Brain & Language 283 (2026) 105850

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Normative clinical language data and task specific effects☆

Brielle C. Stark [a,*], Charalambos Themistocleous [b]

[a] Indiana University Bloomington, Department of Speech, Language and Hearing Sciences, USA
[b] University of Oslo, Department of Special Needs Education, Norway

ARTICLE INFO

Dataset link: TalkBank Clinical Banks (Reference data)

Keywords:
Clinical linguistics
Aphasia
Right hemisphere disorder
Mild cognitive impairment
Dementia
Discourse
Task
Traumatic brain injury

ABSTRACT

Language production in clinical populations varies by both neurological condition and discourse task. This study analyzes connected speech from 1394 individuals with left or right hemisphere damage, Traumatic Brain Injury, Mild Cognitive Impairment, dementia, or no neurological diagnosis. Participants completed narrative, procedural, and picture-based tasks. Using Open Brain AI to extract linguistic features across phonology, morphology, syntax, semantics, lexicon, and readability, we applied mixed-effects models to examine effects of diagnosis, task, and their interaction. Results showed that task type strongly shapes linguistic output and interacts with diagnostic group across nearly all domains. Narrative tasks elicited more complex syntax and aspectual morphology, while procedural tasks prompted simpler grammatical structures and picture-based descriptions produced greater lexical diversity. Uniform Manifold Approximation and Projection further revealed clear clustering by task type, reflecting a continuum of contextual constraint ranging from structured picture descriptions and procedural instructions to free speech and open-ended narratives. This gradient illustrates a shift in linguistic scaffolding: highly structured tasks tightly constrain the target lexicon and syntax, whereas open-ended tasks require speakers to independently generate and organize narrative content. These findings show that clinical language tasks are not interchangeable. Treating them as equivalent may reduce diagnostic accuracy and weaken comparisons with healthy norms. We propose an Open Brain AI task-specific framework for clinical language assessment to improve identification of linguistic impairments and guide targeted intervention.

1. Introduction

Language production in neurogenic communication disorders is shaped not only by lesion location or disease process but also by the discourse task used to elicit speech. Research in left hemisphere damage, aphasia, Traumatic Brain Injury (TBI), dementia, and Mild Cognitive Impairment (MCI) has consistently demonstrated that task structure drives systematic variation in connected speech, often interacting with diagnostic profile to reveal or obscure impairments (Fridriksson et al., 2018; Stark & Fukuyama, 2021). For example, narrative tasks (e.g., story retell) typically elicit greater syntactic complexity, denser propositional content (i.e., packing more distinct ideas and details into a single utterance), and increased use of story grammar elements (such as establishing a setting, an initiating event, and a resolution) (Richardson et al., 2021; Stark & Fukuyama, 2021). Procedural tasks (e.g., explaining how to make a sandwich) yield simplified syntax, frequent imperative constructions, and a higher noun-to-verb ratio, alongside heavy use of prepositions and "light" verbs (Stark & Fukuyama, 2021; Ulatowska et al., 1981). Picture descriptions emphasize lexical retrieval of core scene items and promote lexical variety, but they often elicit fewer words overall, rely on simpler clause structures, and show a high proportion of nouns relative to verbs, with limited cohesion (Fergadiotis & Wright, 2011; Stark, 2019). Picture sequences add demands for temporal sequencing and causal reasoning, eliciting more global coherence markers and causal connectives than single pictures, but with relatively simple syntax (Stark & Fukuyama, 2021).

Critically, the expression of impairment depends on task-by-diagnosis interactions rather than on broad genre alone. Although robust discourse-type effects have been reported in both healthy adults and individuals with aphasia, the magnitude and profile of impairment vary across specific elicitation contexts, including tasks that nominally belong to the same genre. A patient may show pronounced syntactic disruption in a fictional or autobiographical narrative, where discourse generation, referential maintenance, and clause-level integration are

☆ This article is part of a Special issue entitled: 'language processing in the wild' published in Brain and Language.

* Corresponding author at: Indiana University Bloomington, Department of Speech, Language and Hearing Sciences, BC Stark. 2631 E Discovery Parkway, Bloomington, IN 47408, USA
E-mail address: bcstark@iu.edu (B.C. Stark).

https://doi.org/10.1016/j.bandl.2026.105850
Received 5 September 2025; Received in revised form 2 September 2026; Accepted 1 October 2026
Available online 7 October 2026
0093-934X/© 2026 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

heavily taxed, yet appear relatively less impaired in a procedural task, where familiar action schemas and predictable sequential structure reduce demands on discourse framing and long-range planning (Stark, 2019).

At the same time, even superficially similar tasks are not linguistically equivalent: prior work has shown that narrative retells tend to be richer in content but slower in rate, procedural discourse often supports simpler grammatical organization, and even closely matched picture description tasks yield distinct linguistic profiles (Stark & Fukuyama, 2021). Moreover, linguistic markers themselves differ by task: people with aphasia show reduced lexical diversity, informativeness, and efficiency, but the extent of these deficits varies depending on elicitation context (Fergadiotis & Wright, 2011; Stark, 2019; Stark, Alexander, et al., 2023; Stark & Fukuyama, 2021). The findings in clinical research are corroborated further by research within corpus linguistics (especially after the publication of the Brown Corpus in the 1960s by Kučera, Henry and W. Nelson Francis) showing that linguistic genres impose certain linguistic patterns pertaining most linguistic domains, e.g., phonology, morphology, lexicon and syntax (Francis & Kucera, 1982; Gries, 2016). In sum, we hypothesize that this marked variability in an individual's linguistic output across descriptive, procedural, and narrative contexts underscores a danger in task-agnostic assessments. Treating disparate elicitation methods as equivalent complicates accurate diagnosis and undermines the reliability of longitudinal tracking. Specifically, single-task evaluations risk inflating the severity of an impairment when a task inherently suppresses certain linguistic features, or conversely, obscuring subtle deficits when a task affords substantial contextual scaffolding.

More recent evidence shows that task effects extend beyond overt aphasia. Speakers with latent aphasia and MCI were differentiated from cognitively healthy adults on core lexicon measures (Kim et al., 2024), but the sensitivity of those measures depended on task type (Stark et al., 2025). For example, Cinderella retellings revealed between-group differences most reliably, whereas picture descriptions sometimes masked subtle impairments. This pattern aligns with theoretical accounts of discourse processing: highly structured tasks like the Cookie Theft scene provide strong contextual scaffolding; by contrast open-ended narratives demand greater topic generation, sequencing, and memory retrieval, thereby amplifying cognitive-linguistic vulnerabilities (Stark & Fukuyama, 2021). Similarly, patients with TBI, whilst demonstrating some impaired microlinguistic features, typically associate with reduced macrostructural organization and pragmatic appropriateness (Coelho, 2007), whereas those with right hemisphere damage demonstrate preserved sentence-level grammar but impaired inferencing, cohesion, and global coherence (Minga et al., 2022).

Beyond acute neurological events such as stroke and TBI, the analysis of connected speech has proven particularly valuable in characterizing progressive neurogenic communication disorders as seen in various forms of dementia and Primary Progressive Aphasia (PPA). For instance, in Alzheimer's disease and PPA, discourse analysis has successfully identified early, subtle shifts toward ‘empty speech’, characterized by an overreliance on pronouns (rather than specified subjects), reduced semantic density, and pragmatic tangentiality, often preceding overt performance drops on standardized clinical tests (Fraser et al., 2014; Themistocleous et al., 2021; Themistocleous, Webster, et al., 2020). Picture description tasks have been widely used in studies of MCI and Alzheimer's disease dementia (Kim et al., 2024; Themistocleous et al., 2018; Themistocleous, Eckerström and Kokkinakis, 2020; Tippett et al., 2025), where they offer a standardized method for eliciting connected speech and identifying changes in semantic content, syntactic complexity, and fluency across the course of progressive cognitive decline (Mueller et al., 2018). For PPA, naturalistic discourse production has been instrumental in distinguishing between the semantic, nonfluent/agrammatic, and logopenic variants. For example, early studies utilized narrative and descriptive tasks to quantify variant-specific breakdowns in real-time, such as the profound syntactic simplification and motor speech errors in the nonfluent variant versus the semantic naming failures and intact grammar of the semantic variant (Wilson et al., 2010). The research in populations with progressive neurological conditions underscores the utility of discourse analysis as a sensitive, ecologically valid measure across the entire spectrum of neurogenic communication disorders.

The clinical implications are significant. Despite multiple elicitation contexts improving the validity of a comprehensive linguistic assessment, most protocols rely on a single task (Bryant et al., 2016; Cruice et al., 2020; Stark & Fukuyama, 2021), often the Cookie Theft picture description, because transcription and coding remain time-consuming. Reliability studies likewise show that test-retest stability of discourse measures is task- and sample-specific, rather than a fixed property of a measure (Stark, Alexander, et al., 2023). Thus, relying on one probe both obscures diagnosis and undermines sensitivity to treatment-related change. Taken together, existing work supports three conclusions. First, discourse tasks are not interchangeable; genre and stimulus constrain lexical, syntactic, and discourse patterns in predictable ways. Second, diagnosis interacts with task such that impairment profiles emerge differently depending on discourse context. Third, clinical validity and reliability depend on aligning tasks with the linguistic construct of interest, rather than assuming generalization across elicitation methods. In practice, this means that clinicians should select tasks that directly place pressure on the feature they wish to evaluate. For example, if the goal is to assess a patient's ability to produce complex, hierarchical syntax, a narrative task will generally be more informative than a single-picture description because narratives require the speaker to generate and organize events across clauses, maintain reference over time, and build more extended syntactic structures. A single-picture description, by contrast, often permits successful performance with shorter, simpler utterances anchored to immediately available visual content. As a result, relatively preserved syntax in a picture-description task should not be assumed to indicate comparable ability when the patient is asked to produce a narrative. Failing to make this distinction risks underestimating deficits in higher-level syntactic organization and, in turn, selecting treatment targets that do not adequately address the patient's actual areas of difficulty.

2. Linguistic measures

Computational language analysis offers comprehensive, quantitative, and objective measures of discourse microstructure and macrostructure and accounts for the breakdown of language faculty in patients with aphasia (Callegari et al., 2024; Charniak & Johnson, 2005; Devlin et al., 2018; Fraser et al., 2015; Goldstein et al., 2022). For example, computational analysis quantifies word usage like lexical richness (Kim et al., 2023) and word-finding difficulties (Efstratiadou et al., 2018). Simple automated measures like the total word count may additionally reflect overall severity (Stark, Alexander, et al., 2023). Also, it can reveal phonological deficits by quantifying measures like the number of syllables and syllable complexity (Ball et al., 2014; Leyton et al., 2014). Comparing these measures across different clinical groups reveals how neurological impairments disrupt the cognitive representation and motor execution of speech (Barbieri et al., 2018; Croot et al., 2000). Automated morphological measures evaluate word structure, the distribution of parts of speech (POS), and inflectional categories such as tense, number, gender, and case, which can serve as indicators of pathologies like agrammatism and anomia (Badecker et al., 1990; Caramazza & Hillis, 1991; Fridriksson et al., 2018; Hillis, 1989; Hillis et al., 2018; Stockbridge et al., 2021). By quantifying sentence structure (e.g., subject-verb-object order), grammatical rules (e.g., subject-verb agreement), and phrase hierarchy, computational language analysis can account for syntactic deficits, which are critical in aphasia (Bastiaanse, 2013; Caramazza & Hillis, 1989; Mack et al., 2021; Thompson & Mack, 2014; Wilson et al., 2016). Automated semantic measures can map how meaning is constructed by quantifying

predefined semantic categories—such as the names of persons, organizations, locations, times, numeric quantities, and monetary values. Identifying semantic breakdown is critical for assessment, diagnosis, and setting therapeutic targets (Antonucci, 2009; Boyle, 2010; Hashimoto & Frome, 2011; Kiran & Thompson, 2003). Readability metrics quantify text accessibility and comprehension difficulty based on features such as average word length, sentence length, and syllable counts (Dale & Chall, 1948; Fitzsimmons et al., 2010; Klare, 1974; Themistocleous, 2024). To facilitate the computational analysis of language, we have developed Open Brain AI (Themistocleous, 2024). As normative output varies significantly by elicitation method, what constitutes a "low" or "normal" range must be evaluated against task-specific baselines; to facilitate this, automated frameworks such as Automated Language Biomarker Analysis (ALBA; (Themistocleous & Stark, 2026) have been developed to benchmark individual patient metrics against standardized normative corpora.

3. This study

In clinical practice, it is well established that the specific discourse tasks used to assess language significantly influence patient output. However, it is often unclear precisely how an individual's performance on a picture description differs from their ability to tell a story or explain a procedure. This ambiguity makes it challenging to select the optimal tasks for revealing specific deficits or setting targeted treatment goals. Furthermore, the manual transcription and linguistic coding of these language samples is an incredibly laborious and time-consuming process, widely cited as the primary barrier to implementing comprehensive discourse analysis in both routine clinical practice and large-scale research (Bryant et al., 2017; Pritchard et al., 2017; Stark et al., 2020). Although there is an extensive and rich body of literature investigating connected language production across multiple clinical populations, comprehensive, multi-level analyses remain rare because of these manual constraints. Consequently, previous research has frequently focused on isolated subsets of linguistic features, such as syntactic complexity (e.g., using Mean Length of Utterance; Stark, 2019) or lexical diversity, yet it was potentially overlooking parallel effects in domains like phonology or readability.

This study aims to bridge this gap by utilizing Open Brain AI (Themistocleous, 2024), an automated computational platform, to analyze an exceptionally large dataset of 9900 clinical language samples. This extensive compilation is drawn from established TalkBank resources, including AphasiaBank, DementiaBank, TBIBank, and RHDBank (MacWhinney, 2025). By leveraging this corpus, we extract a comprehensive suite of 227 linguistic features spanning phonology, morphology, syntax, semantics, lexicon, and readability. Some of these metrics—such as cross-task phonological profiling, clinical readability indices, and highly granular morphological and syntactic markers—are applied here at an unprecedented scale.

The dataset encompasses fifteen distinct tasks, categorized into four broad elicitation types: 1) Picture description (e.g., the "Cookie Theft" scene, "Broken Window" sequence); 2) Narrative retell (e.g., the "Cinderella" story); 3) Procedural explanation (e.g., making a peanut butter and jelly sandwich); and 4) Personal event narratives (e.g., recounting an illness or recovery). This varied collection captures the inherent heterogeneity within clinical populations, allowing us to delineate the linguistic "fingerprints" of different task types across diverse neurological etiologies (e.g., Left Hemisphere Damage [LHD], Right Hemisphere Damage [RHD], Traumatic Brain Injury [TBI], Mild Cognitive Impairment [MCI], and dementia). For clarity, throughout this paper, the term diagnosis is used to refer to these broad etiological categories rather than specific aphasia classifications (e.g., fluent vs. non-fluent).

By systematically testing how task-by-diagnosis interactions modulate linguistic output, our research is guided by three core hypotheses:

First, we hypothesized that task type will exert a pervasive and predictable influence on linguistic output, independent of neurological etiology. Specifically, narrative tasks will produce speech that is highly cohesive and rich in story grammar but may feature less diverse vocabulary. Procedural tasks will yield the simplest grammatical structures, characterized by shorter sentences and a direct, noun-heavy style. Conversely, picture description tasks will prompt the greatest lexical diversity as speakers attempt to label specific visual elements.

Second, we predicted that although these task effects are universal, they will be significantly modulated by a patient's specific neurological etiology. For instance, the high demand for complex sentence structure in a narrative task will likely expose severe syntactic breakdowns in a patient with LHD (e.g., agrammatism), whereas the same task might instead reveal prominent breakdowns in semantic coherence and topic maintenance in a patient with dementia.

Third, we hypothesized that these distinct linguistic profiles would allow us to map a "clinical roadmap" for task selection, demonstrating, for example, that picture descriptions are superior for rapidly assessing word-finding deficits. In stark contrast personal narratives are more sensitive to higher-level pragmatic and discourse organization skills. Ultimately, we aim to provide an evidence-based framework that moves the field beyond task-agnostic assessments toward a task-specific, etiology-informed model of clinical discourse evaluation.

4. Methodology

4.1. Participants and data

The individuals for this study were drawn from the TalkBank database (MacWhinney, 2025), a comprehensive resource featuring language samples from people with various neurological conditions alongside a critical baseline group of healthy controls. The databases utilized within the TalkBank consortium; namely AphasiaBank (MacWhinney et al., 2011), RHD Bank (Minga et al., 2022), TBI Bank (Elbourn et al., 2019; Steel et al., 2017), and DementiaBank (Lanzi, et al., 2023), adhere to shared, standardized protocols for assessment. All these databases share an established discourse protocol that elicits a variety of genres, including picture descriptions, story narratives, procedural explanations, and personal narratives. Nevertheless, these source databases are characterized by a significant clinical heterogeneity. For example, the LHD cohort encompasses various clinical subtypes (e.g., anomic, Wernicke's, and Broca's aphasia); the dementia subgroup, however, from the Pitt corpus consists of patients with dementia alongside individuals with MCI. However, we decided to incorporate these databases in their entirety to preserve the ecological validity of the data and reflect the natural heterogeneity inherent in clinical populations. Overall, the present analysis is based on a comprehensive corpus of 9900 language samples. Because many participants completed several different tasks, they often contributed multiple samples to this dataset. The detailed demographic characteristics and sample sizes of the diagnostic groups (LHD, RHD, TBI, Dementia, MCI, and Healthy Controls) are provided in Table 1 (with further details in Supplementary Data).

4.2. Measures

The speech samples were subjected to automated linguistic analysis using Open Brain AI (http://openbrainai.com; Fig. 1), a custom clinical linguistics platform developed to address the limitations of standard computational tools (Themistocleous, 2024). Unlike generic models, Open Brain AI is specifically designed for the granular phenotyping of language features, enabling hypothesis-driven research into speech pathology. The platform executes a cascade of Natural Language Processing (NLP) techniques, including tokenization, part-of-speech tagging, and dependency parsing. For each extracted feature, both raw counts and normalized ratios; which are stated after the category name were computed to account for variations in text length, and the data was

Table 1
Demographic Characteristics of Participant Groups by Diagnosis and Data Source.

Diagnostic Group	Data Source	N (Speakers)	Age in Years (M, SD)	Education in Years (M, SD)
Healthy Control (HC)	AphasiaBank	359	56.89 (15.91)	15.91 (2.64)
	RHD Bank	40	47.95 (13.54)	17.09 (2.93)
	DementiaBank (Pitt)	99	63.70 (7.90)	13.90 (2.50)
Left Hemisphere Damage (LHD)	AphasiaBank	536	61.04 (12.40)	15.70 (2.91)
Dementia	DementiaBank (Pitt)	193	71.00 (8.60)	12.20 (2.90)
Mild Cognitive Impairment (MCI)	DementiaBank (Delaware)	71	73.50 (8.03)	*
Right Hemisphere Damage (RHD)	RHD Bank	38	57.40 (12.33)	17.10 (3.99)
Traumatic Brain Injury (TBI)	TBI Bank	58	36.25 (13.47)	13.91 (3.05)

Note. Values for Age and Education are presented as mean (M) and standard deviation (SD). * Education for the MCI group was reported as percentages of highest attainment: PhD (10.81%), Bachelor's/Master's (67.57%), and Vocational Training (21.62%).

exported for statistical analysis (e.g., "Person Third Ratio" is the estimated ratio of this inflectional category over all pronoun counts whereas "Person Third Count" is the raw count of this measure; for operational definitions, see the Open Brain AI documentation: https://openbrainai.com/docs).

A comprehensive suite of 227 linguistic measures was extracted across six core domains: lexicon, phonology, morphology, syntax, semantics, and readability. The online application ALBA has been developed to provide access to these measures (Themistocleous & Stark, 2026). A complete listing of the specific metrics extracted for each of these domains and their clinical definitions is provided in Appendix 1.

Finally, the analysis incorporated measures to evaluate pragmatic dimensions, such as discourse cohesiveness and informational efficiency. Because different discourse tasks inherently demand different semantic patterns (e.g., narrative tasks require specific "person" and "location" entities, whereas procedural tasks rely heavily on inanimate objects), we utilized the extracted semantic Named Entity Recognition (NER) features to analyze the ratio of relevant task-specific entities to total words. This allowed us to evaluate a patient's topic maintenance, appropriateness, and ability to convey essential information without excessive tangential speech.

4.2.1. *Unsupervised machine learning*

The breadth of measures resulted in a very high-dimensional dataset, with each row representing a speech-language sample tied to a specific task and containing hundreds of linguistic features. This richness makes direct representation and visualization of the data structure impossible without dimensionality reduction.

We evaluated two main ML models. First, we used Principal Component Analysis (PCA), a linear dimensionality reduction method widely used in neurolinguistic research (Fridriksson et al., 2018). PCA detects orthogonal axes that capture the greatest variance, preserving

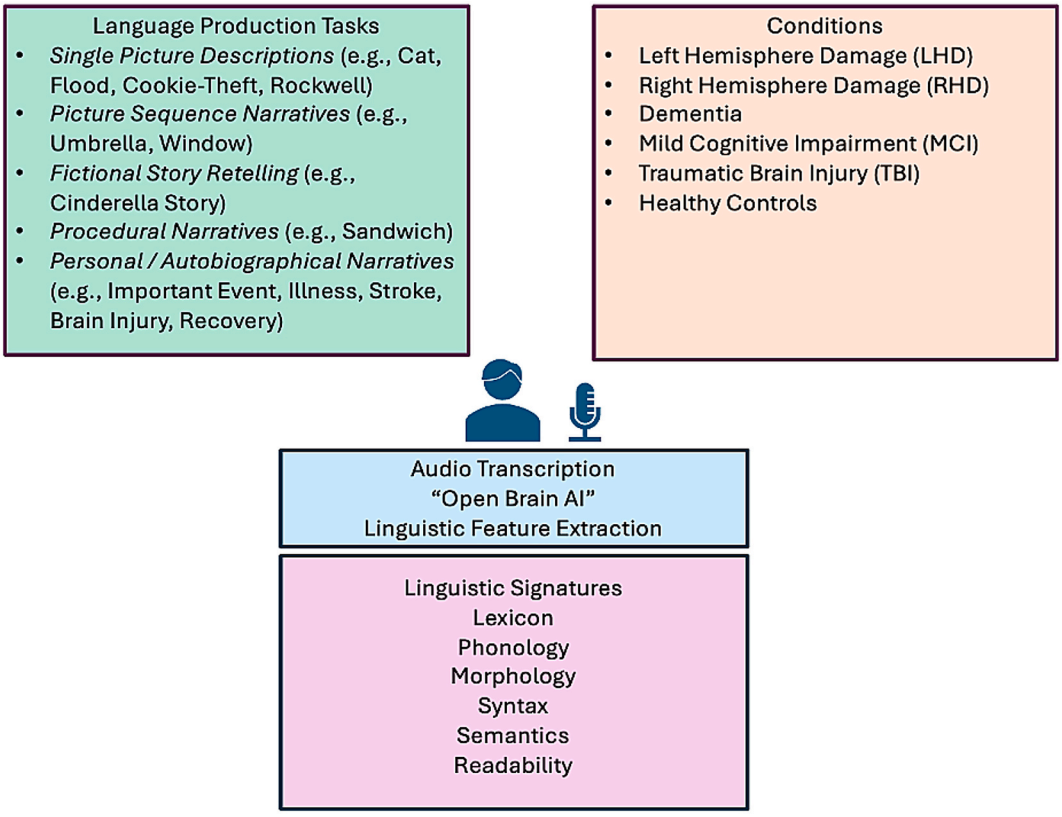


Fig. 1. This diagram illustrates the methodology for studying language production in neurological conditions. The study involves participants from multiple diagnostic groups (LHD, RHD, Dementia, MCI, TBI, and Healthy Controls) performing various language production tasks, including narrative tasks ("Cinderella"), procedural tasks ("how to make a sandwich"), and picture description tasks ("Cookie Theft"). The participants' recorded speech transcripts are processed through the Open Brain AI platform for automated linguistic feature extraction across multiple domains (lexicon, phonology, morphology, syntax, semantics, and readability). Data from the TalkBank consortium clinical banks is available by becoming a member of the TalkBank project.

global data structures but potentially missing non-linear patterns. To address this limitation, we conducted a second approach using Uniform Manifold Approximation and Projection (UMAP). UMAP is a non-linear dimensionality reduction method that constructs a high-dimensional graph of data points and optimizes a low-dimensional embedding to preserve local neighborhood structures (Ghojogh et al., 2023). To evaluate the quality of this non-linear embedding, we calculated the model's "trustworthiness", a metric that measures the extent to which local data neighborhoods in the high-dimensional space are accurately preserved in the lower-dimensional projection (Venna & Kaski, 2001). Trustworthiness on the full dataset for the UMAP model was 0.868, indicating a strong, reliable preservation of the original data structure.

4.2.2. Mixed effects models

To assess the influence of clinical diagnosis on each linguistic outcome variable, we utilized an automated mixed-effects modeling pipeline (a scripted sequence of data processing and statistical evaluation steps developed in R; R Core Team (2025). This pipeline was designed to be flexible, data-driven, and robust to the violations of statistical assumptions common in linguistic data.

A primary objective of this analysis was to evaluate the Task x Diagnosis interaction to identify features that vary significantly across communicative contexts within specific clinical groups. Although the Dementia cohort was included in the initial data cleaning and unsupervised analysis, it was excluded from the interaction modeling and post-hoc pairwise task comparisons because this group provided data for only a single task (Cookie Theft).

For each linguistic variable, a Generalized Linear Mixed-Effects Model (GLMM) was evaluated. The core of the analysis specified Task, Diagnosis, and their interaction (Task x Diagnosis) as fixed effects; in Diagnosis healthy controls was employed as the baseline. This allows the model to determine whether the linguistic differences between tasks depend on the patient's underlying diagnosis. To appropriately account for the non-independence of data (as multiple observations originated from the same individual), a random intercept was included for each participant via the (1 | SpeakerID) term. Given the unequal cohort sizes across diagnostic groups, mixed-effects modeling was selected in part for its suitability for unbalanced designs, while recognizing that such imbalance may still introduce differences in statistical power across groups. Including Speaker ID as a random intercept allowed the models to account for repeated observations within speakers and unequal numbers of observations across diagnostic categories while estimating the fixed effects of Task and Diagnosis.

The analytical pipeline systematically selected the most appropriate statistical model based on the distribution of each dependent variable. This adaptive process involved fitting Gaussian Linear Mixed-Effects Models (LMMs) for continuous variables, using robust LMMs if residual diagnostics (via the DHARMa package; Hartig (2016) indicated assumption violations, and employing GLMMs with appropriate distributions (e.g., binomial, Poisson, or negative binomial) for binary or count data. If a suitable model could not be fitted through these steps, a rank-based LMM was applied as a robust fallback.

When a significant Task x Diagnosis interaction was found ($p < 0.05$) via a Type III ANOVA, post-hoc pairwise comparisons were conducted to understand the nature of the interaction. These comparisons were made between tasks within each diagnostic group using estimated marginal means (via the emmeans package; Russell (2020), with Tukey's method applied to adjust for multiple comparisons.

5. Results

Our first research question aimed to determine whether linguistic measures differ significantly between tasks, and which features are most sensitive to task variation. To address this, we conducted an UMAP to visualize global data structures, followed by generalized linear mixed-effects models (GLMMs) to isolate the effects of Task, Diagnosis, and their interaction on each linguistic measure. Statistical models, full regression and post-hoc results are available in the Supplementary Materials and our Open Science Framework (OSF; https://osf.io/vud95/) repository.

5.1. Global linguistic variance across tasks

Initial dimensionality reduction using PCA revealed that the first two principal components captured ~33% of the total variance (PC1 = 28%, PC2 = 5%)—a substantial proportion given the high dimensionality of the dataset. PCA captures linear, global variance whereas the UMAP, we utilized better preserves local, non-linear neighborhood structures. The UMAP projection of all linguistic features (Figs. 2 and 3) yielded a trustworthiness score of 0.868. As seen in Fig. 2, the unsupervised algorithm naturally clustered the language samples based on the elicitation task.

When these specific tasks are grouped by their broader genres (Fig. 3), a clear continuum emerges. Tasks cluster from left to right, transitioning from highly constrained contexts (Single Pictures and Procedural tasks) to less constrained, generative contexts (Autobiographical and Fictional Retelling).

5.2. Most sensitive linguistic features

To determine which specific variables drove these task distinctions, we ranked the top ten most significant post-hoc contrasts for each diagnostic group based on absolute test statistics (Table 2). Higher absolute statistics indicate a stronger, more robust separation between two tasks for a given variable, while the estimate reflects the magnitude of that difference.

Table 2 highlights the top 10 most significant linguistic differences between speaking tasks per diagnostic group. Features are ranked from 1 to 10 by the absolute strength of their test statistic, either a *t*- or *z*-ratio, with Rank 1 representing the most robust difference. The table details the clinical group, the specific linguistic variable, and the direction of the difference, where > or < indicates higher or lower frequency in the first task. The estimated marginal mean difference reflects the magnitude of the gap between tasks, and all reported findings are highly significant, $p < 0.001$. Both the magnitude and configuration of these task effects were highly phenotype-specific: The HC group showed the most robust differentiation in phonological complexity, especially CCV structures, and overall lexical productivity, including *Types* and *Function Words*. The *Cinderella* narrative task elicited substantially higher complexity and third-person usage than procedural or descriptive tasks such as *Sandwich* and *Umbrella*; for example, the CCV contrast showed an Estimate of 71.68 and a test statistic of 47.43. Patients with LHD demonstrated the strongest task modulations in perspective and overall vocabulary richness, particularly Type Token Ratio. Notably, the autobiographical *Speech* task elicited significantly higher rates of first-person pronoun usage and higher TTR compared to structured narratives such as *Cinderella*, with a test statistic of −50.98. In patients with RHD, task sensitivity was almost entirely dominated by raw productivity and vocabulary volume, including *Types*, *Content Words*, and *Function Words*. The *Stroke* recovery narrative acted as a particularly strong elicitor for this group, producing significantly more diverse vocabulary than *Speech* or *Sandwich*; for example, the *Types* contrast showed an Estimate of −126.71. Patients with MCI displayed a subtler differentiation across tasks compared to the other cohorts, with test statistics in the range of approximately 14 to 16. Their task distinctions were driven more by verb morphology and tense, such as *Cookie Theft* eliciting higher Present Tense Ratios, and by specific syntactic ratios, including Direct Object Ratio. Patients with TBI showed a mixed profile of task sensitivity, balancing morphological features, such as Determiner Ratio and Infinitive Verbs, with perspective shifts, including Person Third Ratio, particularly when contrasting structured description, such as *Cat*, with unscripted monologue, such as *Speech*.

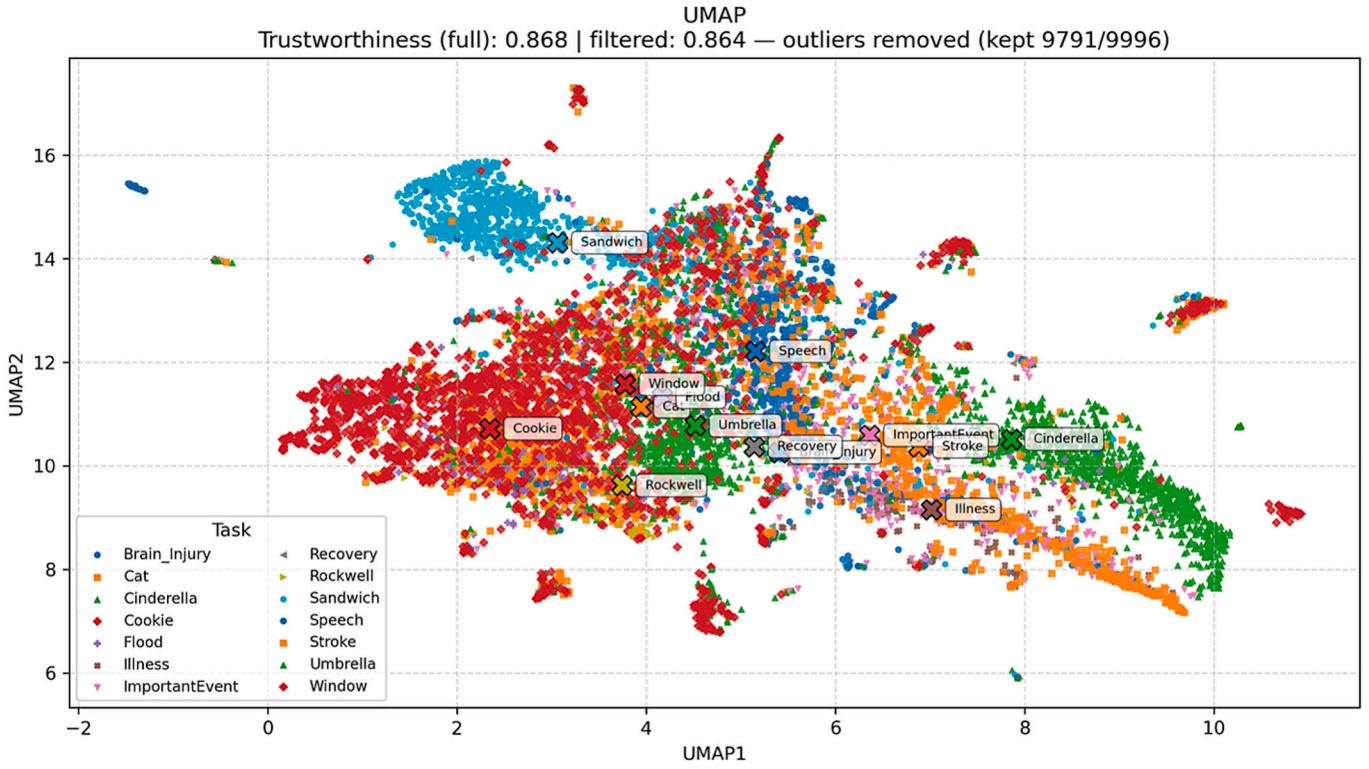


Fig. 2. UMAP projection of linguistic features (Colored by Specific Task). Each point represents a single language sample, colored and shaped by the specific assessment task. Outliers were excluded using IQR filtering. The 'x' symbol indicates the centroid of each task group.

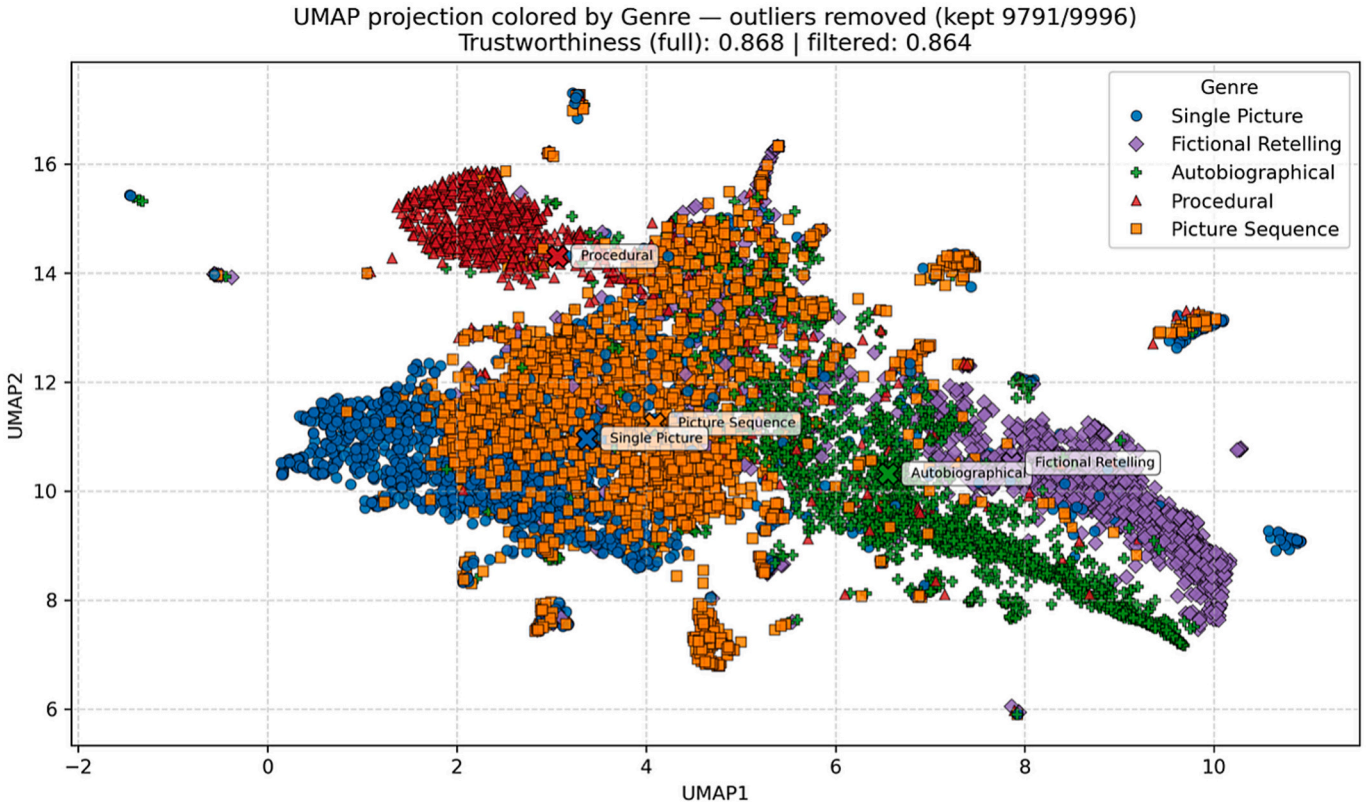


Fig. 3. UMAP projection of linguistic features (Colored by Task Genre). The same UMAP projection as Fig. 2, re-colored to highlight the broader genre categories: Single Picture Descriptions, Picture Sequence Narratives, Procedural Narratives, Personal/Autobiographical Narratives, and Fictional Retelling.

Table 2
Ranked Intra-Diagnostic Task Contrasts: Top 10 Discriminating Linguistic Features.

Rank	Group	Variable	Direction	EMM difference	Test Statistic	*p* value
1		CCV	Cinderella > Umbrella	71.68	47.43	<0.001
2		Person Third Count	Cinderella > Sandwich	85.81	46.48	<0.001
3		CCV	Cinderella > Window	69.24	46.09	<0.001
4		Person Third Ratio	Sandwich < Umbrella	−0.08	−45.36	<0.001
5		CCV	Cinderella > Sandwich	67.07	45.07	<0.001
6		Corrected TTR/CTTR	Cinderella > Sandwich	1.88	44.42	<0.001
7		Function Words (Unique)	Cinderella > Sandwich	31.38	43.89	<0.001
8		CCV	Cookie < Cinderella	−66.94	−43.8	<0.001
9		Types	Cinderella > Sandwich	121.12	43.73	<0.001
10	HC	Types	Cinderella > Window	119.97	42.86	<0.001
1		Person First Ratio	Cinderella < Speech	−0.05	−50.98	<0.001
2		Type Token Ratio/TTR	Cinderella < Speech	−0.27	−49.84	<0.001
3		Person First Ratio	Cinderella < Stroke	−0.05	−46.65	<0.001
4		CCV	Cinderella > Window	42.4	45.79	<0.001
5		Type Token Ratio/TTR	Speech > Stroke	0.26	45.5	<0.001
6		Function Words (Unique)	Sandwich < Stroke	−25.15	−45.14	<0.001
7		Person First Ratio	Cat < Speech	−0.05	−44.44	<0.001
8		Determiner Ratio	Cat > Speech	0.08	44.39	<0.001
9		Date Count	Stroke > Window	10.39	44.07	<0.001
10	LHD	Case Nominative Ratio	Cat < Speech	−0.04	−44.02	<0.001
1		Types	Speech < Stroke	−126.71	−14.24	<0.001
2		Types	Sandwich < Stroke	−123.14	−14.15	<0.001
3		Person First Ratio	Cinderella < Stroke	−0.06	−13.93	<0.001
4		Corrected TTR/CTTR	Sandwich < Stroke	−1.85	−13.93	<0.001
5		Content Words (Unique)	Speech < Stroke	−96.88	−13.76	<0.001
6		Function Words (Unique)	Sandwich < Stroke	−30.76	−13.69	<0.001
7		Content Words (Unique)	Sandwich < Stroke	−93.93	−13.64	<0.001
8		Person First Ratio	Cookie < Stroke	−0.06	−13.4	<0.001
9		Corrected TTR/CTTR	Speech < Stroke	−1.82	−13.37	<0.001
10	RHD	Person First Ratio	Cinderella < Speech	−0.06	−13.21	<0.001
1		Direct object Ratio	Rockwell < Sandwich	−0.08	−16.96	<0.001
2		Tense Present Ratio	Cookie > Cinderella	0.06	15.57	<0.001
3		Corrected TTR/CTTR	Cinderella > Sandwich	1.49	15.51	<0.001
4		Person Third Ratio	Cinderella > Sandwich	0.06	15.45	<0.001
5		Tense Present Ratio	Cookie > Sandwich	0.06	15.19	<0.001
6		VerbForm Infinitive Ratio	Rockwell < Sandwich	−0.05	−15.16	<0.001
7		Gender Feminine Ratio	Cinderella > Rockwell	0.03	14.76	<0.001
8		Mood Indicative Ratio	Cookie > Sandwich	0.04	14.72	<0.001
9		Function Words (Unique)	Cinderella > Sandwich	23.63	14.61	<0.001
10	MCI	Aspect Progressive Ratio	Cookie > Cinderella	0.03	14.17	<0.001
1		Determiner Ratio	Cat > Speech	0.1	16.78	<0.001
2		Gender Masculine Ratio	Cinderella < Umbrella	−0.04	−16.74	<0.001
3		VerbForm Infinitive Ratio	Sandwich > Window	0.06	16.44	<0.001
4		Type Token Ratio/TTR	Cinderella < Speech	−0.31	−16.36	<0.001
5		Direct object Ratio	Sandwich > Speech	0.08	16.3	<0.001
6		Person Third Ratio	Recovery < Umbrella	−0.07	−16.2	<0.001
7		Person Third Ratio	Brain_Injury < Umbrella	−0.07	−15.74	<0.001
8		Person First Ratio	Cinderella < Speech	−0.06	−15.6	<0.001
9		Person Third Ratio	Sandwich < Umbrella	−0.07	−15.59	<0.001
10	TBI	VerbForm Infinitive Ratio	ImportantEvent < Sandwich	−0.05	−15.12	<0.001

Note: Features are ranked by the absolute magnitude of their test statistics (t or z ratios) derived from post-hoc pairwise comparisons of tasks within each diagnostic group. Only features exhibiting a significant Task x Diagnosis interaction ($p < 0.001$) are included. The Dementia cohort was excluded from this specific analysis as data were limited to a single task (Cookie), precluding within-group task contrasts. LHD = Left hemisphere damage; RHD = Right hemisphere damage; MCI = Mild Cognitive Impairment; TBI = Traumatic Brain Injury; EMM difference = Estimated Marginal Mean Difference.

5.3. Main and interaction effects on linguistic domains

To address our second research question regarding how specific neurological conditions influence unique linguistic patterns within each task, we evaluated the ANOVA outputs from the GLMMs. For presentation purposes, we group these findings by linguistic domain. (Note: Because separate models were fitted for each variable, multicollinearity between dependent variables did not impact individual model estimates).

5.3.1. Phonology

Table 3 presents the ANOVA output for word lengths and syllable structures. The results highlight a pervasive main effect of Task across all variables, indicating that task demands heavily influence phonological output. A main effect of Diagnosis was present for total syllables and specific word lengths. Widespread Task x Diagnosis interaction effects were also observed, indicating that task demands differentially amplify phonological vulnerabilities depending on the patient's specific neurological profile.

5.3.2. Morphology - part of speech measures

Due to the extensive number of variables analyzed, the full results for Part of Speech measures are shown in Appendix 2 to maintain the readability of the main text. Task was significant in all measures; the diagnosis was also significant for most measures, except for Conjunct Count, Coordinating conjunction Count, Expletive Count, Noun Ratio, Numeral Count, Numeral Ratio, Pronoun Count, Pronoun Ratio, Proper noun Count (n.s.). The effect of the task x diagnosis interaction on Part of Speech Measures was significant in most distributions, except for Expletive Counts.

Table 3
Phonological analysis results by syllable count and syllable structure patterns: results from the ANOVA output on the glmer models.

Variable	Task Effect	Diagnosis Effect	Interaction Effect
1 syllable word	***	n.s.	***
2 syllables word	***	***	***
3 syllables word	***	**	***
4 syllables word	***	n.s.	***
5 syllables word	***	n.s.	**
CCCV	***	n.s.	n.s.
CCCVC	***	*	n.s.
CCV	***	n.s.	***
CCVC	***	n.s.	***
CCVCC	***	n.s.	***
CCVCCC	***	n.s.	n.s.
CV	***	***	***
CVC	***	***	***
CVCC	***	***	***
CVCCC	***	***	***
V	***	n.s.	***
VC	***	n.s.	***
VCC	***	n.s.	***
Syllables	***	**	***
Total Characters in Text All symbols	***	*	***
Total Characters in Text Letters Only	***	**	***

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). "n.s." indicates non-significant results. C = consonant, V = vowel in syllable structure notation.

Appendix 2. Morphological analysis results from the Anova output on the glmer models.

Variable	Task	Diagnosis	Interaction
Adjective Count	***	***	***
Adjective Ratio	***	***	*
Adposition Count	***	***	***
Adposition Ratio	***	***	***
Adverb Count	***	*	***
Adverb Ratio	***	***	***
Auxiliary Count	***	**	***
Auxiliary Ratio	***	***	***
Conjunct Count	***	n.s.	***
Conjunct Ratio	***	***	***
Coordinating conjunction Count	***	n.s.	***
Coordinating conjunction Ratio	***	***	***
Determiner Count	***	**	***
Determiner Ratio	***	***	***
Expletive Count	***	n.s.	n.s.
Expletive Ratio	***	***	***
Interjection Count	***	***	***
Interjection Ratio	***	***	***
Noun Count	***	**	***
Noun Ratio	***	n.s.	***
Numeral Count	***	n.s.	**
Numeral Ratio	***	n.s.	***
Other Count	***	***	***
Other Ratio	***	***	***
Particle Count	***	***	***
Particle Ratio	***	***	***
Pronoun Count	***	n.s.	***
Pronoun Ratio	***	n.s.	***
Proper noun Count	***	n.s.	***
Proper noun Ratio	***	***	***
Subordinating conjunction Count	***	***	***
Subordinating conjunction Ratio	***	***	***
Verb Count	***	***	***
Verb Ratio	***	***	***

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

5.3.3. Morphology - inflectional morphology

For readability, the comprehensive statistical results table for inflectional morphology are provided in Appendix 3. Across nearly all categories of inflectional morphology, task effects were highly significant: verb aspect, tense, mood, voice, form, modality, number, gender, definiteness, possession, case, comparison, degree, pronoun type, and polarity all varied systematically by the discourse task. This means that the linguistic context provided by the task strongly shapes how morphological features appear, even when diagnosis alone was not significant.

At the same time, diagnosis and its interaction with task modulated these effects. For example, groups differed in how they marked tense and aspect, produced passive forms, used complex verb forms, and handled pronouns, gender, and possession. These differences often appeared in the ratios (relative frequency of forms), highlighting that diagnostic groups vary not only in whether features are present but in how heavily they are relied upon relative to other structures. Taken together, this shows that task type is a powerful driver of inflectional morphology, while the specific diagnostic group (including the healthy control baseline) shapes the strength and distribution of these effects.

5.3.4. Syntax

Syntactic measures represent the constructions of sentences and phrases providing information about their role in the syntactic hierarchy (Table 4). For presentation purposes we break these measures down into Sentence Complexity and Structural Measures, Clauses, Sentences, Phrase Types and Units Counts and Ratios, and Grammatical Dependency Types (Modifiers and Complements) (Appendix 4).

These features in Table 4 indicate robust task effects for structural complexity, with more complex syntax in certain task types. Also, we observed significant Task × Diagnosis interaction effects across all structural measures. Although effects of diagnosis do emerge in measures related to syntactic embedding depth (e.g., Yngve Load, tree height) and passive constructions, these differences are highly task dependent, suggesting that even though some diagnostic groups produce simpler or flatter syntactic trees or struggle with complex sentential structures, the structural limitations are dependent on the specific task demands rather than being an independent, overarching deficits.

There is strong evidence that syntactic packaging—how clauses and phrases are organized into T-units, sentences, and phrase structures—is shaped by the interaction between task demands and diagnostic group. Although some of the measures do not show an effect of diagnosis (Table 5); there are significant effects with respect to clausal complexity (e.g., dependent clauses) and phrasal expansion (complex nominals), depending on the task being performed.

5.3.5. Semantics

Like in the previous measures, the interaction (Table 6) between task and diagnosis results in significant effects for Cardinal Numbers, Dates, Organizations, and Persons, suggesting that semantics are affected by the interaction between Diagnosis and Task. That said, diagnosis had no independent significant effects for measures regarding Numbers, the total number of Dates and Persons in speech productions.

5.3.6. Readability

Table 7 presents the results of ANOVA tests conducted on GLMMs examining the effects of task, diagnosis, and their interaction on ten established readability metrics. These indices capture complementary dimensions of textual complexity, including syntactic sophistication (e.g., Gunning Fog, SMOG), lexical difficulty (e.g., Dale–Chall, Difficult Words), and composite readability formulas (e.g., Flesch–Kincaid Grade Level, Automated Readability Index).

Task type exerted a highly significant main effect ($p < 0.001$) across all readability measures, confirming that textual complexity varied systematically as a function of task demands. The main effect of diagnosis was significant for nine of the ten indices ($p < 0.001$ for most; $p <$

Appendix 3. Inflectional morphology results from the Anova output on the glmer models.

Variable	Task	Diagn.	Inter.	Variable	Task	Diagn.	Inter
Aspect Perfective Count	***	*	***	Number Singular Ratio	***	***	***
Aspect Perfective Ratio	***	***	***	NumType Cardinal Count	***	n.s.	n.s.
Aspect Progressive Count	***	***	***	NumType Cardinal Ratio	***	**	n.s.
Aspect Progressive Ratio	***	n.s.	***	Person First Count	***	**	***
Auxiliary passive Count	***	n.s.	***	Person First Ratio	***	***	***
Auxiliary passive Ratio	***	***	***	Person Second Count	***	n.s.	**
Case Accusative Count	***	n.s.	***	Person Second Ratio	***	n.s.	***
Case Accusative Ratio	***	*	***	Person Third Count	***	n.s.	***
Case marker Count	***	n.s.	***	Person Third Ratio	***	***	***
Case marker Ratio	***	***	***	Polarity Negative Count	***	n.s.	n.s.
Case Nominative Count	***	n.s.	***	Polarity Negative Ratio	***	**	n.s.
Case Nominative Ratio	***	***	***	Poss Possessive Count	***	***	***
ConjType Cmp Count	***	n.s.	***	Poss Possessive Ratio	***	***	***
ConjType Cmp Ratio	***	*	***	PronType Article Count	***	n.s.	***
Dative Count	***	n.s.	n.s.	PronType Article Ratio	***	n.s.	***
Dative Ratio	***	n.s.	***	PronType Demonstrative Count	***	n.s.	***
Definite Definite Count	***	n.s.	***	PronType Demonstrative Ratio	***	***	***
Definite Definite Ratio	***	***	***	PronType Indefinite Count	***	n.s.	n.s.
Definite Indefinite Count	***	*	***	PronType Indefinite Ratio	***	n.s.	n.s.
Definite Indefinite Ratio	***	n.s.	***	PronType Personal Count	***	*	***
Degree Cmp Count	***	n.s.	n.s.	PronType Personal Ratio	***	n.s.	***
Degree Cmp Ratio	***	n.s.	n.s.	PronType Relative Count	***	n.s.	n.s.
Degree Positive Count	***	**	***	PronType Relative Ratio	***	*	*
Degree Positive Ratio	***	n.s.	n.s.	Tense Past Count	***	**	***
Gender Feminine Count	***	n.s.	***	Tense Past Ratio	***	***	***
Gender Feminine Ratio	***	n.s.	***	Tense Present Count	***	n.s.	***
Gender Masculine Count	***	**	***	Tense Present Ratio	***	**	***
Gender Masculine Ratio	***	**	***	VerbForm Finite Count	***	n.s.	***
Gender Neuter Count	***	n.s.	***	VerbForm Finite Ratio	***	n.s.	***
Gender Neuter Ratio	***	***	***	VerbForm Infinitive Count	***	*	***
Mood Indicative Count	***	n.s.	***	VerbForm Infinitive Ratio	***	**	***
Mood Indicative Ratio	***	***	***	VerbForm Participle Count	***	***	***
Number Plural Count	***	n.s.	***	VerbForm Participle Ratio	***	***	***
Number Plural Ratio	***	***	***	VerbType Mod Count	***	n.s.	n.s.
Number Singular Count	***	*	***	VerbType Mod Ratio	***	n.s.	n.s.

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Table 4
Sentence Complexity and Structural Measures results from the Anova output on the *glmer* models.

Variable	Task	Diagnosis	Interaction
Average Sentence Length	***	***	***
Mean Sentence Length in Words	***	***	***
Max Sentence Length	***	***	***
Min Sentence Length	***	***	***
Average Tree Height	***	***	***
Total Tree Height	***	n.s.	***
Max Classical Yngve Load	***	*	***
Mean Classical Yngve Load	***	***	***
Total Classical Yngve Load	***	n.s.	***
Max Left Branching Depth	***	n.s.	***
Mean Left Branching Depth	***	***	***
Passive Sentences Percent	***	***	***

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

0.05 for Estimated Reading Time), with only the Difficult Words metric failing to reach significance ($p > 0.05$). This pattern suggests that, overall, diagnostic groups differed in the readability profiles of their output, though group differences were less consistent for measures focused exclusively on low-frequency vocabulary. Critically, a highly significant task x diagnosis interaction ($p < 0.001$) emerged for every readability metric. This consistent interaction indicates that group differences in textual complexity are not stable across contexts but are instead contingent on task characteristics. For example, one diagnostic group may produce texts of comparable readability to the other in a structured task but show markedly greater (or lesser) complexity in a free-production task.

5.3.7. Lexical

Table 8 presents the results of ANOVA tests conducted on glms to examine the effects of task, diagnosis, and their interaction on a range of lexical measures. The variables are organized into three conceptual categories: (1) overall lexical output (total words, average word length, and total tokens), (2) content versus function word distribution (including total and unique counts for both categories, as well as automated lexical density), and (3) lexical richness/diversity indices (Maas's TTR, CTTR, and standard TTR). We conducted independent measures to evaluate how diagnosis and task influence each one of these measures independently, rather than examining how each measure correlated with each other. Inter-correlation is expected since they measure similar constructs. Across all measures, task type exerted a highly significant main effect ($p < 0.001$), indicating that lexical output varied substantially depending on task demands. The main effect of diagnosis was non-significant for global output measures (total words, total tokens, and total function words) but reached significance for average word length, unique content and function words, lexical density, and all three richness indices. Most notably, a highly significant task x diagnosis interaction emerged across every lexical variable. This consistent interaction suggests that group differences in lexical complexity, diversity, and composition are not uniform but are systematically modulated by task and diagnosis characteristics.

6. Discussion

In this study, we examined the comprehensive linguistic profiles of patient-produced discourse and the profound influence that the elicitation task exerts on these metrics. A central question we addressed is the validity of comparing speech productions across different tasks to draw conclusions about an underlying pathology. Consequently, we question

Appendix 4. Grammatical Dependency Types (Modifiers and Complements) results from the Anova output on the glmer models.

Variable	Task	Diagnosis	Interaction
Adjectival complement Count	***	n.s.	***
Adjectival complement Ratio	***	***	***
Adjectival modifier Count	***	***	***
Adjectival modifier Ratio	***	***	***
Adverbial clause modifier Count	***	**	***
Adverbial clause modifier Ratio	***	***	***
Adverbial modifier Count	***	*	***
Adverbial modifier Ratio	***	***	***
Appositional modifier Count	***	n.s.	***
Appositional modifier Ratio	***	***	n.s.
Attribute Count	***	n.s.	***
Attribute Ratio	***	***	***
Clausal complement Count	***	n.s.	***
Clausal complement Ratio	***	*	***
Clausal modifier of noun Count	***	n.s.	*
Clausal modifier of noun Ratio	***	***	***
Complement of preposition Count	***	n.s.	n.s.
Complement of preposition Ratio	***	***	***
Compound modifier Count	***	n.s.	***
Compound modifier Ratio	***	***	***
Direct object Count	***	***	***
Direct object Ratio	***	***	***
Marker Count	***	n.s.	***
Marker Ratio	***	***	***
Negation modifier Count	***	n.s.	n.s.
Negation modifier Ratio	***	n.s.	n.s.
Nominal subject Count	***	n.s.	***
Nominal subject Ratio	***	**	***
Nominal subject passive Count	***	n.s.	***
Nominal subject passive Ratio	***	***	***
Number modifier Count	***	*	n.s.
Number modifier Ratio	***	n.s.	***
Object of preposition Count	***	***	***
Object of preposition Ratio	***	***	***
Open clausal complement Count	***	***	***
Open clausal complement Ratio	***	***	***
Possession modifier Count	***	***	***
Possession modifier Ratio	***	***	***
Prepositional modifier Count	***	***	***
Prepositional modifier Ratio	***	***	***
Relative clause modifier Count	***	n.s.	***
Relative clause modifier Ratio	***	*	***
Root Count	***	n.s.	***
Root Ratio	***	***	***
Unclassified dependent Count	***	n.s.	**
Unclassified dependent Ratio	***	***	n.s.

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Table 5
Clauses, Sentences, Phrase Types and Units Counts and Ratios results from the Anova output on the *glmer* models.

Variable	Task	Diagnosis	Interaction
Total Sentences	***	n.s.	***
Sentences Alphabetic Only	***	n.s.	***
Total T units	***	*	***
Total Complex T units	***	***	***
Total Clauses	***	*	***
Clauses Total Clauses in Text	***	n.s.	***
Total Dependent Clauses	***	***	***
Total Coordinate Phrases	***	n.s.	***
Total Complex Nominals	***	***	***
Adjective Phrases	***	***	***
Adverbial Phrases	***	*	***
Noun Phrases	***	*	***
Verb Phrases	***	n.s.	***
Prepositional Phrases	***	***	***

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

the customary practice of using generalized normative corpora without

Table 6
Semantic results from the Anova output on the *glmer* models.

Variable	Task	Diagnosis	Interaction
Cardinal Number Count	***	n.s.	***
Cardinal Number Ratio	***	n.s.	***
Date Count	***	n.s.	***
Date Ratio	***	***	***
Organization Count	***	***	***
Organization Ratio	***	***	***
Person Count	***	n.s.	***
Person Ratio	***	***	***

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Table 7
Readability results from the Anova output on the *glmer* models.

Variable	Task	Diagnosis	Interaction
Automated Readability Index	***	***	***
Coleman Liau Index	***	***	***
Dale Chall Readability Score	***	***	***
Difficult Words	***	n.s.	***
Estimated Reading Time *sec*	***	*	***
Flesch Kincaid Grade Level	***	***	***
Flesch Reading Ease	***	***	***
Gunning Fog Index	***	***	***
Linsear Write Formula	***	***	***
Smog Index	***	***	***

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Table 8
Lexical results from the Anova output on the *glmer* models.

Variable	Task	Diagnosis	Interaction
Overall Measures			
Total Words	***	ns	***
Average Word Length (in Characters)	***	***	***
Number of Tokens	***	ns	***
Content vs. Function Words			
Content Words Total	***	ns	***
Content Words Unique	***	***	***
Function Words Total	***	ns	***
Function Words Unique	***	***	***
Automated Lexical Density Content Words / Total Words	***	*	***
Lexical Richness			
Maas s TTR A2	***	***	***
Corrected TTR CTTR	***	***	***
Type Token Ratio (TTR)	***	***	***

Note: Asterisks (*) indicate statistical significance levels, with more asterisks representing stronger significance (e.g., * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

rigorously controlling for the confounding effect of the task itself. Currently, tasks as varied as the 'Cookie Theft' picture description, procedural explanations, and fictional story retells are sometimes treated interchangeably in both clinical assessment and research. However, it is imperative to model how specific tasks shape language production before drawing clinical conclusions from comparisons between diagnostic groups. As the findings make evident, comparing linguistic output across individuals or groups without accounting for the specific task or even the specific visual stimulus creates a significant measurement confound. This is not just a methodological problem; it is a fundamental barrier to achieving reliable, reproducible measures across diagnostic groups and longitudinal timepoints from discourse and other narrative data (Stark et al., 2025; Stark, Bryant, et al., 2023; Stark & Fukuyama, 2021). These confounds are only magnified when

considering cross-linguistic typological differences, where the grammatical features elicited by the same conceptual task can differ fundamentally depending on the language's reliance on word order versus morphology. This also suggests that failing to control for task-induced variability risks the development of flawed machine-learning algorithms and unreliable diagnostic markers. Furthermore, studies like Senthilkumar et al. (2026) show that simple picture description tasks from the WAB-R are not sensitive enough to catch mild impairments; patients score well on the test but still struggle in real life. This study suggests that to further evaluate mild impairments, we must explore different micro-linguistic features, because mild deficits only reveal themselves under differential cognitive load in different language tasks.

A systematic investigation of these task effects is therefore essential for advancing the accuracy and validity of language assessment in neurology. Even though previous literature has increasingly recognized that connected speech varies significantly depending on the elicitation method (Stark & Fukuyama, 2021; Wright & Capilouto, 2009) the present study scales this concept dramatically. We advance beyond prior work—which has typically relied on a narrow subset of measures per domain, such as mean length of utterance for syntax or type-token ratio for the lexicon—by demonstrating that task-diagnosis interactions permeate a much wider set of linguistic structures than previously documented. By applying an automated computational pipeline to a large number features spanning phonology, morphology, syntax, semantics, lexicon, and readability, we show that specific discourse tasks generate distinct, reproducible linguistic "fingerprints" (sometimes called 'profiles') across multiple domains. This reinforces the argument that different tasks cannot simply be collapsed into a single "connected speech" category and underscores the critical need for task-specific normative data in both research and clinical interpretation.

6.1. *The interaction of task and diagnosis in linguistic production*

Our findings show significant interaction effects between task and diagnosis across all major linguistic domains. This indicates that the linguistic and cognitive profile of a neurological condition is not static (Aamodt et al., 2021; Cousins et al., 2021; Delaby et al., 2022); rather, it manifests differently depending on the specific cognitive and linguistic demands of the task at hand (Ash et al., 2013; Bose et al., 2021; Giannini et al., 2017). This is demonstrated well in Fig. 2, where the spatial arrangement of tasks suggests a graded continuum. *Cookie Theft* (a simple single picture description task with few actions and characters) and *Sandwich* (a simple procedural narrative with relatively few steps) occupied the most constrained end of the embedding, followed by other visually or procedurally scaffolded tasks such as *Cat* (single picture description), *Rockwell* (single picture descriptions that range in complexity), *Window* (four-panel picture sequence description), Umbrella (six-panel picture sequence description), and *Flood* (single picture description of emotionally-charged, real picture); guided dialogue-like tasks such as *Speech* ("how is your speech these days?") and *Recovery* ("tell me, what have you been doing to help with your recovery?") occupied an intermediate position; and the least constrained region was defined by autobiographical narratives (like *Important Event*) and fictional retell (*Cinderella*). This distribution suggests that task effects are shaped not only by broad genre labels (e.g., picture description, narrative), but by the degree of contextual support (e.g., visual present vs. not present), complexity (e.g., number of actors, actions), discourse planning (e.g., number of events, story grammar), and memory-based generation required by each elicitation condition.

In the domain of phonology—which is frequently impacted first by a bilateral temporal damage (Hickok & Poeppel, 2007) and as it integrates with higher level language processing by left hemisphere fronto-temporal damage (Best et al., 2002; Meteyard & Bose, 2018). We found that both word length (by syllable count) and the complexity of syllable structures are significantly influenced by the interplay of task and diagnosis. More demanding tasks appear to amplify phonological deficits, revealing vulnerabilities in motor planning and articulatory control that might be masked in simpler contexts. For example, a patient with underlying motor-speech or phonological deficits might produce simple consonant-vowel (CV) structures adequately during a straightforward naming or descriptive task, but exhibit marked simplification of complex clusters (e.g., CCCV) when burdened by the cognitive load of a demanding narrative task. Although highly complex clusters like CCCV (e.g., *spry, straw, scream,* and *splash)* are naturally infrequent in English, retaining these metrics allows our models to detect when a patient's avoidance of such structures deviates significantly from the healthy control baseline. These findings suggest that phonological performance is not task-invariant but may deteriorate when more complex discourse demands place additional pressure on an already vulnerable production system.

Core morphological measures are traditionally critical for detecting patterns of linguistic breakdown, such as those seen in clinical agrammatism or anomia (Berndt et al., 1996; Caramazza & Zurif, 1976; Friedmann, 2002). Our findings demonstrate that the distribution of parts of speech and inflectional morphology are highly sensitive to the interaction between task and diagnosis; consequently, studies employing different discourse tasks cannot be assumed to sample equivalent linguistic behavior. As demonstrated by the unsupervised ML models (Fig. 2), the arrangement of linguistic clusters follows the degree of contextual constraint: highly structured activities (e.g., *Cookie Theft*) are followed by instruction-based tasks (e.g., procedural narratives like sandwich making), dialogic exchanges (e.g., *Speech* and *Recovery,* in this sample), and finally open-ended narratives. From a clinical perspective, this progression reflects a shift in contextual scaffolding, a pattern that has also been observed in earlier, more limited modeling studies focused on aphasia and healthy controls (Stark, 2019; Stark & Fukuyama, 2021). Structured tasks provide clear, shared visual or procedural cues, which inherently constrain the expected lexicon and syntactic frameworks, whereas unscripted tasks require the speaker to independently establish the discourse structure and referential context. Crucially, the degree of shared context also appears to modulate morphosyntactic load, such that more constrained tasks permit successful communication with simpler lexical and grammatical means, whereas less constrained tasks require greater reliance on explicit inflectional marking, referential maintenance, and clause-level organization. In the weakliest scaffolded narratives (e.g., *open-ended narratives)*, speakers must establish the discourse context independently, increasing the need for morphosyntactic markers such as tense, aspect, and more complex pronominal reference to specify temporal, spatial, and interpersonal relations. Under these conditions, failures in grammatical morphology may become evident even when they remain relatively masked in simple picture description.

Even though syntactic deficits were not in the most distinguishing features for groups when collapsed across task (Table 2), independent main effects of syntactic measures (Tables 4, 5) did become significant when task was considered. Specifically, measures of syntactic complexity show significant interaction effects as measures like sentence length, clausal embedding (e.g., dependent clauses), and hierarchical structure (e.g., tree height) varied based on the specific pairing of task and diagnosis (Agmon et al., 2024; Cheung & Kemper, 1992; Kemper, 1987; Kyle & Crossley, 2017; Lan et al., 2022). This suggests that tasks with higher narrative demands force the production of more complex syntax, thereby revealing latent deficits in sentence planning. This has been shown in aphasia and neurotypical controls for simple measures like noun/verb ratio (Stark, 2019). It should also be noted that these syntactic effects on task should not be assumed to result from the same underlying mechanism. For instance, the reduced production of complex nominals or dependent clauses in a narrative task by a patient with TBI may reflect impaired executive functions impacting global discourse planning, representing a fundamentally different underlying mechanism than grammatical simplification seen in LHD.

However, a keen eye may have also noticed that, in Fig. 2, the 'least'

complex procedural narrative (Sandwich) appeared distinctively different from the more 'complex' narratives (*Cinderella, Stroke*). Although procedural discourse occupied a relatively constrained position in the global task space, selected syntactic features nonetheless differentiated procedural discourse from autobiographical narrative and story retell within specific phenotypes, particularly LHD. One possible interpretation is that procedural discourse offers a distinct scaffold for speakers with LHD: because it relies on familiar, overlearned action schemas and predictable linear organization, it may reduce demands on episodic retrieval, discourse framing, and long-range referential maintenance while still supporting certain forms of clause linkage. Thus, the apparent advantage of procedural discourse for selected syntactic features should not be interpreted as evidence that the task is globally more complex, but rather that its predictable action schema reduces higher-level discourse and memory demands, allowing speakers to maintain certain local syntactic linkages despite broader language impairment.

Regarding semantics, the ability to retrieve specific content, as measured by named-entity recognition, was also significantly modulated by task context (Boyle, 2010; Kiran & Thompson, 2003). Even when a broad diagnosis alone did not have a main effect, its interaction with a specific task often did. A personal event narrative directly probes the recall of temporal and personal information, making it a highly sensitive tool for detecting impaired temporal orientation or proper-name retrieval deficits; vulnerabilities that a generic, visually constrained picture description might completely miss.

Readability metrics, which synthesize features such as word choice, sentence length, and syntactic complexity, offer a broad index of textual accessibility. Our analyses showed that the apparent "readability" of a patient's discourse fluctuated substantially as a function of task. A speaker may produce relatively simple and therefore superficially "readable" output in one context because of pathological syntactic simplification, yet produce denser but error-prone, disorganized, or neologistic language in another. This underscores that readability scores do not map straightforwardly onto communicative success or linguistic health in neurogenic populations; rather, they behave as emergent properties of the interaction between task demands and impairment profile. Accordingly, a single readability score is insufficient for characterizing communicative effectiveness or tracking impairment across contexts unless interpreted in relation to elicitation type and the underlying linguistic features from which it is derived.

Lexical production differed substantially across tasks. Task effects were noted for every measure examined, across the broad categories of total output, content and function word use, and lexical richness. Diagnostic groups did not differ consistently in total output but did differ in average word length, unique content and function words, lexical density, and lexical richness. Together, these findings suggest that broad measures of how much a person says may be less diagnostically informative than measures capturing the variety and composition of the vocabulary produced. Further, every lexical variable measured showed a significant interaction between task and diagnosis, suggesting that these lexical differences identified for diagnostic groups depended on the eliciting task. As with other variables discussed above, lexical performance therefore cannot be interpreted independently of task context (which has been shown before in aphasia, e.g., Stark, 2019).

Specifically, prior discourse research has shown that elicitation context systematically shapes language output, that different discourse genres place different demands on lexical, syntactic, and discourse-level processing, and that even psychometric reliability can vary by task (Stark, 2019; Stark, Alexander, et al., 2023; Stark & Fukuyama, 2021). This study has established that discourse tasks are not interchangeable and that clinicians should be cautious about drawing broad conclusions from a single elicitation context. The present study builds on that literature by demonstrating the scope of these task effects across a substantially broader clinical and linguistic landscape. Across 15 elicitation tasks, multiple diagnostic groups, and a large number of features spanning phonology, morphology, syntax, semantics, readability, and lexicon, we found pervasive Task x Diagnosis interactions and clear clustering by task in the unsupervised analyses. Importantly, these differences were not limited to broad genre contrasts: even ostensibly similar tasks within genre, such as Cookie Theft, Cat, and Rockwell, or Cinderella retell and autobiographical narrative, yielded distinct linguistic profiles.

Clinically, these findings suggest that task selection should be guided by the question the clinician is trying to answer and by the degree of contextual support the patient is likely to need. For example, when the goal is to evaluate lexical retrieval under conditions of high external support, single-picture descriptions may be most informative because they constrain the topic while still requiring access to target vocabulary. When the goal is to examine temporal-causal linkage or event sequencing, picture-sequence tasks may be preferable because they add demands on connecting events across panels. When the goal is to assess morphosyntactic production in a familiar and relatively scaffolded context, procedural discourse may be especially useful, as it reduces discourse-framing demands while preserving pressure on local grammatical organization.

By contrast, when the goal is to evaluate higher-level discourse abilities such as syntactic complexity and referential maintenance, autobiographical and fictional narrative tasks may be more sensitive because they require the speaker to generate and sustain structure with far less external scaffolding. To reverse this logic, our data suggests that it would not be goal-aligned to use a single picture description task to comprehensively evaluate syntactical complexity, for example. Thus, rather than treating discourse sampling as a single interchangeable procedure, we recommend that clinicians select tasks according to the linguistic domain of interest, the amount of contextual support required (e.g., whether the client requires the support of a picture to aid in their production), and the nature of the suspected impairment. Further, when the situation allows for acquisition of discourse through the strategic pairing of tasks, such as acquiring spoken language data from a single picture description and a narrative retell, offer enhanced ability to evaluate discourse ability more comprehensively.

Our results indicate that tasks cannot be treated as interchangeable even within the same broad genre, as apparently similar elicitation contexts such as *Cookie Theft, Cat,* and *Rockwell,* or *Cinderella* retell and personal event narrative, yielded distinct linguistic profiles. Thus, clinically meaningful assessment requires not only matching task type to the linguistic construct of interest but also recognizing that specific stimuli within a genre impose different lexical, syntactic, and discourse-planning demands, albeit subtle. Developing genre- and task-specific normative data will allow clinicians to select the probe that best aligns with their diagnostic question.

Furthermore, the use of automated computational pipelines makes this level of profiling feasible at scale, eliminating the time-prohibitive burden of manual transcription. Although this study successfully mapped the comprehensive linguistic landscape across broad diagnostic categories (LHD, RHD, TBI, MCI) using a large number of features, the inherent statistical and clinical redundancy among these numerous metrics means that moving the field forward requires distilling this complexity into practical clinical applications. Future research must utilize feature-selection techniques to isolate the 10 to 15 most highly discriminative markers for each specific pathology to eliminate this redundancy. Moreover, because substantial heterogeneity exists within our broad groups, applying this computational framework to narrowly defined clinical subtypes (e.g., Broca's versus Wernicke's aphasia, or amnestic versus non-amnestic MCI) is the critical next step (Kim et al., 2024). At the same time, although distilling high-dimensional linguistic data into smaller clinically usable feature sets is an important goal, such reduction should be approached cautiously. Narrowing analysis to subtype- or task-specific markers may improve interpretability, but it also risks obscuring meaningful individual variability, particularly in clinically heterogeneous populations. Features that appear statistically redundant at the group level may still capture important person-specific

differences in lexical choice, syntactic organization, discourse planning, or compensatory strategy use. Thus, the next step is not simply to identify a fixed set of markers for each pathology, but to determine which features are both robustly discriminative and sufficiently flexible to accommodate within-group variability. Future work should therefore aim to develop clinically interpretable models that combine reduced feature sets with individualized profiling, allowing discourse assessment to remain sensitive both to diagnostic group patterns and to the unique linguistic presentation of each speaker. Translating these dense, high-dimensional linguistic profiles into streamlined, interpretable diagnostic dashboards will eventually empower clinicians to conduct precision, task-matched discourse assessments in routine practice.

6.2. Generalizing beyond the tasks studied

These findings and earlier research led to the assumption that the results have consequences that extend beyond the tasks and potentially the populations studied in this work including affecting the acoustic and linguistic measures (Themistocleous, 2016, 2017, 2019). For example, earlier research has examined written and spoken language production in various populations, including individuals with post-stroke aphasia (e.g., writing; (Mortensen, 2005), Traumatic Brain Injury (TBI; e.g., writing; (Wilson & Proctor, 2002), Alzheimer's disease (writing and speech; (Croisile et al., 1996), non-fluent progressive aphasia (writing and speech; Graham et al., 2004; Tippett et al., 2025), and even healthy young adults (writing and speech; Behrns et al., 2009; Croisile et al., 1996; Ulatowska et al., 1983). Overall, these studies demonstrate that the modality of expression—whether written or spoken—acts as another powerful contextual constraint on language output. Across both healthy and clinical groups, written tasks typically elicit denser, more syntactically complex language than spoken tasks, while simultaneously exposing distinctly different patterns of pathological breakdown. This reinforces the fundamental premise of our study: an individual's linguistic profile is inextricably linked to, and modulated by, the specific demands of the elicitation method.

6.3. Limitations and future directions

Although this study's scale and computational depth provide novel insights, several limitations must be acknowledged, which in turn pave the way for future research.

First, the analysis relies on aggregating large, publicly available clinical corpora. A significant strength of this approach is its ecological validity and the large sample size; however, it also introduces limitations. The dataset is inherently unbalanced, as not all participants completed every task. Although our use of mixed-effects models is statistically robust for managing such missing data, the statistical power to detect interaction effects may be lower for specific task-and-diagnosis combinations where data is sparse. Furthermore, combining data from various sources (e.g., AphasiaBank, DementiaBank) means there could be subtle, unmeasured variations in data collection protocols, recording quality, or examiner instructions across the original study sites that contribute to noise in the data. One consequence of these difficulties was that patients with Dementia did not undergo the whole battery of tasks as the other groups and were excluded from the post-hoc analysis. However, their data were critical for the unsupervised machine learning analysis and most importantly provided data for their picture description tasks in the clustering analysis.

Second, the diagnostic categories used in this study (e.g., LHD, TBI, Dementia) are broad and encompass significant internal heterogeneity. For instance, the LHD group includes individuals with varying aphasia subtypes (e.g., Broca's, Wernicke's, anomic), each with a distinct linguistic profile. Our analysis successfully identifies overarching patterns at the group level, but it may mask more granular, subtype-specific interactions between task and diagnosis. Future research could apply this framework to more narrowly defined clinical subgroups to uncover finer-grained distinctions.

Third, our findings are based on data from English-speaking participants. The specific linguistic features that are most sensitive to the task-diagnosis interaction (e.g., reliance on word order, specific inflectional morphemes) are tied to the typological structure of English. The broader principle—that task demands interact with underlying deficits—is likely universal, but the specific “linguistic fingerprints" will almost certainly differ in other languages, particularly those with richer morphological systems or different syntactic rules. Cross-linguistic replication of this study is a critical next step to establishing a more universal model of task effects in clinical assessment.

Fourth, our analysis is contingent on an automated computational pipeline. Although Open Brain AI enables a large-scale analysis that would be impossible to perform manually, no NLP tool is perfect. The accuracy of part-of-speech tagging, dependency parsing, and semantic entity recognition can be lower for atypical, error-prone, or disfluent speech common in neurogenic disorders. These potential inaccuracies could influence the precise values of the linguistic measures, though the large-scale patterns observed are likely robust. Future work should include targeted validation of the platform's performance on diverse clinical speech samples. Looking ahead, these limitations highlight clear future directions. The clinical takeaway that assessment tasks are not interchangeable is paramount. Future research should focus on developing task-specific normative databases, allowing for more precise and valid comparisons. Clinicians can use the framework presented here to make hypothesis-driven choices about assessment, selecting tasks strategically to probe specific linguistic abilities. For example, a picture description is ideal for a quick lexical assessment, whereas a personal narrative is a superior tool for evaluating high-level discourse organization and syntactic complexity. By embracing the interaction between task and diagnosis, we can refine our diagnostic tools, set more targeted therapeutic goals, and ultimately deepen our understanding of the multifaceted nature of neurogenic language disorders.

Fifth, the current automated analysis is focused on microlinguistic features and does not fully capture macrolinguistic and functional aspects of discourse. Even though certain extracted metrics can serve as valuable, pragmatic assessment broader measures of macrostructural organization, global coherence, and overall communicative effectiveness remain challenging to extract reliably at scale. Because certain clinical profiles, such as Traumatic Brain Injury, are typically associated with reduced macrostructural organization and pragmatic appropriateness despite relatively preserved microlinguistic features, examining these higher-level dimensions is essential. Future research must prioritize extending automated computational frameworks to evaluate these functional elements, as the specific demands of a task likely exert profound influences on language performance well beyond the microlinguistic level. Relatedly, the present results are limited to largely monologic data and free-speech data, which was likewise largely monologic with limited interlocutor involvement. Conversational communication introduces additional demands related to turn-taking, topic negotiation, adaptation to a conversational partner, and real-time response generation that are not fully captured in monologic discourse tasks or the data captured here. Accordingly, the present findings should not be assumed to extend directly to dialogic conversation.

Building on these principles, three translational priorities emerge. First, cross-linguistic validation is essential to determine whether the task-diagnosis “fingerprints” identified here generalize to languages with richer morphology or different syntactic structures. Second, linking linguistic profiling with neuroimaging (e.g., lesion-symptom mapping, functional connectivity) will help clarify mechanistic underpinnings of task effects and support biomarker development. Third, there is a pressing need to translate these findings into implementable clinical tools. This includes developing a modernized library of discourse tasks that are co-designed with patients and clinicians, systematically tiered by genre (extending into dialogic) and complexity, and validated for

psychometric rigor. Parallel efforts should create streamlined, clinician-informed scoring frameworks that capture both linguistic and pragmatic features but can be completed efficiently in routine practice.

CRediT authorship contribution statement

Brielle C. Stark: Writing – original draft, Resources, Investigation, Funding acquisition, Data curation, Conceptualization. **Charalambos Themistocleous:** Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Author BCS is a Co—I on R01-DC008524 (PI: MacWhinney), which contributes data to AphasiaBank. Author BCS is a TalkBank Board Member. Author BCS has no other financial or non-financial conflict of interest].

Acknowledgment

This research was suppported by National Institute on Deafness and Other Communication Disorders Grant R01DC008524 awarded to Co-I Brielle C. Stark (PI: MacWhinney).

Appendix A. Appendix 1. Inclusive table of metrics

The measure categories elicited from the picture description tasks are provided below, with a working definition. In addition to the measures listed below, individual phonological and morphological measures that fall under the categories, like the number of syllables (one, two, three etc. syllable words), different syllable types (e.g., CV, VC, V, CCV) and POS morphological inflectional features, such as nominal number, gender and case, which are operationalized using the Universal Dependencies framework (https://universaldependencies.org/u/feat/index.html), are not listed. Also, several measures include both counts and ratios.

Readability Measures

1. *Estimated Reading Time (sec)*: Predicts time to read a text, typically calculated by dividing word count by average reading speed (e.g., 200–300 words/min) and converting to seconds
2. *Flesch Reading Ease*: Scores text 0–100 based on sentence length and syllables per word Higher scores mean easier reading (e.g., 90–100 = ~5th grade)
3. *Flesch-Kincaid Grade Level*: Converts Flesch scores to U S grade levels (e.g., 8 0 = 8th-grade comprehension), using sentence length and syllables
4. *Gunning Fog Index*: Estimates education level needed via sentence length and complex words (≥3 syllables) A score of 12 = 12th-grade level
5. *Coleman-Liau Index*: Uses characters per word and sentences per one hundred words to determine grade level, avoiding syllable counts
6. *Automated Readability Index (ARI)*: Relies on characters per word and sentence length to output grade levels (e.g., 10 = 10th grade)
7. *SMOG Index*: Focuses on polysyllabic words (≥3 syllables) to gauge education level required, often used in healthcare contexts
8. *Linsear Write Formula*: Tailored for technical writing; weights "hard" words (≥3 syllables) and sentence length, adjusting results for grade level
9. *Passive Sentences Percent*: Tracks passive voice usage; lower percentages enhance clarity and readability
10. *Dale-Chall Readability Score*: Uses a predefined "easy word" list and sentence length to assign grade levels (e.g., 9–10 = college level)
11. *Difficult Words*: Measures percentage of complex terms (e.g., multisyllabic or rare words), with higher percentages indicating denser text

Phonological Measures

1. Syllable counts count of one, two, three, etc. syllable words
2. Structure of syllables: a combination of Consonant (C) and Vowel (V)

Morphological Measures

Part of Speech Measures

POS Morphological features are operationalized using the Universal Dependencies framework to classify inflectional categories, such as nominal number, gender and case.

1. *Adjectives:* words that describe or modify nouns or pronouns (e.g., "red," "happy")
2. *Adpositions:* words that indicate direction, location, or time (e.g., "in," "on")
3. *Adverbs:* words that modify verbs, adjectives, or other adverbs (e.g., "quickly," "very")
4. *Auxiliaries:* helping verbs that express tense, aspect, or modality (e.g., "is," "will")
5. *Coordinating conjunctions:* words that connect words, phrases, or clauses of equal importance (e.g., "and" "but")
6. *Determiners*: words that specify or limit a noun's reference (e.g., "the," "this")
7. *Interjections*: exclamatory words expressing emotion (e.g., "wow," "oh")
8. *Nouns*: words that represent people, places, things, or ideas (e.g., "dog," "city")
9. *Numerals*: words that quantify nouns or indicate order (e.g., "one," "first")
10. *Particles*: function words that modify the meaning of a sentence or verb (e.g., "not," "to")
11. Pronouns: words that replace nouns (e.g., "he," "they")
12. Proper nouns: words that name unique entities (e.g., "John," "Paris")
13. Subordinating conjunctions: words that connect dependent clauses to independent clauses (e.g., "because" "although")
14. Symbols: characters representing symbols (e.g., "$," "%")
15. Verbs: words that express actions, occurrences, or states of being (e.g., "run," "is")
16. Other: words that do not fit into any other POS category

Syntactic Measures

Core Syntactic Relations

1. *Clausal modifier of noun (acl)*: A clause that modifies a noun (e.g., *"The book [that I read] was fascinating"*)
2. *Adjectival complement (acomp)*: An adjective functioning as the complement of a verb (e.g., *"She seems [happy]"*)
3. *Adverbial clause modifier (advcl)*: A clause acting as an adverb to modify a verb (e.g., *"He left [because it was late]"*)
4. *Adverbial modifier (advmod)*: A word or phrase modifying a verb, adjective, or adverb (e.g., *"She ran [quickly]"*)
5. *Adjectival modifier (amod)*: An adjective directly modifying a noun (e.g., *"[Red] apple"*)
6. *Appositional modifier (appos)*: A noun phrase providing additional info about another noun (e.g., *"My friend [the doctor] is here"*)
7. *Attribute (attr)*: A nominal predicate linked to the subject via a copula (e.g., *"She is [a teacher]"*)

Verb-Related Roles

1. *Auxiliary: (aux)*: A helping verb (e.g., *"She [has] eaten"*)
2. *Auxiliary: (passive) (aux: pass)*: Auxiliary: verb in a passive construction (e.g., *"The cake [was] baked"*)
3. *Clausal complement (ccomp)*: A clause acting as the direct object of a verb (e.g., *"He said [she left]"*)
4. *Open clausal complement (xcomp)*: A clausal complement missing an overt subject (e.g., *"She wants [to leave]"*)

Subject/Object Roles

1. *Clausal subject (csubj)*: A clause acting as the subject (e.g., *"[That he lied] is obvious"*)
2. *Nominal subject (nsubj)*: A noun phrase acting as the subject (e.g., *"[The cat] sleeps"*)
3. *Nominal subject (passive) (nsubj: pass)*: Subject of a passive verb (e.g., *"[The cake] was eaten"*)
4. *Direct object (dobj)*: The noun phrase directly receiving the action (e.g., *"She ate [the cake]"*)
5. *Object of preposition (pobj)*: Noun governed by a preposition (e.g., *"in [the house]"*)

Modifiers & Connectives

1. *Case: marker (case)*: A preposition/postposition marking grammatical Case: (e.g., *"[in] the box"*)
2. *Coordinating conjunction (cc)*: Connects words/clauses (e.g., *"tea [and] coffee"*)
3. *Compound modifier (compound)*: Multi-word expressions (e.g., *"[coffee] cup"*)
4. *Conjunct (conj)*: Elements linked by coordination (e.g., *"apples [and] oranges"*)
5. *Possession modifier (nmod: poss)*: Marks possession (e.g., *"[John's] book"*)
6. *Prepositional modifier (prep)*: A prepositional phrase modifying a noun/verb (e.g., *"the book [on the table]"*)

Special Constructions

1. *Expletive (expl)*: A placeholder subject with no semantic role (e.g., *"[There] is a problem"*)
2. *Interjection (intj)*: An exclamation (e.g., *"[Wow], that's amazing!"*)
3. *Marker (mark)*: Subordinating conjunction (e.g., *"She said [that] it's true"*)
4. *Negation modifier (neg)*: Negation word (e.g., *"She [did not] go"*)
5. *Parataxis*: Loosely connected clauses (e.g., *"He left, [she stayed]"*)
6. *Relative clause modifier (relcl)*: Clause introduced by a relative pronoun (e.g., *"The man [who called]"*)

Phrasal Categories

1. *Noun Phrases (NP)*: A phrase headed by a noun (e.g., *"[The quick brown fox]"*)
2. *Verb Phrases (VP)*: A phrase headed by a verb (e.g., *"[jumps over the dog]"*)
3. *Prepositional Phrases (PP)*: A phrase headed by a preposition (e.g., *"[in the house]"*)
4. *Adjective Phrases (ADJP)*: A phrase headed by an adjective (e.g., *"[very happy]"*)
5. *Adverbial Phrases (ADVP)*: A phrase headed by an adverb (e.g., *"[quite slowly]"*)

Miscellaneous

1. *Determiner (*det*)*: Articles or quantifiers (e.g., *"[The] cat"*)
2. *Number: modifier (nummod)*: Numerals modifying nouns (e.g., *"[three] books"*)
3. *Particle (prt)*: Verb particle in phrasal verbs (e.g., *"give [up]"*)
4. *Punctuation (punct)*: Marks like commas or periods
5. *Root*: The main predicate of the sentence (e.g., the central verb)
6. *Unclassified dependent (dep)*: A dependency that defies standard classification

Basic Syntax Measures

1. *Total Sentences*: The Number: of sentences in a text
2. *Total Words*: The total word count (tokens) in a text
3. *Average Sentence Length (tokens)*: Mean Number: of words per sentence (Total Words / Total Sentences)
4. *Minimum Sentence Length (tokens)*: Shortest sentence in the text (by word count)
5. *Maximum Sentence Length (tokens)*: Longest sentence in the text (by word count)

Syntactic Tree Metrics

1. *Average Tree Height*: Mean depth of syntactic parse trees (hierarchical structure) across sentences. Higher values indicate more embedded/clauses
2. *Total Tree Height*: Sum of the heights of all syntactic parse trees in the text (if analyzing multiple sentences)
3. *Yngve Depth*: A measure of syntactic complexity based on how deeply phrases are nested in a parse tree. For example, in *"The cat [on the mat [in the room]]"*, the second prepositional phrase increases depth, we provide the following:
   a. *Mean Yngve Depth*: Average Yngve score per sentence, measuring syntactic complexity via right-branching structures (higher = more nested phrases)
   b. *Max Yngve Depth*: Highest Yngve score in the text, indicating the most complex sentence's nesting depth

Clause and Sentence Complexity

1. *Total Clauses*: Count of all independent and dependent clauses (e.g., *"She laughed [when he fell]"*= 2 clauses)
2. *Total T-units*: *T-units* are used to assess writing maturity; longer/more complex T-units suggest advanced syntax. We calculate the Number: of "minimal terminable units"—a main clause plus any attached subordinate clauses. Measures sentence segmentation (e.g., *"He ran [because he was late]"* = 1 T-unit)
3. *Total Complex T-units*: T-units containing at least one subordinate clause (e.g., *"She said [that it was true]"*)
4. Total Dependent Clauses: Clauses that cannot stand alone (e.g., "[After she left], it rained")

Phrase-Level Metrics

1. Total Verb Phrases (VPs): Phrases centered around a verb (e.g., "[is running quickly]")
2. *Total Coordinate Phrases*: Phrases joined by coordinating conjunctions (e.g., *"apples [and] oranges"*)
3. *Total Complex Nominals*: Noun phrases with modifiers (e.g., *"the [quick brown] fox"*) Counts nouns with adjectives, prepositional phrases, or clauses. Frequent use correlates with formal/academic writing

Semantics

Numerical and Quantitative Entities

1. *Cardinal Number:* Numerals (e.g., *"three"*, *"100"*)
2. *Ordinal Number:* Positions in a sequence (e.g., *"first"*, *"third"*)

3. *Percentage*: Values with a percent symbol (e.g., *"75%"*)
4. *Monetary value*: Currency amounts (e.g., *"$50"*, *"€200"*)
5. *Quantity*: Measurements with units (e.g., *"5 kg"*, *"10 l"*)

Time-Related Entities

1. *Date*: Absolute or relative dates (e.g., *"January 1, 2023"*, *"next Monday"*)
2. *Time*: Time of day or duration (e.g., *"3:00 PM"*, *"two hours"*)

Geographical and Political Entities

1. *Location*: Physical places (natural or human-made) (e.g., *"Mount Everest"*, *"Pacific Ocean"*)
2. *Geopolitical entity (GPE)*: Countries, cities, states, or regions (e.g., *"France"*, *"New York"*)
3. *Facility*: Man-made structures (e.g., *"Kennedy Airport"*, *"Eiffel Tower"*)

Organizations and Groups

1. *Organization*: Companies, institutions, or groups (e.g., *"Google"*, *"United Nations"*)
2. *Nationalities/Religious/Political groups*: Ethnicities, faiths, or parties (e.g., *"Christians"*, *"Democrats"*)

People and Cultural Entities

1. *Person*: Individual names (e.g., *"Marie Curie"*, *"Elon Musk"*)
2. *Language*: Named languages (e.g., *"Mandarin"*, *"Swahili"*)
3. *Work of art*: Creative works (e.g., *"Mona Lisa"*, *"Hamlet"*)

Events and Abstract Concepts

1. *Event*: Historical, cultural, or planned occurrences (e.g., *"World War II"*, *"Olympic Games"*)
2. *Law*: Legal documents, statutes, or treaties (e.g., *"Civil Rights Act"*, *"GDPR"*)

Commercial Entities

Product: Branded goods or services (e.g., *"iPhone"*, *"Tesla Model S"*).

Data availability

https://osf.io/vud95/overview

TalkBank Clinical Banks (Reference data) (Open Science Framework)

References

Aamodt, W. W., Waligorska, T., Shen, J., Tropea, T. F., Siderowf, A., Weintraub, D., … Chen-Plotkin, A. S. (2021). Neurofilament light chain as a biomarker for cognitive decline in Parkinson disease. *Movement Disorders, 36*(12), 2945–2950. https://doi.org/10.1002/mds.28779

Agmon, G., Pradhan, S., Ash, S., Nevler, N., Liberman, M., Grossman, M., & Cho, S. (2024). Automated measures of syntactic complexity in natural speech production: Older and younger adults as a case study. *Journal of Speech, Language, & Hearing Research, 67*(2), 545–561. https://doi.org/10.1044/2023_JSLHR-23-00009

Antonucci, S. M. (2009). Use of semantic feature analysis in group aphasia treatment. *Aphasiology, 23*(7–8), 854–866. https://doi.org/10.1080/02687030802634405

Ash, S., et al. (2013). Differentiating primary progressive aphasias in a brief sample of connected speech. *Neurology, 81*(4), 329–336. https://doi.org/10.1212/WNL.0b013e31829c5d0e

Badecker, W., Hillis, A., & Caramazza, A. (1990). Lexical morphology and its role in the writing process: Evidence from a case of acquired dysgraphia. *Cognition, 35*(3), 205–243.

Ball, M. J., Müller, N., & Rutter, B. (2014). *Phonology for communication disorders*. Psychology Press.

Barbieri, Z., Fernández, M. A., Newbury, D. F., & Villanueva, P. (2018). Family aggregation of language impairment in an isolated Chilean population from Robinson Crusoe Island. *International Journal of Language & Communication Disorders, 53*(3), 643–655. https://doi.org/10.1111/1460-6984.12377

Bastiaanse, R. (2013). Why reference to the past is difficult for agrammatic speakers [conference paper]. *Clinical Linguistics & Phonetics, 27*(4), 244–263. https://doi.org/10.3109/02699206.2012.751626

Behrns, I., Wengelin, Å., Broberg, M., & Hartelius, L. (2009). A comparison between written and spoken narratives in aphasia. *Clinical Linguistics & Phonetics, 23*(7), 507–528.

Berndt, R. S., et al. (1996). Comprehension of reversible sentences in "agrammatism": A meta-analysis. *Cognition, 58*(3), 289–308. https://doi.org/10.1016/0010-0277(95)00682-6

Best, W., et al. (2002). Phonological and orthographic facilitation of word-retrieval in aphasia: Immediate and delayed effects. *Aphasiology, 16*(1–2), 151–168. https://doi.org/10.1080/02687040143000483

Bose, A., et al. (2021). Connected speech characteristics of Bengali speakers with Alzheimer's disease: Evidence for language-specific diagnostic markers. *Frontiers in Aging Neuroscience, 13*, Article 707628. https://doi.org/10.3389/fnagi.2021.707628

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: What's in a name? *Topics in Stroke Rehabilitation, 17*(6), 411–422. https://doi.org/10.1310/tsr1706-411

Bryant, L., Ferguson, A., & Spencer, E. (2016). Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical Linguistics & Phonetics, 30*(7), 489–518. https://doi.org/10.3109/02699206.2016.1145740

Bryant, L., Spencer, E., & Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology, 31*(10), 1105–1126. https://doi.org/10.1080/02687038.2016.1239013

Callegari, E., et al. (2024). *Automatic extraction of language-specific biomarkers of healthy aging in Icelandic*.

Caramazza, A., & Hillis, A. E. (1989). The disruption of sentence production: Some dissociations. *Brain and Language, 36*(4), 625–650.

Caramazza, A., & Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature, 349*(6312), 788–790.

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language, 3*(4), 572–582. https://doi.org/10.1016/0093-934X(76)90048-1

Charniak, E., & Johnson, M. (2005). *Coarse-to-fine n-best parsing and MaxEnt discriminative reranking*.

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied PsychoLinguistics, 13*(1), 53–76. https://doi.org/10.1017/S0142716400005427

Coelho, C. A. (2007). Management of discourse deficits following traumatic brain injury: Progress, caveats, and needs. *Seminars in Speech and Language, 28*(2), 122–135.

Cousins, K. A. Q., Phillips, J. S., Irwin, D. J., Lee, E. B., Wolk, D. A., Shaw, L. M., … Grossman, M. (2021). ATN incorporating cerebrospinal fluid neurofilament light chain detects frontotemporal lobar degeneration. *Alzheimer's & Dementia, 17*(5), 822–830. https://doi.org/10.1002/alz.12233

Croisile, B., Ska, B., Brabant, M.-J., Duchene, A., Lepage, Y., Aimard, G., & Trillet, M. (1996). Comparative study of oral and written picture description in patients with Alzheimer's disease. *Brain and Language, 53*(1), 1–19. https://www.sciencedirect.com/science/article/abs/pii/S0093934X96900334?via%3Dihub.

Croot, K., Hodges, J. R., Xuereb, J., & Patterson, K. (2000). Phonological and articulatory impairment in Alzheimer's disease: A case series. *Brain and Language, 75*(2), 277–309. https://doi.org/10.1006/brln.2000.2357

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M., & Dipper, L. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *International Journal of Language & Communication Disorders, 55*(3), 417–442.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, 37–54.

Delaby, C., Bousiges, O., Bouvier, D., Fillée, C., Fourier, A., Mondésert, E., Nezry, N., Omar, S., Quadrio, I., Rucheton, B., Schraen-Maschke, S., van Pesch, V., Vicca, S., Lehmann, S., & Bedel, A. (2022). Neurofilaments contribution in clinic: State of the art. *Frontiers in Aging Neuroscience, 14*, Article 1265. https://doi.org/10.3389/fnagi.2022.1034684

Devlin, J., et al. (2018). Pre-training of deep bidirectional transformers for language understanding. *arXiv* Doi:10.48550/arXiv.1810.04805.

Efstratiadou, E. A., Papathanasiou, I., Holland, R., Archonti, A., & Hilari, K. (2018). A systematic review of semantic feature analysis therapy studies for aphasia. *Journal of Speech, Language, and Hearing Research, 61*(5), 1261–1278. https://doi.org/10.1044/2018_JSLHR-L-16-0330

Elbourn, E., Kenny, B., Power, E., & Togher, L. (2019). Psychosocial outcomes of severe traumatic brain injury in relation to discourse recovery: A longitudinal study up to 1 year post-injury. *American Journal of Speech-Language Pathology, 28*(4), 1463–1478. https://doi.org/10.1044/2019_AJSLP-18-0204

Fergadiotis, G., & Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology, 25*(11), 1414–1430. https://doi.org/10.1080/02687038.2011.603898

Fitzsimmons, P. R., Michael, B. D., Hulley, J. L., & Scott, G. O. (2010). A readability assessment of online Parkinson's disease information. *The Journal of the Royal College of Physicians of Edinburgh, 40*(4), 292–296. https://doi.org/10.4997/JRCPE.2010.401

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage*. Houghton-Mifflin Company.

Fraser, K. C., Ben-David, N., Hirst, G., Graham, N., & Rochon, E. (2015). Sentence segmentation of aphasic speech. In *Proceedings of the 2015 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 862–871).

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. https://doi.org/10.1016/j.cortex.2012.12.006

Fridriksson, J., den Ouden, D. B., Hillis, A. E., Hickok, G., Rorden, C., Basilakos, A., & Bonilha, L. (2018). Anatomy of aphasia revisited. *Brain, 141*(3), 848–862. https://doi.org/10.1093/brain/awx363

Friedmann, N. (2002). Question production in agrammatism: The tree pruning hypothesis. *Brain and Language, 80*(2), 160–187. https://doi.org/10.1006/brln.2001.2587

Ghojogh, B., Crowley, M., Karray, F., & Ghodsi, A. (2023). Uniform manifold approximation and projection (UMAP). In B. Ghojogh, M. Crowley, F. Karray, & A. Ghodsi (Eds.), *Elements of dimensionality reduction and manifold learning* (pp. 479–497). Springer International Publishing. https://doi.org/10.1007/978-3-031-10602-6_17.

Giannini, L. A. A., et al. (2017). Clinical marker for Alzheimer disease pathology in logopenic primary progressive aphasia. *Neurology, 88*(24), 2276–2284. https://doi.org/10.1212/wnl.0000000000004034

Goldstein, A., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience, 25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Graham, N. L., Patterson, K., & Hodges, J. R. (2004). When more yields less: Speaking and writing deficits in nonfluent progressive aphasia. *Neurocase, 10*(2), 141–155. https://doi.org/10.1080/13554790490497256

Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction*. Taylor & Francis.

Hartig, F. (2016). *DHARMa: Residual diagnostics for hierarchical (multi-level/mixed) regression models*. CRAN: Contributed Packages.

Hashimoto, N., & Frome, A. (2011). The use of a modified semantic features analysis approach in aphasia. *Journal of Communication Disorders, 44*(4), 459–469. https://doi.org/10.1016/j.jcomdis.2011.02.004

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

Hillis, A. E. (1989). Efficacy and generalization of treatment for aphasic naming errors. *Archives of Physical Medicine and Rehabilitation, 70*(8), 632–636.

Hillis, A. E., Beh, Y. Y., Sebastian, R., Breining, B., Tippett, D. C., Wright, A., … Basilakos, A. (2018). Predicting recovery in acute post-stroke aphasia. *Annals of Neurology, 83*(3), 612–622.

Kemper, S. (1987). Life-span changes in syntactic complexity. *Journal of Gerontology, 42*(3), 323–328. https://doi.org/10.1093/geronj/42.3.323

Kim, H., Berube, S., & Hillis, A. E. (2023). Core lexicon in aphasia: A longitudinal study. *Aphasiology, 37*(10), 1679–1691. https://doi.org/10.1080/02687038.2022.2121598

Kim, H., Hillis, A. E., & Themistocleous, C. (2024). Machine learning classification of patients with amnestic mild cognitive impairment and non-amnestic mild cognitive impairment from written picture description tasks. *Brain Sciences, 14*(7).

Kiran, S., & Thompson, C. K. (2003). The role of semantic complexity in treatment of naming deficits. *Journal of Speech, Language, and Hearing Research, 46*(4), 773–787. https://doi.org/10.1044/1092-4388(2003/061)

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly, 10*(1), 62–102. https://doi.org/10.2307/747086

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Lan, G., Zhang, Q., Lucas, K., Sun, Y., & Gao, J. (2022). A corpus-based investigation on noun phrase complexity in L1 and L2 English writing. *English for Specific Purposes, 67*, 4–17. https://doi.org/10.1016/j.esp.2022.02.002

Lanzi, A., .M., Saylor, A., .K., Fromm, D., Liu, H., MacWhinney, B., & Cohen, M., .L. (2023). DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology, 32*(2), 426–438. https://doi.org/10.1044/2022_AJSLP-22-00281

Leyton, C., Ballard, K., Piguet, O., & Hodges, J. (2014). Phonological errors as a clinical marker of the logopenic variant of PPA. *Neurology, 82*(18), 1620–1627.

Mack, J. E., Barbieri, E., Weintraub, S., Mesulam, M. M., & Thompson, C. K. (2021). Quantifying grammatical impairments in primary progressive aphasia: Structured language tests and narrative language production. *Neuropsychologia, 151*, Article 107713. https://doi.org/10.1016/j.neuropsychologia.2020.107713

MacWhinney, B. (2025). Understanding language through TalkBank. *Current Directions in Psychological Science, 34*(2), 75–81. https://doi.org/10.1177/09637214241304345

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

Meteyard, L., & Bose, A. (2018). What does a cue do? Comparing phonological and semantic cues for picture naming in aphasia. *Journal of Speech, Language, and Hearing Research, 61*(3), 658–674. https://doi.org/10.1044/2017_JSLHR-L-17-0214

Minga, J., Stockbridge, M., .D., Durfee, A., & Johnson, M. (2022). Clinical guidelines for eliciting discourse using the RHDBank protocol. *American Journal of Speech-Language Pathology, 31*(5), 1949–1962. https://doi.org/10.1044/2022_AJSLP-22-00097

Mortensen, L. (2005). Written discourse and acquired brain impairment: Evaluation of structural and semantic features of personal letters from a systemic functional linguistic perspective. *Clinical Linguistics & Phonetics, 19*(3), 227–247. https://doi.org/10.1080/02699200410001698652

Mueller, K. D., Hermann, B., Mecollari, J., & Turkstra, L. S. (2018). Connected speech and language in mild cognitive impairment and Alzheimer's disease: A review of picture description tasks. *Journal of Clinical and Experimental Neuropsychology, 40*(9), 917–939. https://doi.org/10.1080/13803395.2018.1446513

Pritchard, M., Hilari, K., Cocks, N., & Dipper, L. (2017). Reviewing the quality of discourse information measures in aphasia. *International Journal of Language & Communication Disorders, 52*(6), 689–732. https://doi.org/10.1111/1460-6984.12318

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, J. D., Dalton, S. G., Greenslade, K. J., Jacks, A., Haley, K. L., & Adams, J. (2021). Main concept, sequencing, and story grammar analyses of Cinderella narratives in a large sample of persons with aphasia. *Brain Sciences, 11*(1), Article 110.

Russell, L. (2020). *Emmeans: Estimated marginal means, aka least-squares means*. R Foundation for Statistical Computing. https://CRAN.R-project.org/package=emmeans.

Senthilkumar, H., Kelly, L., Neal, V., Diaz-Carr, I., Colavito, M., Hillis Argye, E., & Stockbridge Melissa, D. (2026). Revisiting Western aphasia battery–revised in the context of mild impairment. *American Journal of Speech-Language Pathology*. https://doi.org/10.1044/2026_AJSLP-25-00361

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. https://doi.org/10.1044/2019_AJSLP-18-0265

Stark, B. C., Alexander, J. M., Hittson, A., Doub, A., Igleheart, M., Streander, T., & Jewell, E. (2023). Test–retest reliability of microlinguistic information derived from spoken discourse in persons with chronic aphasia. *Journal of Speech, Language, and Hearing Research, 66*(7), 2316–2345.

Stark, B. C., Bryant, L., Themistocleous, C., den Ouden, D.-B., & Roberts, A. C. (2023). Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology, 37*(5), 761–784. https://doi.org/10.1080/02687038.2022.2039372

Stark, B. C., Dalton, S. G., & Lanzi, A. M. (2025). Access to context-specific lexical-semantic information during discourse tasks differentiates speakers with latent aphasia, mild cognitive impairment, and cognitively healthy adults [original research]. *Frontiers in Human Neuroscience, 18*. https://doi.org/10.3389/fnhum.2024.1500735

Stark, B. C., Dutta, M., Murray Laura, L., Bryant, L., Fromm, D., MacWhinney, B., … Sharma, S. (2020). Standardizing assessment of spoken discourse in aphasia: A working group with deliverables. *American Journal of Speech-Language Pathology*, 1–12. https://doi.org/10.1044/2020_AJSLP-19-00093

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. https://doi.org/10.1080/23273798.2020.1862258

Steel, J., Ferguson, A., Spencer, E., & Togher, L. (2017). Language and cognitive communication disorder during post-traumatic amnesia: Profiles of recovery after TBI from three cases. *Brain Injury, 31*(13–14), 1889–1902. https://doi.org/10.1080/02699052.2017.1373200

Stockbridge, M. D., Matchin, W., Walker, A., Breining, B., Fridriksson, J., Hickok, G., & Hillis, A. E. (2021). One cat, two cats, red cat, blue cats: Eliciting morphemes from individuals with primary progressive aphasia. *Aphasiology, 35*(12), 1–12. https://doi.org/10.1080/02687038.2020.1852167

Themistocleous, C. (2016). The bursts of stops can convey dialectal information. *The Journal of the Acoustical Society of America, 140*(4), EL334–EL339. https://doi.org/10.1121/1.4964818

Themistocleous, C. (2017). The nature of phonetic gradience across a dialect continuum: Evidence from modern Greek vowels [journal article]. Phonetica, 74(3), 157–172. doi:https://doi.org/10.1159/000450554)https://doi.org/10.1159/000450554).

Themistocleous, C. (2019). Dialect classification from a single Sonorant sound using deep neural networks. *Frontiers in Communication, 4*, 1–12. https://doi.org/10.3389/fcomm.2019.00064

Themistocleous, C. (2024). Open brain AI and language assessment. *Frontiers in Human Neuroscience, 18*. https://doi.org/10.3389/fnhum.2024.1421435

Themistocleous, C., Eckerström, M., & Kokkinakis, D. (2018). Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks [https://doi.org/10.3389/fneur.2018.00975]. Frontiers in Neurology, 9, 975. doi: 10.3389/fneur.2018.00975.

Themistocleous, C., Eckerström, M., & Kokkinakis, D. (2020). Voice quality and speech fluency distinguish individuals with mild cognitive impairment from healthy controls. *PLoS One, 15*(7), Article e0236009. https://doi.org/10.1371/journal.pone.0236009

Themistocleous, C., Ficek, B., Webster, K., den Ouden, D.-B., Hillis, A. E., & Tsapkini, K. (2021). Automatic subtyping of individuals with primary progressive aphasia. *Journal of Alzheimer's Disease, 79*, 1185–1194. https://doi.org/10.3233/JAD-201101

Themistocleous, C., & Stark, B. C. (2026). *ALBA: An automated framework for benchmarking clinical language biomarkers against standardized corpora*. LREC 2026,. Palma de Mallorca.

Themistocleous, C., Webster, K., Afthinos, A., & Tsapkini, K. (2020). Part of speech production in patients with primary progressive aphasia: An analysis based on natural language processing. *American Journal of Speech-Language Pathology, 1-15*. https://doi.org/10.1044/2020_AJSLP-19-00114

Thompson, C. K., & Mack, J. E. (2014). Grammatical impairments in PPA. *Aphasiology, 28*(8–9), 1018–1037. https://doi.org/10.1080/02687038.2014.912744

Tippett, D. C., Surrao, K., Neophytou, K., Kim, H., Gallegos, J., Themistocleous, C., … Tsapkini, K. (2025). Written picture descriptions distinguish variants of primary progressive aphasia. *Journal of Alzheimer's Disease, 108*(1), 334–352. https://doi.org/10.1177/13872877251376381

Ulatowska, H. K., Doyel, A. W., Stern, R. F., Haynes, S. M., & North, A. J. (1983). Production of procedural discourse in aphasia. *Brain and Language, 18*(2), 315–341. https://www.sciencedirect.com/science/article/abs/pii/0093934X83900238?via%3Dihub.

Ulatowska, H. K., North, A. J., & Macaluso-Haynes, S. (1981). Production of narrative and procedural discourse in aphasia. *Brain and Language, 13*(2), 345–371. https://www.sciencedirect.com/science/article/abs/pii/0093934X81901000?via%3Dihub.

Venna, J., & Kaski, S. (2001). Neighborhood preservation in nonlinear projection methods: An experimental study. In *Artificial neural networks — ICANN 2001, Berlin, Heidelberg*.

Wilson, B. M., & Proctor, A. (2002). Written discourse of adolescents with closed head injury. *Brain Injury, 16*(11), 1011–1024. https://doi.org/10.1080/02699050210147248

Wilson, S. M., DeMarco, A. T., Henry, M. L., Gesierich, B., Babiak, M., Miller, B. L., & Gorno-Tempini, M. L. (2016). Variable disruption of a syntactic processing network in primary progressive aphasia. *Brain, 139*(11), 2994–3006. https://doi.org/10.1093/brain/aww218

Wilson, S. M., Dronkers, N. F., Ogar, J. M., Jang, J., Growdon, M. E., Agosta, F., & Gorno-Tempini, M. L. (2010). Neural correlates of syntactic processing in the nonfluent variant of primary progressive aphasia. *Journal of Neuroscience, 30*(null), 16845.

Wright, H. H., & Capilouto, G. J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology, 23*(10), 1295–1308. https://doi.org/10.1080/02687030902826844