A Preliminary Evaluation of Semantic Coherence and Cohesion in Aphasic and Non-Aphasic Discourse Across Test and Retest

Snigdha Khanna, Brielle Caserta Stark
Indiana University Bloomington
snkhanna@iu.edu, bcstark@indiana.edu

Abstract

This paper evaluates global and local semantic coherence in aphasic and non-aphasic discourse tasks using the Tool for the Automatic Analysis of Cohesion (TAACO). The motivation for this paper stems from a lack of automatic methods to evaluate discourse-level phenomena, such as semantic cohesion, in transcripts derived from persons with aphasia. It leverages existing test-retest data to evaluate two main objectives: (1) Test-Retest Reliability, to identify if variables significantly differ across test and retest time points for either group (aphasia, control), and (2) Inter-Group Discourse Cohesion, where aphasic discourse is expected to be less cohesive than control discourse, resulting in lower cohesion scores for the aphasia group. Exploratory analysis examines correlations between variables for both groups, identifying any relationships between word-level and sentence-level semantic variables. Results verify that semantic cohesion and coherence are generally preserved in both groups, except for word-level and a few sentence-level semantic measures,w which are higher for the control group. Overall, variables tend to be reliable across time points for both groups. Notably, the aphasia group demonstrates more variability in cohesion than the control group, which is to be expected after brain injury. A close relationship between word-level indices and other indices is observed, suggesting a disconnection between word-level factors and sentence-level metrics.

Keywords: semantics, coherence, cohesion, aphasia, discourse, automatic scoring, correlation

1. Introduction

Spoken discourse, which is verbal language beyond a single sentence elicited for a specific purpose, is a compelling way of evaluating linguistic, propositional, macrostructural, and pragmatic aspects of language. This has been especially true in populations with typical speech and language, but the evaluation of spoken discourse has occurred less commonly in clinical populations. Yet, spoken discourse is a sensitive way of evaluating impairments arising at each level (i.e., linguistic, propositional, macrostructural, pragmatic).

Discourse coherence is categorized into global and local coherence. Global coherence broadly refers to how discourse units maintain the overall topic. Researchers have examined global coherence in various populations, including individuals with neurogenic disorders like aphasia. Different methods have been developed to measure coherence ability, including rating scales, measures of coherence violations, and assessment of global coherence errors. Glosser and Deser (1992) developed a five-point rating scale to measure global coherence, focused on different types of cohesive ties, including appropriate closed class lexical cohesion (personal pronouns, demonstrative pronouns, and definite articles), appropriate open class lexical cohesion (repetitions, synonyms, superordinates, and subordinates), and incomplete cohesion. Cohesion was assessed by identifying occurrences of these cohesive ties within the preceding three verbalizations. Coherence, on the other hand, was evaluated based on raters' impressions of the overall meaning and content of the discourse, considering global and local coherence separately using a five-point rating scale. Other studies have used measures of coherence violations and degree of global coherence. Leer and Turkstra (1999) adapted the Glosser and Deser's scale for discourse samples from adolescents with brain injury. The mean global coherence score across discourse tasks was computed for participants. More recently, Wright et al. (2013) conducted a study aimed at determining the feasibility and validity of a four-point global coherence scale. The study used both the four-point scale and Glosser and Deser's five-point scale in storytelling discourse samples from cognitively healthy adults. Reliability estimates for both scales were high, indicating their effectiveness in measuring global coherence.

However, these existing methods have several limitations. Firstly, the reliance on rating scales introduces subjectivity and potential inter-rater variability. Secondly, manual rating scales are time-consuming and resource-intensive, hindering scalability in analyzing large datasets. Additionally, the limited granularity provided by manual scales restricts the depth of analysis, while the lack of standardization and replicability across studies hampers comparisons and meta-analyses. To address these limitations, the incorporation of automatic scoring of semantics in discourse offers potential solutions.

A recent study by Stark et al. (2023) on spoken discourse evaluated whether linguistic performance in individuals with and without aphasia was reliable in a short test-retest time frame (one week) and

Proceedings of the RaPID-5 @LREC-COLING 2024, pages 77–86

across several different tasks, including a picture description, picture sequence description, fictional narrative, and procedural narrative. The study provided data on test-retest as well as rater reliability of established and commonly used spoken discourse measures (e.g. mean length of utterance, correct information units, words per minute) in aphasia, across a battery of tasks. They found that across groups and tasks, rater reliability was excellent and that the lexical, informative, and fluency measures were most reliable when averaged across tasks, though measure reliability varied considerably by task. Further, they found that individuals without aphasia were not necessarily producing more reliable language than those with aphasia, though there was a small effect of aphasia severity and sample length (number of words per sample) on reliability.

As has been the case for most measures extracted from spoken discourse in aphasia Bryant et al. (2016), the Stark et al. (2023) study focused primarily on lexico-syntactic linguistic measures. Linguistic measures like words per minute and mean length of utterance are relatively easy to extract using automatic measures, though the measure that Stark et al. (2023) found to be most reliable in persons with and without aphasia, correct information units, is hand-scored.

This requires establishing inter- and intra-rater reliability and is also very time-consuming. Of particular interest to the current study is the extent to which cohesion measures (propositional, macrostructural, or pragmatic), rather than lexical-syntactic linguistic measures, can be automatically extracted from transcripts of persons with and without aphasia. Unfortunately, the most widely available means of scoring discourse measures, such as cohesion (a propositional metric) and coherence (a macrostructural metric), are hand-scored and time-consuming (Wright et al., 2010; Glosser and Deser, 1991). Further, the test-retest reliability of these discourse measures has rarely, if ever, been evaluated in aphasia. As such, the preliminary results presented in this paper are in response to the need for automatic scoring methods to extract and evaluate meaningful information from discourse.

This paper builds on the Stark et al. (2023) study to evaluate automatically extracted propositional and macrostructural components at test and retest for one discourse task in persons with and without aphasia using the Tool for the Automatic Analysis of Cohesion (TAACO) (Crossley et al., 2019). For this paper, coherence, as defined by Halliday and Hasan (1976), is based on the cohesion in a text, which in turn is a semantic relation. Semantic coherence, by this definition, captures the general content of the text and can be interpreted as a macrostructural measure.

Crossley et al. (2019) also describe this as their basis for TAACO 2.0, through which they target explicit as well as implicit levels of semantic coherence in English writing tests. This tool could greatly reduce manual efforts in scoring, and highlight discourse-level patterns (and possible impairments) without requiring time-consuming human-scoring. We address the lack of research on cohesion and coherence in aphasia by validating the use of this tool, to differentiate aphasia and control transcripts based on semantic cohesion. Additionally, we explore the relationship between local and global coherence variables for semantic cohesion.

2. Automatic Scoring of Discourse-level Metrics

Earlier literature in text cohesion analysis includes the use of WordNet Teich and Fankhauser (2004), to automatically annotate texts that had potential cohesive ties. Since then, improvements have been made in the annotation methods and scoring methods. Martinez and Lapshinova-Koltunski, 2016 compared manual and automatic procedures to annotate lexical cohesion in GECCo Kunz et al. (2014), a corpus of English and German data, including textual and spoken data. Their findings suggest that there is a need for better automatic methods for annotating lexical cohesion. The manual correction of automatic system output was found to be more time-consuming than starting from scratch, indicating that the automatic system's output required substantial post-editing. This highlights the difficulty of the annotation task and the challenges in achieving high agreement scores, even for human annotators. The complexity of the annotation process, along with the linguistic analysis involved, underscores the necessity for improved automatic methods that can accurately capture and represent lexical cohesion in text.

More recent work on text coherence analysis has relied on extracting semantic information and relations from a given input text using supervised methods. Notably, Cui et al. (2017) proposed a deep coherence model (DCM) using a convolutional neural network model that combined a sentence distribution representation with text coherence modeling. The model was trained on report-based corpora from aviation accidents and earthquakes, and evaluated on a sentence-ordering task. These results were promising, with the DCM showing a 5.3% average improvement gain over existing methods. Despite having a good performance in deriving abstract semantic representations, the model does not accurately categorize semantic features.

In this paper, we suggest a novel approach to extending the use of TAACO (Crossley et al., 2019) to evaluate semantic coherence in other forms of

textual data, namely aphasic and non-aphasic transcripts. For our analyses, we have interpreted cohesion as local ties, usually within a sentence, and coherence as more global ties, across sentences. Hence, cohesion in this paper, is a highly semantic, propositional measure while coherence is a macrostructural measure.

A crucial component of Crossley et al. (2019) results was that the global semantic similarity reported by word2vec Church (2017) was an important predictor of coherence, which is consistent with existing theories of coherence in this school of thought (Kintsch, 1992; Gernsbacher and Talmy, 1995). TAACO 2.0 also has an additional feature for calculating lexical and semantic overlap between a source and response text. This aligns perfectly with our intended goal for comparing test-retest transcripts across time points, to evaluate how much cohesion and coherence is retained.

3. Research Hypotheses

This is a preliminary study using automatic semantic scoring methods in aphasic and non-aphasic transcripts across tests and retest. The goals for this paper are twofold:

1. **Test-Retest Reliability:** The hypothesis is that, if the variables are reliable in a short test-retest format, the variables should not be significantly different across the time points, for either group. This would suggest that the measures evaluated in this study are stable across time, for each group.

2. **Inter-Group Discourse Cohesion:** The hypothesis is that aphasic discourse would be less cohesive than control discourse, which is a validity check given much research establishing that persons with aphasia produce less coherent and cohesive speech (Galetto et al., 2013; Hazamy and Obermeyer, 2020; Leaman and Edmonds, 2021). Hence, the group with aphasia should exhibit overall lower scores than the control group across variables.

We also hoped to highlight any significant correlations between word- and sentence-level variables, as described below in Section 5.1.

4. Data

4.1. Transcripts

All text transcripts were obtained from the NEURAL Research Lab Corpus (talkbank.org). The corpus comprises 24 pairs of test-retest transcripts from persons without aphasia and 23 pairs of test-retest transcripts for persons with aphasia. We have skipped 1 aphasic transcript where the task being analyzed was missing at the test or retest time point, leaving 22 total pairs of test-retest transcripts for persons with aphasia.

4.2. Participants

All participants were part of a larger study (Stark et al., 2023) that aimed to compare test-retest reliability for discourse measures between two groups: individuals with aphasia and individuals without aphasia. The test and retest was spaced approximately 7.79 ± 1.72 days apart. The sample size estimation was determined based on a pilot sample of n = 7 individuals with aphasia and n = 9 speakers without brain damage. The final sample included n = 25 persons with aphasia and n = 24 age- and education-matched adults without brain injury.

Subject recruitment was conducted virtually, and participants were screened using an online survey. The inclusion criteria for the non-brain-damaged group were being native English speakers, aged between 45 and 80, with at least 10 years of education and no history of brain injury or neurological or developmental language disorder. The inclusion criteria for individuals with aphasia were being native English speakers, aged 18 or older, with a diagnosis of aphasia due to an acquired brain injury at least 6 months prior to the study and without any other neurological disorder or neurodegenerative disease.

All samples were collected under Indiana University IRB #1904590484. All data used in this study is available for free to members of AphasiaBank (MacWhinney, 2000). Informed consent was obtained, and neuropsychological tests were administered to verify eligibility, including the Montreal Cognitive Assessment (et al., 2005) for the non-brain-damaged group and the Bedside version of the Western Aphasia Battery-Revised (Kertesz, 2007) for the aphasia group. For detailed demographic and neuropsychological information about included participants, please refer to (Stark et al., 2023).

5. Methods

We have used TAACO 2.0.4 (Crossley et al., 2016) for the semantic analysis of text transcripts in our study. TAACO uses 194 indices in seven main categories: Type-Token Ratio (TTR) and TTR Density, Lexical Overlap (sentences), Lexical Overlap (paragraphs), Semantic Overlap, Connectives, Givenness, and Source Text Similarity. We refer to these as our linguistic variables of interest. These have been further described below and summarized in Table 1.

For the pilot run of the automatic semantic scoring, we defaulted to processing all available analyses in TAACO, to extract semantic information, except the paragraph analysis feature, as each text transcript comprises one paragraph chunk after pre-processing. Additionally, we have used the bag-of-words approaches available by default in TAACO as measures of semantic similarity across test and retest time points for each pair of transcripts. Below we describe indices we have used for within and between text comparisons.

5.1. Semantic Indices

Word-level metrics: Type-Token-Ratio (TTR) is a measure of lexical diversity. TAACO 2.0.4 calculates TTR for various part of speech categories, one-word, 2-word (bigram) and 3-word (trigram) phrases, including a moving average TTR (MATTR) that reports the average TTR score for an overlapping sequence of 50 words. At the word level, we chose the lemma MATTR and the function MATTR as measures of semantic cohesion. MATTR is a known measure of lexical diversity, and is known to be much better than TTR in handling populations where the speech sample sizes are vastly different and in very short samples like what we see in aphasia (Cunningham and Haley, 2020).

Discourse-level metrics: Since we do not have true paragraphs in our samples, we chose the adjacent sentences' overlapping indices to look at local coherence at the sentence level. These metrics essentially measure the semantic overlap of variables from one sentence to the next. We chose binary overlap for all words, content-word overlap, function words, and arguments.

Givenness: Next, we look at givenness in text cohesion, which is an approximation of the ratio of given information to new information, examined through pronoun density, pronoun-to-noun ratios, and repeated content lemmas and pronouns.

Semantic overlap: Finally, we evaluate semantic overlap across test and retest time points using the Latent Semantic Analysis (LSA) and word2vec options available in TAACO. These are indices of both local and global cohesion that use the WordNet database to measure overlap between words and between sentences. This is a broader way of measuring topic maintenance and more general coherence across the sentences by looking at the semantic similarity of words across sentences.

5.2. Task

We start with analyzing the "Broken Window" task in this paper. This is a picture sequence description task comprising a set of four pictures, where there is a logical progression of events (Fig. 1). Participants describe what they can see in the picture

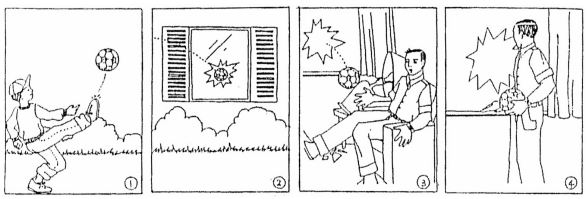


Figure 1: Picture Sequence task: "Broken Window". A four-panel visual stimulus is given to participants, who describe the logical progression of the events.

"Now I'm going to show you these pictures. Take a little time to look at these pictures. They tell a story. Take a look at all of them, and then I'll ask you to tell me the story with a beginning, a middle, and an end. You can look at the pictures as you tell the story."

If no response is received after 10 seconds, they are prompted as follows:

"Take a look at the first picture and tell me what you think is happening."

If necessary, they are continued to be prompted for each of the panels 2, 3, and 4, as follows:

"And what happens in the second panel?
Again, if no response is received in 10 seconds, they are prompted, as follows:

"Can you tell me anything about this picture?"

And again, if no response is received in 10 seconds, they prompted, as follows:

"Is the boy kicking the ball through the window?"

Figure 2: The figure shows what a typical exchange between the invigilator and the participant might look like. Annotations would be added to the recorded speeches.

sequence. They tend to go in order (from left to right) in describing the pictures. The task instructions are always given in the same way to each participant (Fig. 2). The participants' speeches were recorded and transcribed using specific annotation guidelines employed by AphasiaBank (MacWhinney, 2000).

This task was ideal for investigating automatic semantic cohesion from the aphasia dataset given that individuals tend to go in order and produce logical sequences of language because of the available visual information from the pictures. Low scores on the metrics evaluated in this study could reflect an impairment or inability to connect the pictures

Semantic Category	Semantic (TAACO) Index	Semantic Information
Word Level	lemma_maTTR	Lexical diversity with a moving average window
	function_mattr	Lexical diversity with a moving average window
Discourse Level (Local Indices)	all_sentence_overlap	Cohesion for all words
	content_word_overlap	Cohesion for all content words
	function_word_overlap	Cohesion for all words
	argument_sent_overlap	Cohesion for all arguments
Semantic Overlap (Global Indices)	LSA_all_pairs	Similarity across adjacent sentence pairs
	LSA_combined_pairs	Similarity between combined sentence pairs
	word2vec_all	Similarity across adjacent sentence pairs
	word2vec_combined_pairs	Similarity between combined sentence pairs
Givenness	repeated_content	Cohesion index for all repeated content
	repeated_pronouns	Cohesion index for repeated content, pronouns

Table 1: The table summarizes the semantic levels chosen for our analysis using TAACO 2.0.4 indices.

through language, a lack of vocabulary sufficient to connect pictures, or a lack of logical progression.

5.3. Pre-processing

TAACO works on plain text and does not account for the transcription annotations in the data. Hence, extensive pre-processing was needed for these text files. All text transcripts were pre-processed using an automated script in Python. Of these, two control transcripts and three aphasic transcripts were manually cleaned, owing to discrepancies in transcription annotations. We have summarized the data cleaning decisions in Table 2.

5.4. Tools

We have used TAACO 2.0.4 (Crossley et al., 2019), which runs on Python2 for extracting semantic info and comparing semantic information across the test and retest time points. Pre-processing and automation for the text transcripts were done in Python3. Statistical analysis of the data was conducted using the JASP software (JASPTeam, 2024).

6. Results

The data was not normally distributed, and therefore non-parametric statistics were computed to evaluate the two hypotheses.

To evaluate test-retest reliability, a paired Wilcoxon signed-rank test was conducted on the variables for each group (aphasia, control). Significant ($p < 0.05$) findings should be interpreted as a difference between test and retest for that specific measure, suggesting unreliable metrics.

A Mann-Whitney U test was conducted on the variables to compare across the 2 groups (aphasia vs. control), corrected for multiple comparisons ($\alpha = 0.05$, JASP default = Bonferroni correction). Significant ($p < 0.05$) findings should be interpreted as a difference between the groups, where it is anticipated that the control group produces more cohesive language.

6.1. Test-Retest Reliability

Few significant differences were found for variables between test and retest for either group, the exception being a significant difference identified for repeated pronouns in the control group (Table 5).

6.2. Inter-Group Discourse Cohesion

There was no significant difference between the two groups, with the exception of word-level semantics and two sentence-level measures (Table 3). We can also see from Table 4 that indices for the control group were higher than the aphasia group. It is evident from this and Fig. 4, that the aphasia group had noticeable variability across test and retest time points, whereas the control group was more concentrated across the semantic indices at each time point. This also follows from similar findings in (Stark et al., 2023). Individual plots for semantic categories across indices can be found in Appendix A.

6.3. Word- and Other-Level Semantic Correlations

Word-level indices were negatively related to discourse-level metrics in both subject groups, such that greater lexical diversity for lemmas as well as function words tend to be strongly negatively related to givenness variables (repeated content words and repeated pronouns). Generally, the discourse-level metrics were positively related to the semantic overlap variables derived from LSA and word2vec for both groups, but especially for the aphasia group.

Annotation	Process	Example	
	Input	*Output*	
CHAT notations	Removed	And then &=points he	*And then he*
Neologisms	Removed	grb@n	—
Repetition	Retained	<goes in> goes in	*goes in goes in*
Dialectal variation	Replaced	durn [:darn]	*darn*
Phonological errors	Replaced	gos [:got]	*got*
Morphological errors	Replaced	He looks [: look] [* m:03s]	*He looks*
Semantic errors	Retained	kick her [: his] [* s:r:gc:pos]	*kick her*
Other errors	Removed	he s@l	*he*

Table 2: The table shows a summary of pre-processing decisions taken to facilitate processing by TAACO, as it explicitly functions on plain text data.

Semantic Index	W	p
lemma_maTTR	499.000	$<$ .001
function_maTTR	696.000	0.005
adjacent_overlap_all_sent	1368.500	0.015
all_sentence_overlap	898.500	0.208
content_word_overlap	852.500	0.112
function_word_overlap	861.500	0.124
argument_sent_overlap	636.000	0.001
LSA_all_pairs	1056.000	1.000
LSA_combined_pairs	1038.000	0.891
word2vec_all_pairs	1007.000	0.706
word2vec_combined_pairs	980.000	0.557
repeated_content	1094.000	0.769
repeated_pronouns	1021.500	0.790

Table 3: Mann-Whitney U test for Aphasia and Control across Test and Retest timepoints, suggesting a significant difference at word-level semantics and 2 measures at the discourse level

Positive relationships, some of which were significant, existed between all discourse-level, givenness, and semantic variables. Therefore, lexical diversity at the word level cannot be assumed to be a good metric for positively predicting discourse-level cohesion for persons with or without brain injury.

The heatmap shown in Fig. 3 suggests that word-level metrics are generally negatively correlated across the two groups. It is also evident that givenness is positively correlated in either group. Interestingly, semantic similarity overlap is more correlated in aphasia than in control.

7. Discussion and Conclusions

Cohesion and coherence in aphasia have only ever been investigated using hand-scoring methods, which are inconvenient and time-consuming. As such, little is known about the relationship between word-level variables (which are much more commonly evaluated) and discourse-level variables related to semantic cohesion. Hence, this paper is an effort to validate that an automatic metric, TAACO, can differentiate aphasia and control transcripts at the level of semantic cohesion. The hypothesis was that the discourse extracted from persons with aphasia would be less coherent than the discourse extracted from persons without aphasia (a control sample). We also evaluated these semantic metrics across test and retest, which is a novel investigation, especially in aphasia.

Cohesion and coherence were generally preserved across test and retest points in both groups, except for word-level semantics and two sentence-level TAACO measures. This is contrary to our second hypothesis. Individuals with aphasia may experience difficulties in lexical retrieval, accessing or processing word meanings, impaired semantic network, or employ compensatory strategies and reliance on alternative word choices content or function words. Generally, both groups had relatively stable semantic cohesion metrics, which follows from prior work in the field (Shekim and LaPointe, 1984; Ulatowska et al., 1983). As the body of Stark et al. (2023)'s work has shown, this should be interpreted cautiously as being relevant to only the task at hand (a picture sequence description) and must be thoroughly investigated in new samples and new tasks.

There has been ongoing debate regarding the relationship between coherence and cohesion in language. In this paper, coherence is defined based on cohesion, which is a semantic relation within a text. We have specifically evaluated these for aphasic and non-aphasic transcripts. Global semantic similarity, measured using word2vec, was a significant predictor of discourse-level and givenness metrics, aligning with existing theories in the field. However, word-level metrics of lexical diversity were negatively (often significantly) not related to discourse-level, givenness, or semantic overlap in either group. This finding suggests caution in extrapolating word- to discourse-level metrics of

Semantic Index	Aphasia			Control		
	$\bar{x}$	σ	v	$\bar{x}$	σ	v
lemma_maTTR	0.598	0.112	0.187	0.667	0.049	0.074
function_maTTR	0.443	0.137	0.309	0.489	0.083	0.171
all_sentence_overlap	0.771	0.296	0.384	0.875	0.138	0.157
content_word_overlap	0.455	0.299	0.658	0.589	0.223	0.378
function_word_overlap	0.736	0.294	0.399	0.851	0.149	0.175
argument_sent_overlap	0.498	0.301	0.604	0.709	0.234	0.330
LSA_all_pairs	0.327	0.164	0.502	0.334	0.107	0.319
LSA_combined_pairs	0.595	0.191	0.321	0.637	0.082	0.129
repeated_content	0.213	0.093	0.435	0.211	0.053	0.253
repeated_pronouns	0.278	0.129	0.465	0.290	0.074	0.257
word2vec_all_pairs	0.778	0.081	0.105	0.795	0.049	0.062
word2vec_combined_pairs	0.783	0.162	0.207	0.805	0.103	0.128

Table 4: Statistical Descriptives for semantic indices ($\bar{x}$=Mean, σ=Standard Deviation, v=Coefficient of Variation). Mean coherence was found to be higher for control than those for aphasia, as shown in Table 4

Semantic Index	Aphasia			Control		
	W	z	p	W	z	p
lemma_maTTR	93.000	−1.088	0.290	195.000	1.286	0.208
function_maTTR	113.000	−0.438	0.679	208.000	1.657	0.101
all_sentence_overlap	79.500	0.142	0.906	86.500	−0.342	0.747
content_word_overlap	123.000	0.261	0.808	133.000	0.211	0.846
function_word_overlap	84.500	−0.423	0.687	112.500	0.280	0.794
argument_sent_overlap	95.500	−0.355	0.737	112.000	−0.122	0.917
LSA_all_pairs	127.000	0.016	1.000	207.000	1.629	0.107
LSA_combined_pairs	118.000	−0.276	0.799	165.000	0.429	0.684
word2vec_all_pairs	130.000	0.114	0.924	153.000	0.086	0.944
word2vec_combined_pairs	132.000	0.179	0.874	119.000	−0.886	0.390
repeated_content	138.000	0.373	0.726	125.000	−0.714	0.491
repeated_pronouns	124.000	−0.081	0.949	76.000	−2.114	0.034

Table 5: Wilcoxon signed-rank test for Aphasia and Control across Test and Retest timepoints, suggesting no significant difference in these variables at retesting

semantic cohesion.

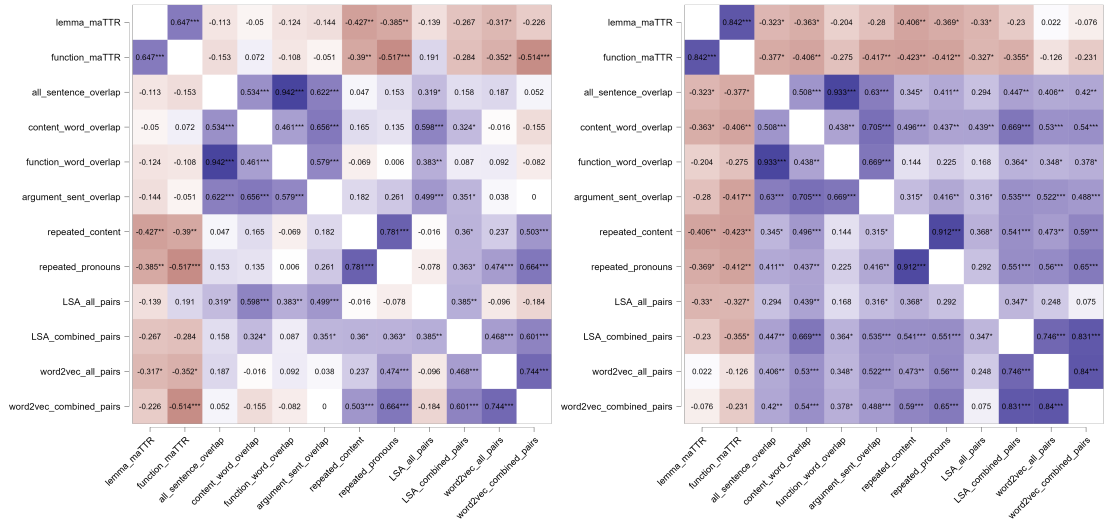
Overall, the discussed findings and methodology contribute to demonstrating the applicability of TAACO (Crossley et al., 2019) in assessing semantic coherence. Such automatic approaches can provide more objective and consistent measures of global coherence, reduce analysis time and resources, enable fine-grained analysis of coherence, ensure standardization and replicability, and facilitate broader investigations across diverse populations and contexts.

8. Future Work

We explored methods for evaluating TAACO's performance and determining if it is effective for evaluating semantic coherence, and differences between the control and aphasia data provide early validation. A clear next step is establishing how well TAACO performs with ground truth, such as a hand-scored validity check. This could be done by comparing TAACO-extracted metrics to the scales used to evaluate cohesion and coherence in aphasia, as discussed in the introduction.

It is important to note that TAACO works on plain text and does not consider transcription annotations in the data, which should be considered for replication of the study. We have discussed the importance of enhancing pre-processing and cleaning techniques to improve the overall performance of TAACO. In this regard, we aim to expand the scope of semantic coherence to cover more tasks in aphasia, to evaluate the performance across tasks as a follow-up to the (Stark et al., 2023). Additionally, as JASP does not have t-test corrections directly built in, we plan to apply this outside of the software to ensure that we can apply desired corrections for multiple comparisons.

Another aspect to consider is how TAACO performs specifically for more heterogeneous aphasia groups. Greater aphasia severity may impact co-



(a) Control

(b) Aphasia

Figure 3: Spearman's rho correlation between semantic variables in Control and Aphasia using, collapsed across test and retest. [Blue squares = positive correlation, Red squares = negative correlation]

hesion and coherence in non-linear ways. Future work should carefully evaluate the impact of aphasia severity, and specific aspects of aphasia (such as anomia or semantic errors) on cohesion and coherence, especially its relationship with automatic scoring of these metrics. To this end, we could also consider building a statistical model or classifier to distinguish between the two groups.

9. Acknowledgements

The data for this study was collected as part of the New Investigator Award from the American Speech Language and Hearing Sciences Foundation, awarded to Dr. Brielle Stark, Assistant Professor at the Department of Speech, Language and Hearing Sciences (Indiana University Bloomington).

We would like to express our sincere gratitude to Dr. Timothy J. Pleskac, Professor at the Department of Psychological and Brain Sciences (Indiana University Bloomington), for his assistance with statistical analysis using JASP. Additionally, we extend our appreciation to Dr. Sandra Kuebler, Professor of Linguistics (Indiana University Bloomington), and Dr. Francis M. Tyers, Assistant Professor of Linguistics (Indiana University Bloomington), for their feedback and insight throughout the development of this work.

Finally, we would like to acknowledge our anonymous reviewers for their valuable comments and suggestions that greatly contributed to improving this paper.

A. Plots

Fig. 4 shows one linguistic variable for the word and sentence level semantic categories across test and retest in aphasia and control groups.

B. Bibliographical References

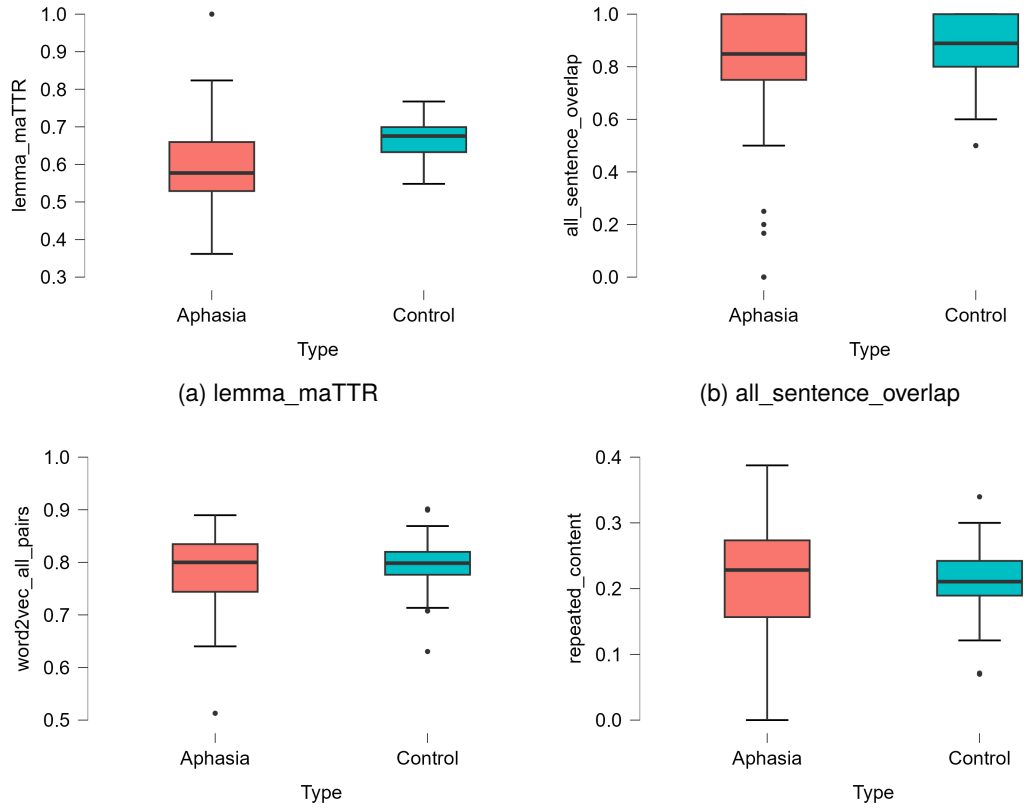
Z. S. Nasreddine , N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow. 2005. The montreal cognitive assessment, moca: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatric Society*.

Lucy Bryant, Alison Ferguson, and Elizabeth Spencer. 2016. Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics Phonetics*.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*.

S. A. Crossley, K. Kyle, and D. S. McNamara. 2016. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods 48*.

Scott A. Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*.



(a) lemma_maTTR

(b) all_sentence_overlap

(c) word2vec_all_pairs

(d) repeated_content

Figure 4: Intergroup coherence scores for Aphasia and Control collapsed across test and retest

Baiyun Cui, Yingming Li, Yaqing Zhang, and Zhongfei Zhang. 2017. Text coherence analysis based on deep neural network. *CIKM '17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*.

Kevin T Cunningham and Katarina L. Haley. 2020. Measuring lexical diversity for discourse analysis in aphasia: Moving-average type-token ratio and word information measure. *Journal of speech, language, and hearing research*.

V. Galetto, S. Kintz, T. West, Heather Harris Wright H., and G. Fergadiotis. 2013. Measuring global coherence in aphasia. *Procedia - Social and Behavioral Sciences*.

Morton Ann Gernsbacher and Givón Talmy. 1995. *Coherence in Spontaneous Text*. John Benjamins Publishing Company.

G. Glosser and T. Deser. 1991. Patterns of discourse production among neurological patients with fluent language disorders. *Brain and Language*.

G. Glosser and T. Deser. 1992. Patterns of discourse production among neurological patients with fluent language disorders. *Aphasiology*.

M.A.K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Routledge.

Audrey A. Hazamy and Jessica Obermeyer. 2020. Evaluating informative content and global coherence in fluent and non-fluent aphasia. *International journal of language communication disorders*.

JASPTeam. 2024. Jasp (version 0.18.3)[computer software].

A. Kertesz. 2007. Western aphasia battery–revised. *The Psychological Corporation*.

Walter Kintsch. 1992. *How readers construct situation models for stories: The role of syntactic cues and causal inferences*. American Psychological Association.

Kerstin Kunz, Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski, Katrin Menzel, and Erich Steiner. 2014. *English-German contrasts in cohesion and implications for translation*, chapter 9. Mouton de Gruyter.

Marion C. Leaman and Lisa A. Edmonds. 2021. Measuring global coherence in people with aphasia during unstructured conversation. *American Journal of Speech-Language Pathology*.

E. Van Leer and L. S. Turkstra. 1999. Coherence in narratives of adolescents with brain injury. *Brain Injury*.

Brian MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk: Transcription format and programs*. Lawrence Erlbaum Associates Publishers.

Jose Manuel Martinez Martinez and Ekaterina Lapshinova-Koltunski. 2016. Annotation of lexical cohesion in english and german: Automatic and manual procedures. *Proceedings of the 13th Conference on Natural Language Processing*.

L. Shekim and L. L. LaPointe. 1984. Coherence in aphasic and normal elderly narratives.

Brielle C. Stark, Julianne M. Alexander, Anne Hittson, Ashleigh Doub, Madison Igleheart, Taylor Streander, and Emily Jewell. 2023. Test–retest reliability of microlinguistic information derived from spoken discourse in persons with chronic aphasia. *Journal of Speech, Language, and Hearing Research*.

Elke Teich and Peter Fankhauser. 2004. Wordnet for lexical cohesion analysis. *The Second Global Wordnet Conference*.

H. K. Ulatowska, M. J. Allinder, and E. H. Lichtenstein. 1983. Microlinguistic and macrolinguistic aspects of aphasia: Their relationship to linguistic theory and to linguistic intervention. *Journal of Speech and Hearing Research*.

H. H. Wright, A. Koutsoftas, G. Fergadiotis, and G. Capilouto. 2010. Coherence in stories told by adults with aphasia. *Procedia-Social and Behavioral Sciences*.

Heather Harris Wright, Gilson J. Capilouto, and Anthony Koutsoftas. 2013. Evaluating measures of global coherence ability in stories in adults. *International Journal of Language Communication Disorders*.