 29 – 80). This battery included structural and functional neuroimaging, administration of the WAB by licensed speech-language pathologists, discourse collection, and other cognitive and language testing. Mean WAB score was 65.9 (range = 14.5 – 100). Among these participants, 10 did not suffer from aphasia, while the rest had different types of aphasia: Broca's (28), Anomic (14), Conduction (11), Global (4), Wernicke's (3), and Transcortical Motor (1). All participants signed informed consent, and the research was approved by the University of South Carolina Institutional Review Board.

2.2 Behavioral Data

At intake, each participant was prompted by a clinician to narrate the Cinderella story, describe how to make a peanut butter and jelly (PBJ) sandwich, and explain the sequence of events shown a picture, referred to as Broken Window, according to AphasiaBank prompt directions (Macwhinney et al., 2011). Their discourse was video recorded. Videos were manually transcribed and coded by trained research assistants under the supervision of licensed speech language pathologists (for full details and reliability metrics, see Spell et al. (2020)). Using Computerized Language Analysis software (MacWhinney, 2000), various discourse features, as shown in Table 1 were extracted.

2.3 MRI data acquisition and preprocessing

MRI data were obtained with a Siemens 3T Trio System with a 12-channel head coil and a Siemens 3T Prisma System with a 20-channel coil. Participants underwent two anatomical MRI sequences: (i) T1-weighted imaging sequence with a magnetization-prepared rapid-gradient echo (MPRAGE) turbo field echo (TFE) sequence with voxel size = 1 mm3, field of view (FOV) = 256 × 256 mm, 192 sagittal slices, 9° flip angle, repetition time (TR) = 2,250 ms, inversion time (TI) = 925 ms, echo time (TE) = 4.15 ms, generalized autocalibrating partial parallel acquisition (GRAPPA) = 2, and 80 reference lines; and (ii) T2-weighted MRI with a 3D sampling perfection with application optimized contrasts by using different flip angle evolutions protocol with the following parameters: voxel size = 1 mm3, FOV = 256 × 256 mm, 160 sagittal slices, variable flip angle, TR = 3,200 ms, TE = 212 ms, and no slice acceleration. The same slice center and angulation were used as in the T1 sequence.

Lesions were defined in native space by a neurologist in MRIcron (Rorden, Bonilha, Fridriksson, Bender, & Karnath, 2012) on individual T2-weighted images. Preprocessing started with coregistration of the T2-weighted images to match the T-weighted images, allowing the lesions to be aligned to native T1 space. Images were warped to standard space using enantiomorphic (Nachev, Coulthard, Jager, Kennard, & Husain, 2008) segmentation-normalization (Ashburner & Friston, 2005) custom Matlab script (https://github.com/rordenlab/spmScripts/blob/master/nii_enat_norm.m) to warp images to an age-appropriate template image found in the Clinical Toolbox for SPM (https://www.nitrc.org/scm/?group_id=881). The normalization parameters were used to reslice the lesion into standard space using linear interpolation, with subsequent lesion maps stored at 1 × 1 × 1-mm resolution and binarized using a 50% threshold. (Because interpolation can lead to fractional probabilities, this step confirms that each voxel is

categorically either lesioned or unlesioned without biasing overall lesion volume.) Normalized images were visually inspected to verify quality.

2.4 Lesion feature extraction

The resulting images were parcellated according to the Johns Hopkins University atlas (Faria et al., 2012; Mori, Wakana, Van Zijl, & Nagae-Poetscher, 2005; Wakana, Jiang, Nagae-Poetscher, van Zijl, & Mori, 2004). For each participant, the percent of voxels damaged within each of these regions was calculated, and areas that were undamaged in all participants were removed from further analysis, resulting in 64 lesion features considered in this study (Supplementary Materials).

Table 1 *The list of discourse features extracted for each of the prompts. The last 16 features starting from Nouns to WordErrors are included as both absolute numbers and relative percentages, amounting to a total of 45 discourse features per prompt.*

Name	Description
Duration	Total duration of discourse (sec)
Total Utts	Total utterances
MLU Utts	Total #utterances for calculating MLU below
MLU Words	Mean number of words per utterance
MLU Morphs	Mean number of morphemes per utterance
FreqTypes	Number of word types used
FreqTokens	Number of unique words used
FreqTTR	Ratio of types to tokens
Words/Min	Words per minute
Verbs/Utt	Number of verbs per utterance
Density	Propositional idea density
Retracing	Number of self-corrections during speech
Repetition	Number of word repetitions
Nouns	Words that were nouns
Prep	Words that were prepositions
Adj	Words that were adjectives
Adv	Words that were adverbs
Conj	Words that were conjunctions
Det/Art	Words that were determiners or articles
Pro	Words that were pronouns
Aux	Words that were auxiliaries
Verbs	Words that were verbs
3S	Verbs that were 3rd person singular
1S/3S	Verbs with same form for first/third person

Name	Description
Past	Verbs that were past tense
PastPart	Verbs that were past participles
PresPart	Verbs that were present participles
Plurals	Nouns that were plural
WordErrors	Words that had some sort of error

Our goal was to predict the AQ of a participant based on their discourse and/or lesion features with the help of machine learning. Two key design choices in developing a machine learning system are the learning algorithm and the feature set. We chose linear Support Vector Machines (SVM) which is a popular machine learning method that is known to perform well on relatively small datasets, (Mahmoud et al. 2021) and is resistant to overfitting. An appropriate subset of given features was selected through recursive feature elimination and cross-validation. Specifically, we used leave-one-out (LOO) to split the participants into a set of 70 for training and 1 for testing. Using the 70 samples in the training set, all the features were ranked using recursive feature elimination. We then selected a combination of top features through cross-validation as follows. By employing LOO again, the training set of 70 samples was further split into 69 for training and 1 for validation. By training the SVM on the 69 samples, we predicted the AQ for the one in the validation set. This is done 70 times with each participant in the validation set once. The predicted AQ values were compared against the true AQ values to compute an R^2 score. This process was repeated for each combination of top-k features, with k limited to 10. When the features are highly correlated, as in the current study, a feature set close to the square root of the sample size is often ideal for SVM (Hua, Xiong, Lowey, Suh, & Dougherty, 2005), and the inclusion of too many features in a relatively small sample can lead to overfitting. The feature combination with the highest R^2 score was then used to train the model with 70 samples to predict the AQ for the one in the test set. We capped predicted AQ values at 100, and set the minimum to 20. Observed AQs below 20 are exceedingly rare in clinical studies (Walker et al., 2022), and differences in numerical AQ below this cutoff are unlikely to be clinically relevant. This process was repeated with each participant in the test set once. Note that with this procedure, we avoid 'peeking', and no information about the left-out participant is used for feature selection. We then computed Pearson's correlation (r), root mean squared error (RMSE), and mean absolute error (MAE) between predicted AQ and observed AQ to evaluate estimation accuracy. The SVM model uses a hyper-parameter C for regularization. We varied C from 0.01 to 100 and chose the C that yielded the highest correlation coefficient.

We conducted analyses with (1) Discourse features only from each of the three prompts individually, and the combined set of features from the three prompts, (2) Lesion features combined the discourse features in (1), and (3) lesion features only.

3. Results

Table 2 and Figures 1-3 summarize the results. For each prompt, and for all sets of features (discourse only, discourse+lesion, lesion only), the correlation between predicted and actual AQ was significant (all p's < 0.001, Pearson's r range 0.64 – 0.83).

Table 2 Results summary for predicted AQ compared to observed AQ for each model.

Prompt	Discourse	Discourse + Lesion	Lesion Only
Broken Window	r = 0.78 RMSE = 14.63 MAE = 11.50	r = 0.79 RMSE = 14.59 MAE = 10.94	~
Cinderella	r = 0.75 RMSE = 15.50 MAE = 12.14	r = 0.83 RMSE = 12.94 MAE = 9.53	~
PBJ	r = 0.7 RMSE = 16.75 MAE = 13.47	r = 0.72 RMSE = 16.56 MAE = 12.95	~
All Combined	r = 0.83 RMSE = 13.09 MAE = 9.77	r = 0.82 RMSE = 13.29 MAE = 10.03	r = 0.64 RMSE = 18.53 MAE = 14.69

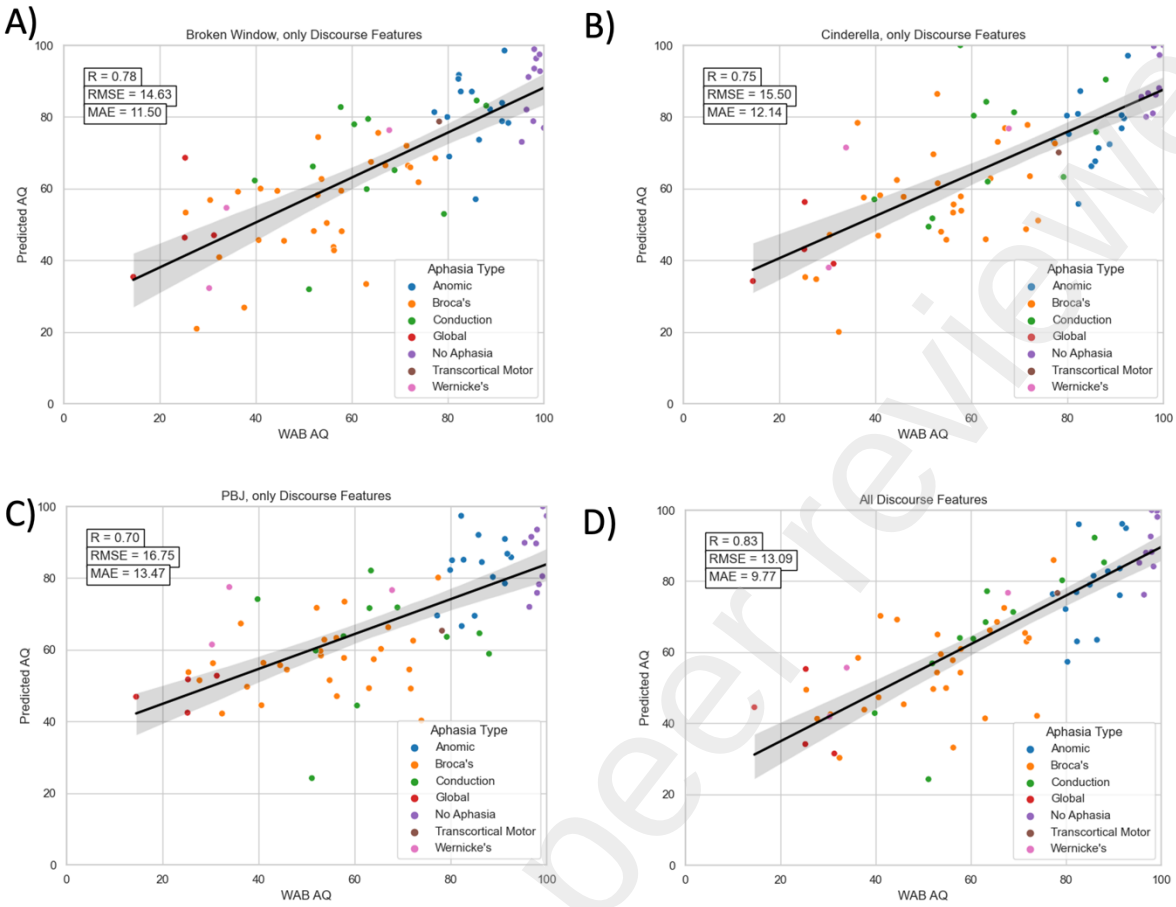

Figure 1 AQ prediction using only discourse features; A) Broken Window, B) Cinderella, C) PBJ, and D) all features combined.

We used a two-tailed Hotelling's t-test for dependent correlations (Weiss, 2011) to examine whether any of the models were significantly better at predicting AQ. This test compares the Pearson's r between predicted and observed AQ for a given pair of models (e.g., Cinderella vs. Broken Window), while considering that the values come from the same group of participants. All features combined was trending towards more accurate predicted AQs than Cinderella ($t(68) = -1.905$, $p = 0.06$), and all features combined was significantly better than PBJ ($t(68) = 2.89$, $p = .005$). No other pairwise tests were significant or trending.

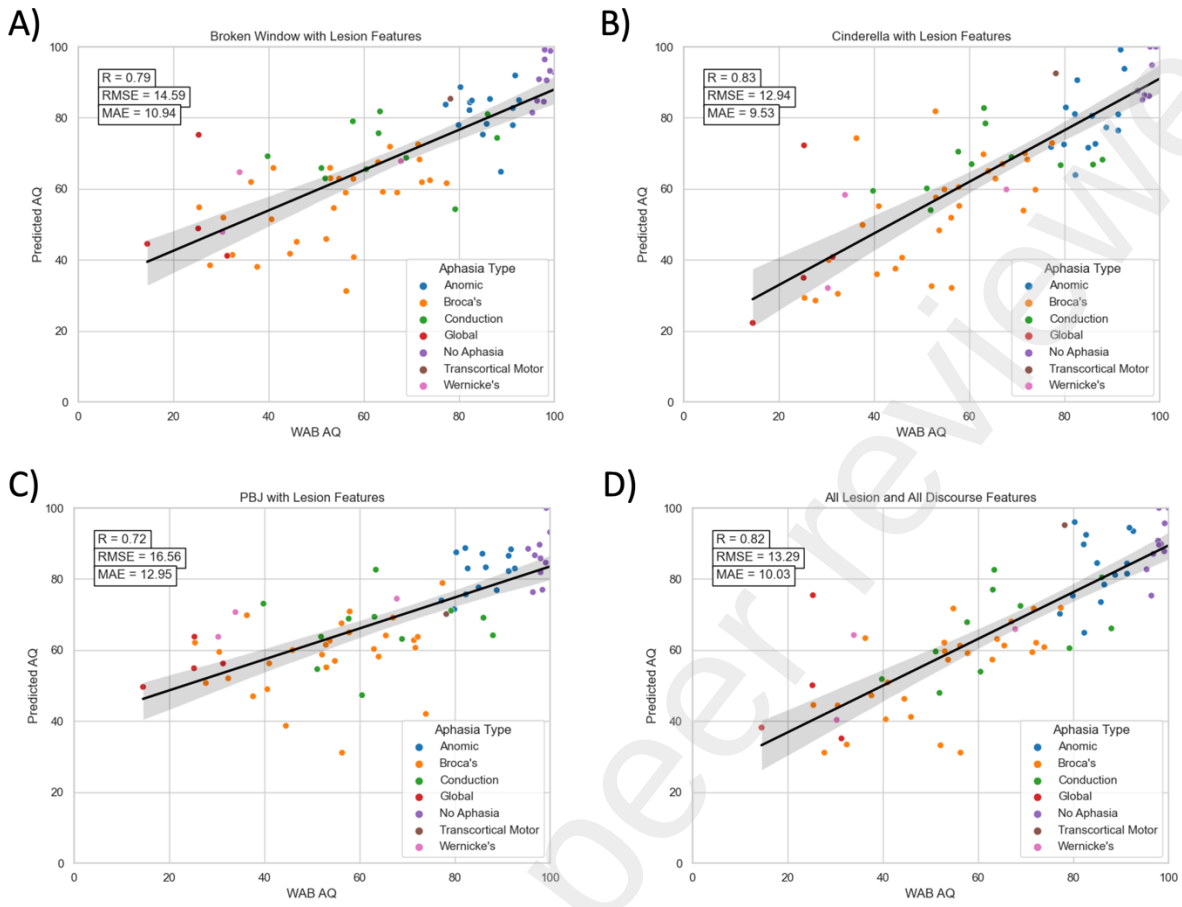

Figure 2 *AQ prediction using discourse plus lesion features*

When lesion features were added to discourse features, the performance for BrokenWindow, PBJ, and all features combined was not altered significantly. However, lesion features significantly boosted the performance of Cinderella, as determined by a Hotelling's t-test for dependent correlations ($t(68) = -2.05$, $p = .04$).

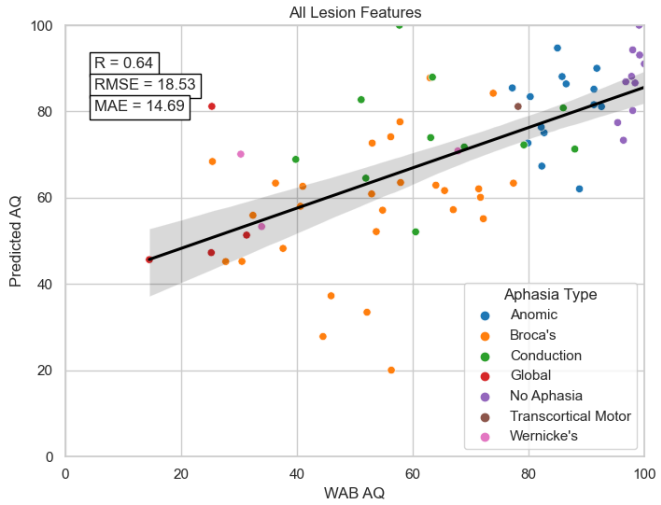

Figure 3 *AQ prediction using only lesion features.*

The lesion-feature only model was significantly outperformed by all models except for PBJ (with or without lesion) and Cinderella without lesion features (all p's < .05).

We also calculated the 10 most informative features for each model (Table 3). Inspecting these features allows us to examine how informative linguistic features change depending on the prompt type.

Table 3 *Top 10 features for each model. In parenthesis, the first number is the percent of times that feature was chosen as a top 10 (across 71 LOO cross validation folds), and the second number is the median rank that feature had. E.g., FreqTypes (100, 1) means that the number of different types of words used was a top 10 feature 100% of the time and had a median importance rank of #1.*

Broken Window	Cinderella	PBJ	All Prompts Combined
FreqTypes (100, 2)	FreqTypes (100, 1)	% Prep (100, 1)	% Prep-PBJ (100, 2)
# WordErrors (100, 3)	% PastPart (100, 2)	MLU Morphs (100, 2)	# WordErrors-BW (82, 3)
# Adv (100, 3)	% Past (100, 3)	# Prep (100, 5)	# Nouns-BW (70, 1)
Density (100, 4)	% Nouns (100, 4)	% 3S (100, 8)	% Nouns-Cind (70, 6)
MLU Utts (86, 7)	% PresPart (97, 6)	# Nouns (99, 3)	% Conj-BW (66, 9)
# Nouns (83, 7)	Words/Min (92, 5)	Words/Min (99, 6)	MLU Utts-BW (59, 7)
% Det/Art (72, 5)	# Repetition (85, 9)	% Nouns (94, 6)	% Past-BW (51, 10)

Broken Window	Cinderella	PBJ	All Prompts Combined
# Adj (62, 8)	% Det/Art (80, 8)	MLU Words (77, 6)	% Det/Art-PBJ (48, 10)
% Conj (54, 9)	# PresPart (70, 7)	Verbs/Utt (49, 10)	MLU Utts-Cind (38, 12)
% Nouns (34, 13)	MLU Morphs (56, 10)	% Past (49, 10)	% Det/Art-BW (34, 14)

Lesion Only	+ Broken Window	+ Cinderella	+ PBJ	+ All Prompts Combined
SLF (100, 1)	# Nouns (100, 1)	% Nouns (100, 3)	SLF (100, 1)	% Nouns-Cind (100, 4)
EC (100, 2)	SLF (100, 2)	% PastPart (100, 4)	MLU Morphs (100, 2)	FreqTypes-Cind (99, 2)
MOG (100, 3)	# WordErrors (100, 3)	FreqTypes (97, 1)	% Prep (100, 3)	% Prep-PBJ (97, 3)
SCC (99, 4)	Density (100, 4)	MFOG (97, 5)	# Nouns (100, 4)	SLF (94, 1)
FUG (99, 5)	% WordErrors (100, 7)	SLF (90, 2)	MOG (100, 8)	PSTG (72, 8)
BCC (79, 10)	EC (99, 5)	% Past (89, 8)	RLIC (99, 7)	# WordErrors-BW (61, 6)
RLIC (77, 6)	# Repetition (93, 9)	% Det/Art (87, 6)	Words/Min (97, 6)	# Nouns-BW (61, 8)
SCR (77, 9)	% Det/Art (80, 6)	PSTG (85, 7)	MLU Words (97, 8)	MFOG (54, 9)
LF (75, 8)	PSTG (54, 10)	CAUD (59, 9)	% WordErrors (96, 6)	% Past-Cind (51, 10)
MFOG (68, 7)	IFO (52, 10)	# WordErrors (30, 11)	# Prep (39, 10)	% Det/Art-BW (25, 18)

4. Discussion

Here, we used linguistic features extracted from discourse analysis to quantify aphasia severity. Using only microstructural (i.e., lexical and grammatical) features, we were able to build models that provided person-specific aphasia severity estimates, with predicted AQ scores being significantly correlated with observed AQ. We also investigated which discourse or lesion features are most predictive of AQ. From these

top features, we can draw conclusions about: 1) which linguistic (or lesion) features are most important for estimating aphasia severity, and 2) differences in how the language system is taxed by different prompt types.

4.1 Prompt types and discourse features

Using only discourse features, Broken Window had the highest prediction accuracy, while PBJ was numerically worse. Although this difference was not significant, it aligns well with the findings of Stark (2019), who suggested that PBJ has lower syntactic demands than narrative or picture description tasks, even when inspecting discourse output from healthy adults. These lower demands may result in not adequately taxing certain syntactic aspects of language, especially related to verb production, that are captured by AQ and other discourse tasks. The top features for the PBJ model demonstrate that it captures somewhat different linguistic properties than Broken Window and Cinderella, especially related to the use of prepositions, which turned out to be the only feature selected 100% of the time when combining all discourse features into a single model. This result is consistent with Stark and Fukuyama (2021), who found that prepositions were one of the main features that separated PBJ from other prompt types when examining discourse output using between-class analysis, a dimension reduction technique. Although not explicitly tested here, these unique prepositional demands may be especially useful in evaluating agrammatic people who are able to use content, but not function, words. The WAB, especially in its fluency subscores, sometimes has difficulty appropriately evaluating agrammatic people who can respond to prompts with simple noun-verb phrases. This results in low inter-rater reliability for this subtest, with nonfluent people sometimes obtaining inflated fluency scores (Clough & Gordon, 2020; Trupe, 1984).

Some patterns emerged relating to the task demands of picture sequence description in the Broken Window prompt. First, the total number of different word types used was the most important feature, demonstrating that it encourages participants to display their general mastery of language by eliciting the use of different types of words (S. G. Dalton & Richardson, 2015). This is perhaps also reflected in the importance of the adverbs feature, as adverbs are not ‘necessary’ per-se when describing a sequence of pictures. Instead, using more adverbs likely reflects an optional level of specificity that measures linguistic competence (Sarno, Postman, Cho, & Norman, 2005), thus aiding model prediction. Propositional density – a measure of content richness - being chosen in 100% of Broken Window models was somewhat surprising, as past research has shown that picture description tasks elicit the lowest amount of content richness of the 3 task types (Stark, 2019), even in healthy adults. However, its inclusion in the model suggests that, while picture description may not elicit particularly rich content, its presence in an individual person’s discourse sample provides information about their aphasia severity. Finally, word errors were also important for Broken Window models, reflecting the naming processes elicited by picture description tasks (e.g., naming various objects or characters in the picture, also called a core lexicon (S. G. Dalton & Richardson, 2015)).

Similar to Broken Window, the most important feature elicited by Cinderella for predicting AQ was the number of different word types used. However, results suggest that the usage of the past tense is what separates Cinderella from the other prompt

types, with past tense and past participle use being among the most important features for predicting AQ in Cinderella-based models. Several studies have found that production and comprehension of past tense can be especially difficult for people with aphasia (Faroqi-Shah & Friedman, 2015; Jonkers & de Bruin, 2009; Ullman et al., 2005). Cinderella, a narrative recall task, forces participants to use the past tense in their retelling of the Cinderella story, while Broken Window and PBJ can be completed using the present tense.

The findings demonstrate that, while each prompt can be used to predict AQ, the top ten features differ – providing insights about the unique linguistic demands of each prompt type. Indeed, combining all features from all prompts into a single model yielded numerically the highest AQ prediction accuracy, suggesting that the prompts make unique contributions to aphasia severity estimation. However, this comes with the drawback that there is less consistency among the top ten features chosen (evidenced by lower percentages and more variable median ranks for the top ten features), due to the expanded feature selection space. This could be ameliorated by simply using each discourse model individually, and then averaging the predicted AQ's together for each participant. This maintains top ten feature consistency from each prompt type, while also allowing each prompt to contribute to prediction.

4.2 Lesion features and aphasia severity

We also investigated the relative importance of the lesion features in AQ assessment. The superior longitudinal fasciculus (SLF) was the most frequently selected top ranked feature when only lesion features was used. Moreover, SLF is among the most frequently selected top two features even in the discourse plus lesion models. The SLF is a white matter tract that connects portions of the occipital, posterior temporal, and parietal lobes to the frontal cortex (Bernal & Altman, 2010; Kamali, Flanders, Brody, Hunter, & Hasan, 2014). Our finding that the SLF is an important feature for predicting aphasia severity aligns with previous research demonstrating that degradation of the SLF in a variety of etiologies has been linked to impaired language or executive abilities that contribute to language (Madhavan, McQueeny, Howe, Shear, & Szaflarski, 2014; Nagae et al., 2012; Rizio & Diaz, 2016; Shinoura et al., 2013).

The other features chosen 100% of the time as a top 10 feature in the lesion-only model, the external capsule (EC) and middle occipital gyrus (MOG), are somewhat surprising as they are not considered classic 'language areas' in most neurobiological models (Desai & Riccardi, 2021; G. Hickok & Poeppel, 2004). However, EC integrity has been implicated in executive dysfunction (Nolze-Charron et al., 2020), and is considered by some to be a part of the ventral language stream (Axer, Klingner, & Prescher, 2013), although this is debated. EC tracts are also adjacent to portions of SLF (Schmahmann, Schmahmann, & Pandya, 2009), raising the possibility that these two pathways are commonly damaged together in stroke affecting the middle cerebral artery. It is also possible that the EC contributes to language via subcortical connections that support language either directly or through domain-general processes (Kuljic-Obradovic, 2003; Sharif, Goldberg, Walker, Hillis, & Meier, 2022). The MOG, on the other hand, may be related to visual identification of items and objects near the 'beginning' of the ventral language stream (Fridriksson et al., 2016; G. Hickok &

Poeppel, 2004; Gregory Hickok & Poeppel, 2016). People with MOG damage likely perform poorly on visual aspects of the WAB such as object naming or picture description, making it an informative feature when predicting aphasia severity.

While it is somewhat surprising that adding lesion features to the discourse models did not boost accuracy of aphasia severity estimation, it demonstrates the effectiveness of discourse task in estimating AQ. The discourse prompts require many of the same language skills that are measured by WAB (indeed, WAB even includes a picture description task), which was sufficient for discourse tasks to have a high predictive value. Lesion features, on the other hand, are comparatively more 'indirect' representatives of language ability. When considering future use of discourse features to estimate aphasia severity, it is a net positive that lesion features do not contribute significantly above and beyond discourse features. If discourse features alone could not estimate aphasia severity and MRI scans were required, then it would negate the advantage of the discourse method as less demanding in terms of resources.

4.3 Limitations and future directions

Here, our focus was on using purely microlinguistic discourse features to estimate aphasia severity. A promising extension may be to investigate how adding macro-level features such as main concept analysis or demographic features could boost model prediction (Johnson et al., 2022). Regarding anatomical features, it is possible that other measures of brain health, such as resting state connectivity (Kristinsson et al., 2021) or brain age (Busby et al., 2023; Kristinsson et al., 2022) could also be useful estimators of aphasia severity. Furthermore, clinical use of discourse-based aphasia severity estimation relies on improved automation of transcription and coding of impaired speech in the coming years, as current automated transcription methods perform relatively poorly in people with aphasia (Mahmoud et al., 2023). Future work could also investigate how, instead of prompts, other naturalistic discourse paradigms could be used to assess language abilities and their neural correlates (Birba et al., 2022; Riccardi & Desai, 2022). Finally, in the current study, even though the prediction was accurate overall with the Pearson's correlation between measured and predicted AQ near 0.8, and the models were highly accurate for majority of the participants, they were relatively inaccurate for a handful of cases. Understanding the characteristics of individuals that lead to lower model prediction performance may help improve models even further.

5. Conclusion

The present study showed that microlinguistic features elicited from three AphasiaBank discourse prompts can be used to estimate aphasia severity. Even a single prompt, containing only a few minutes (or sometimes less than a minute) of speech output, was sufficient to estimate AQ reasonably well for most individuals. Each prompt elicited different informative features, demonstrating potential differences between prompts. Lesion features can also be used to estimate aphasia severity, although with lower accuracy than the discourse-based models. An important role for superior longitudinal fasciculus integrity in aphasia severity is suggested. Discourse-based aphasia severity estimation is promising as a supplemental language measurement that is ecologically

valid and less resource-intensive. The current study provides important first steps towards mapping how discourse features can quantify aphasia severity.

Acknowledgments

This work was supported by R01DC017162, R56DC010783, and R01DC010783 to RHD and P50DC014664 to JF.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage, 26*(3), 839-851. doi:10.1016/j.neuroimage.2005.02.018

Axer, H., Klingner, C. M., & Prescher, A. (2013). Fiber anatomy of dorsal and ventral language streams. *Brain Lang, 127*(2), 192-204. doi:10.1016/j.bandl.2012.04.015

Bernal, B., & Altman, N. (2010). The connectivity of the superior longitudinal fasciculus: a tractography DTI study. *Magn Reson Imaging, 28*(2), 217-225. doi:10.1016/j.mri.2009.07.008

Birba, A., Fittipaldi, S., Cediel Escobar, J. C., Gonzalez Campo, C., Legaz, A., Galiani, A., . . . Garcia, A. M. (2022). Multimodal Neurocognitive Markers of Naturalistic Discourse Typify Diverse Neurodegenerative Diseases. *Cereb Cortex, 32*(16), 3377-3391. doi:10.1093/cercor/bhab421

Boyles, L. (1998). Conversational discourse analysis as a method for evaluating progress in aphasia: A case report. *Journal of communication disorders, 31*, 261-274.

Bryant, L., Ferguson, A., & Spencer, E. (2016a). Linguistic analysis of discourse in aphasia: A review of the literature. *Clin Linguist Phon, 30*(7), 489-518. doi:10.3109/02699206.2016.1145740

Bryant, L., Ferguson, A., & Spencer, E. (2016b). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics, 30*(7), 489-518.

Bullier, B., Cassoudesalle, H., Villain, M., Cogne, M., Mollo, C., De Gabory, I., . . . Glize, B. (2020). New factors that affect quality of life in patients with aphasia. *Ann Phys Rehabil Med, 63*(1), 33-37. doi:10.1016/j.rehab.2019.06.015

Busby, N., Wilmskoetter, J., Gleichgerrcht, E., Rorden, C., Roth, R., Newman-Norlund, R., . . . Bonilha, L. (2023). Advanced Brain Age and Chronic Poststroke Aphasia Severity. *Neurology, 100*(11), e1166-e1176. doi:10.1212/WNL.0000000000201693

Choi, Y., Park, H. K., Ahn, K., Son, Y., & Paik, N. (2015). A telescreening tool to detect apahsia in patients with stroke. *Telemedicine and e-Health, 21*(9), 729-734.

Clough, S., & Gordon, J. K. (2020). Fluent or nonfluent? Part A. Underlying contributors to categorical classifications of fluency in aphasia. *Aphasiology, 34*(5), 515-539.

Dalton, S. G., & Richardson, J. D. (2015). Core-Lexicon and Main-Concept Production During Picture-Sequence Description in Adults Without Brain Damage and Adults With Aphasia. *Am J Speech Lang Pathol, 24*(4), S923-938. doi:10.1044/2015_AJSLP-14-0161

Dalton, S. G., Stark, B. C., Fromm, D., Apple, K., MacWhinney, B., Rensch, A., & Rowedder, M. (2022). Validation of an Automated Procedure for Calculating Core Lexicon From Transcripts. *J Speech Lang Hear Res, 65*(8), 2996-3003. doi:10.1044/2022_JSLHR-21-00473

Dalton, S. G. H., & Richardson, J. D. (2019). A Large-Scale Comparison of Main Concept Production Between Persons With Aphasia and Persons Without Brain Injury. *Am J Speech Lang Pathol, 28*(1S), 293-320. doi:10.1044/2018_AJSLP-17-0166

Desai, R. H., & Riccardi, N. (2021). Cognitive neuroscience of language. In *The Routledge handbook of cognitive linguistics* (pp. 615-642).

Faria, A. V., Joel, S. E., Zhang, Y., Oishi, K., van Zjil, P. C., Miller, M. I., . . . Mori, S. (2012). Atlas-based analysis of resting-state functional connectivity: evaluation for reproducibility and multi-modal anatomy-function correlation studies. *Neuroimage, 61*(3), 613-621. doi:10.1016/j.neuroimage.2012.03.078

Faroqi-Shah, Y., & Friedman, L. (2015). Production of verb tense in agrammatic aphasia: A meta-analysis and further data. *Behavioural neurology, 2015*.

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., & Rochon, E. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex, 55*, 43-60. doi:10.1016/j.cortex.2012.12.006

Fridriksson, J., Yourganov, G., Bonilha, L., Basilakos, A., Den Ouden, D. B., & Rorden, C. (2016). Revealing the dual streams of speech processing. *Proc Natl Acad Sci U S A, 113*(52), 15108-15113. doi:10.1073/pnas.1614038114

Fromm, D., Forbes, M., Holland, A., & MacWhinney, B. (2020). Using AphasiaBank for Discourse Assessment. *Semin Speech Lang, 41*(1), 10-19. doi:10.1055/s-0039-3399499

Galski, T., Tompkins, C., & Johnston, M. (1998). Competence in discourse as a measure of social integration and quality of life in persons with traumatic brain injury. *Brain injury, 12*(9), 769-782.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1-2), 67-99. doi:10.1016/j.cognition.2003.10.011

Hickok, G., & Poeppel, D. (2016). Neural basis of speech perception. *Neurobiology of language*, 299-310.

Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. R. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics, 21*(8), 1509-1515.

Jacks, A., Haley, K. L., Bishop, G., & Harmon, T. G. (2019). Automated speech recognition in adult stroke survivors: Comparing human and computer transcriptions. *Phoniatrica et Logopaedica, 71*(5-6), 286-296.

Johnson, L., Nemati, S., Bonilha, L., Rorden, C., Busby, N., Basilakos, A., . . . Fridriksson, J. (2022). Predictors beyond the lesion: Health and demographic factors associated with aphasia severity. *Cortex, 154*, 375-389. doi:10.1016/j.cortex.2022.06.013

Jonkers, R., & de Bruin, A. (2009). Tense processing in Broca's and Wernicke's aphasia. *Aphasiology, 23*(10), 1252-1265.

Kamali, A., Flanders, A. E., Brody, J., Hunter, J. V., & Hasan, K. M. (2014). Tracing superior longitudinal fasciculus connectivity in the human brain using high resolution diffusion tensor tractography. *Brain Struct Funct, 219*(1), 269-281. doi:10.1007/s00429-012-0498-y

Kertesz, A. (1982). *The Western Aphasia Battery*. New York: Grune and Stratton.

Kertesz, A. (2007). *Western Aphasia Battery-Revised*. San Antonio, TX.: Pearson.

Kristinsson, S., Busby, N., Rorden, C., Newman-Norlund, R., den Ouden, D. B., Magnusdottir, S., . . . Fridriksson, J. (2022). Brain age predicts long-term recovery in post-stroke aphasia. *Brain Commun, 4*(5), fcac252. doi:10.1093/braincomms/fcac252

Kristinsson, S., Zhang, W., Rorden, C., Newman-Norlund, R., Basilakos, A., Bonilha, L., . . . Fridriksson, J. (2021). Machine learning-based multimodal prediction of language outcomes in chronic aphasia. *Hum Brain Mapp, 42*(6), 1682-1698. doi:10.1002/hbm.25321

Kuljic-Obradovic, D. C. (2003). Subcortical aphasia: three different language disorder syndromes? *Eur J Neurol, 10*(4), 445-448. doi:10.1046/j.1468-1331.2003.00604.x

Le, D., Licata, K., & Provost, E. M. (2017). *Automatic Paraphasia Detection from Aphasic Speech: A Preliminary Study*. Paper presented at the Interspeech 2017.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs* (Vol. 1): Psychology Press.

Macwhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for Studying Discourse. *Aphasiology, 25*(11), 1286-1307. doi:10.1080/02687038.2011.589893

Madhavan, K. M., McQueeny, T., Howe, S. R., Shear, P., & Szaflarski, J. (2014). Superior longitudinal fasciculus and language functioning in healthy aging. *Brain Res, 1562*, 11-22. doi:10.1016/j.brainres.2014.03.012

Mahmoud, S. S., Pallaud, R. F., Kumar, A., Faisal, S., Wang, Y., & Fang, Q. (2023). A Comparative Investigation of Automatic Speech Recognition Platforms for Aphasia Assessment Batteries. *Sensors (Basel), 23*(2). doi:10.3390/s23020857

Mori, S., Wakana, S., Van Zijl, P. C., & Nagae-Poetscher, L. (2005). *MRI atlas of human white matter*: Elsevier.

Nachev, P., Coulthard, E., Jager, H. R., Kennard, C., & Husain, M. (2008). Enantiomorphic normalization of focally lesioned brains. *Neuroimage, 39*(3), 1215-1226. doi:10.1016/j.neuroimage.2007.10.002

Nagae, L. M., Zarnow, D. M., Blaskey, L., Dell, J., Khan, S. Y., Qasmieh, S., . . . Roberts, T. P. (2012). Elevated mean diffusivity in the left hemisphere superior longitudinal fasciculus in autism spectrum disorders increases with more profound language impairment. *AJNR Am J Neuroradiol, 33*(9), 1720-1725. doi:10.3174/ajnr.A3037

Nolze-Charron, G., Dufort-Rouleau, R., Houde, J. C., Dumont, M., Castellano, C. A., Cunnane, S., . . . Bocti, C. (2020). Tractography of the external capsule and cognition: A diffusion MRI study of cholinergic fibers. *Exp Gerontol, 130*, 110792. doi:10.1016/j.exger.2019.110792

Riccardi, N., & Desai, R. H. (2022). Discourse and the brain. In *The Routledge Handbook of Semiosis and the Brain* (pp. 174-189).

Rizio, A. A., & Diaz, M. T. (2016). Language, aging, and cognition: frontal aslant tract and superior longitudinal fasciculus contribute toward working memory performance in older adults. *Neuroreport, 27*(9), 689-693. doi:10.1097/WNR.0000000000000597

Rorden, C., Bonilha, L., Fridriksson, J., Bender, B., & Karnath, H. O. (2012). Age-specific CT and MRI templates for spatial normalization. *Neuroimage, 61*(4), 957-965. doi:10.1016/j.neuroimage.2012.03.020

Sarno, M. T., Postman, W. A., Cho, Y. S., & Norman, R. G. (2005). Evolution of phonemic word fluency performance in post-stroke aphasia. *J Commun Disord, 38*(2), 83-107. doi:10.1016/j.jcomdis.2004.05.001

Schmahmann, J. D., Schmahmann, J., & Pandya, D. (2009). *Fiber pathways of the brain*: OUP USA.

Sharif, M. S., Goldberg, E. B., Walker, A., Hillis, A. E., & Meier, E. L. (2022). The contribution of white matter pathology, hypoperfusion, lesion load, and stroke recurrence to language deficits following acute subcortical left hemisphere stroke. *PLOS ONE, 17*(10), e0275664. doi:10.1371/journal.pone.0275664

Shinoura, N., Midorikawa, A., Onodera, T., Tsukada, M., Yamada, R., Tabei, Y., . . . Yagi, K. (2013). Damage to the left ventral, arcuate fasciculus and superior longitudinal fasciculus-related pathways induces deficits in object naming, phonological language function and writing, respectively. *Int J Neurosci, 123*(7), 494-502. doi:10.3109/00207454.2013.765420

Spaccavento, S., Craca, A., Del Prete, M., Falcone, R., Colucci, A., Di Palma, A., & Loverre, A. (2014). Quality of life measurement and outcome in aphasia. *Neuropsychiatr Dis Treat, 10*, 27-37. doi:10.2147/NDT.S52357

Spell, L. A., Richardson, J. D., Basilakos, A., Stark, B. C., Teklehaimanot, A., Hillis, A. E., & Fridriksson, J. (2020). Developing, Implementing, and Improving Assessment and Treatment Fidelity in Clinical Aphasia Research. *Am J Speech Lang Pathol, 29*(1), 286-298. doi:10.1044/2019_AJSLP-19-00126

Stark, B. C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: Implications for language assessment and outcome. *American Journal of Speech-Language Pathology, 28*(3), 1067-1083.

Stark, B. C., & Fukuyama, J. (2021). Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. *Language, Cognition and Neuroscience, 36*(5), 562-585.

Trupe, E. H. (1984). Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification. *Clinical Aphasiology: Proceedings of the Conference 1984*.

Ullman, M. T., Pancheva, R., Love, T., Yee, E., Swinney, D., & Hickok, G. (2005). Neural correlates of lexicon and grammar: evidence from the production, reading, and judgment of inflection in aphasia. *Brain Lang, 93*(2), 185-238; discussion 239-142. doi:10.1016/j.bandl.2004.10.001

Wakana, S., Jiang, H., Nagae-Poetscher, L. M., van Zijl, P. C., & Mori, S. (2004). Fiber tract-based atlas of human white matter anatomy. *Radiology, 230*(1), 77-87. doi:10.1148/radiol.2301021640

Walker, G. M., Fridriksson, J., Hillis, A. E., Den Ouden, D. B., Bonilha, L., & Hickok, G. (2022). The Severity-Calibrated Aphasia Naming Test. *American Journal of Speech-Language Pathology, 31*(6), 2722-2740.

Weiss, B. A. (2011). Hotelling's t Test and Steiger's Z test calculator. Retrieved from https://blogs.gwu.edu/weissba/teaching/calculators/hotellings-t-and-steigers-z-tests/