ALBA: An Automated Framework for Benchmarking Clinical Language Biomarkers against Standardized Corpora

Charalambos Themistocleous, Brielle C. Stark
[1]University of Oslo, 0371 Oslo, Norway
[2]Indiana University Bloomington, 2631 E Discovery Parkway, Bloomington, IN 47408, USA
[1]charalth@uio.no, [2]bcstark@iu.edu

Abstract

Patients with diverse neurocognitive conditions frequently exhibit measurable language deficits that serve as biomarkers for differential diagnosis and therapy decision making. Discourse analysis can offer reliable ecological measures of human communication, yet manual discourse analysis is cumbersome. Recent advances in automated analysis software provide quick and easy extraction of raw language metrics in the clinic. Nevertheless, transforming these measures into actionable clinical insights remains a significant challenge. The aim of this paper is to present the Automated Language Biomarker Application (ALBA), an integrated framework developed within the Open Brain AI ecosystem to bridge the gap between feature extraction and clinical interpretation. ALBA provides clinicians with a robust statistical infrastructure to benchmark individual patient measures against standardized, large-scale clinical corpora. By utilizing a shared elicitation and processing pipeline, the application ensures that user-provided data are directly comparable to population norms for conditions including Aphasia, Mild Cognitive Impairment (MCI), Dementia, and other neurological conditions. The system implements adaptive statistical logic, employing one-sample t-tests and robust non-parametric alternatives to provide real-time significance testing and dynamic visualizations (box, bar, and violin plots). By automating the comparison of "Language Signatures" against healthy controls and specific clinical phenotypes, ALBA facilitates rapid, evidence-based decision-making in both research and rehabilitation contexts.

Keywords: Language Biomarkers, Open Brain AI, TalkBank, Clinical Discourse Analysis

1. Introduction

Clinical discourse analysis has a long history in assessment of neurogenic language, especially aphasia, because language-in-use, such as during picture sequence description or narrative retells, often provides a more nuanced and complex picture of a person's linguistic ability than isolated tests of language, such as confrontation picture naming. While clinical discourse analysis evaluates linguistic, propositional, macrostructural, and pragmatic aspects of discourse, many neurogenic populations struggle with foundational linguistic components. As a result, language biomarkers have tended to be derived from quantitative language measures linked to specific clinical or developmental conditions. Further, in practice, clinicians often extract such measures, including part-of-speech counts, lexical ratios, and word-frequency distributions, to systematically assess and objectively score an individual's communicative efficacy (Bryant et al., 2016).

However, interpreting these findings remains challenging in the absence of reference corpora or standardized norms to facilitate comparative analysis. To bridge this gap, three primary objectives must be met: the establishment of gold-standard metrics, the adoption of shared elicitation methodologies, and the provision of accessible tools for the rapid comparison of results to support informed clinical decision-making.

Although numerous research studies have reported various linguistic measures (Themistocleous and Stark, 2026; Varkanitsa et al., 2023; Kiran et al., 2019), the field currently lacks a unified framework that enables clinicians to generate patient-specific data and compare it with standardized measures derived through identical methodologies. As the need for clinical discourse analysis and its importance in eliciting rich and ecological measures of human communication compared to other clinical tasks becomes more pressing, providing such a unified framework is an exceedingly important need.

This paper presents Automated Language Biomarker Application (ALBA), a web-based tool for producing and comparing clinical research outcomes with standardized measures. ALBA has been developed to serve as a data resource by facilitating the presentation and comparison of automatic measures generated by Open Brain AI, a clinical research platform and computational tool designed for language assessment and analysis (Themistocleous, 2024). These measures originate from both clinical tasks (e.g., cookie theft, Cinderella story, story-telling, and story-retelling) as well as from large-scale text corpora. A key advantage of ALBA is its flexibility: new language measures can be easily added or updated to accommodate emerging clinical conditions and evolving research needs.

Proceedings of the 6th RaPID@MENTAI.ai Workshop @ LREC 2026, pages 34–40
12 May 2026.

2. Previous Research

Developmental or acquired neurocognitive conditions can disrupt language and communication in complex and heterogeneous ways, with impairments often spanning expressive domains such as phonology, morphology, syntax, semantics, and lexical access (Themistocleous and Stark, 2026; Varkanitsa et al., 2023; Kiran et al., 2019). While some aspects of linguistic breakdown can be partially predicted by lesion location or severity, the relationship is far from deterministic due to the distributed and dynamic nature of language networks in the brain. Discourse-level language analysis offers a unique window into these impairments, capturing subtle disruptions in coherence, cohesion, informativeness, and pragmatic appropriateness that are often missed by more constrained or modular assessments.

Through systematic quantification of expressive language—particularly at the discourse level—automated language measures can function as robust biomarkers, helping to characterize, differentiate, and subtype various neurological conditions, including left and right hemisphere stroke, traumatic brain injury, mild cognitive impairment, and dementia. These measures not only distinguish clinical populations from healthy controls but also provide insight into the underlying cognitive and communicative mechanisms affected in each condition (Themistocleous and Stark, 2026).

To enable differential diagnosis and inform treatment planning, language measures must be interpreted against well-characterized reference populations or normative data, which provide essential context for distinguishing clinical profiles. Normative reference data are especially valuable in identifying subtle but clinically meaningful language changes in individuals with the mildest forms of aphasia (e.g., latent aphasia), traumatic brain injury, right hemisphere disorder, mild cognitive impairment, or the earliest stages of dementia—conditions where language impairments may not be immediately obvious but are crucial to differentiate from typical aging. While normative comparisons may be less critical for diagnosing more overt or "frank" aphasias, they still provide meaningful context for characterizing the specific pattern and severity of impairment, guiding tailored treatment approaches, and establishing a baseline for tracking individual progress.

For clinicians, access to normative data can enhance decision-making in several key ways. (1) To characterize the specific pattern and severity of impairment, detailed language measures—such as lexical diversity, syntactic complexity, or informativeness—can help phenotype aphasia presentations beyond broad classifications like Broca's or Wernicke's aphasia. For instance, two individuals with anomic aphasia may show similar naming deficits, but one may produce overly vague narratives with limited cohesion, while another struggles with syntactic formulation—distinctions that are only visible through discourse-level profiling against normative benchmarks. (2) In guiding tailored treatment approaches, knowing which language domains are disproportionately affected relative to healthy controls can help clinicians prioritize intervention targets. For example, if a person with right hemisphere damage demonstrates relatively preserved syntax but poor global coherence, therapy can focus more on narrative structuring and pragmatic use. (3) To establish a baseline, quantified language data at intake provide a reference point for monitoring individual progress over time, allowing clinicians to track meaningful change in discourse abilities, even if those changes do not shift the person's broad diagnostic category.

2.1. Traditional Discourse Analysis: Time Consuming, Resource Intensive, Lacking in Tools

Historically, discourse analysis in clinical and educational settings has relied on manual annotation, a process requiring granular characterization of *micro-structural* features (linguistic and propositional levels, e.g., lexical and sentence level features) and *macro-structural* properties (planning and pragmatic levels, e.g., global coherence/cohesion, thematic evolution, and topic maintenance) (Paltridge, 2006). Although providing deep qualitative insights, manual analysis is resource-intensive, requiring specialized linguistic expertise and extensive labor for scoring (Hansen et al., 2022; Cruice et al., 2020; Bryant et al., 2018), and often clinicians and researchers cite these as significant barriers preventing them from engaging in discourse analysis. Consequently, traditional diagnostic batteries have often been restricted to narrow elicitations of measures from connected speech productions and a substantial lack of standardization outputs (Stark et al., 2023).

2.2. Computational Tools

To address the limitations of manual scoring, several software frameworks have emerged to automate linguistic feature extraction. The *Computerized Language Analysis (CLAN)* system, part of the TalkBank project, established the industry standard by using the CHAT transcription format to calculate morphosyntactic and lexical diversity measures (MacWhinney). Similarly, the *Systematic Analysis of Language Transcripts (SALT)* has become a clinical staple for speech-language pathologists (SLPs), focusing on standardized metrics like Mean Length

of Utterance (MLU) and error coding (Cunningham and Haley, 2020; Fergadiotis, 2011).

The emergence of neural Natural Language Processing along with end-to-end automated pipelines for text processing are gradually finding their way into the clinic (Tippett et al., 2025). Platforms such as *Open Brain AI (OBAI)* (Themistocleous, 2024) and automated tools like *Batchalign* (Liu et al., 2023) have moved beyond keyword counting to leverage Automatic Speech Recognition (ASR), Neural Morphosyntactic Tagging, and Transdiagnostic Biomarkers. For example, Open Brain AI can now extract hundreds of linguistic biomarkers—ranging from acoustics to semantic coherence–allowing for a "Language Biomarker" that characterizes specific neurodegenerative or developmental conditions. Such automated computational language measures are being to describe the various language domains including and contribute to the automatic patient identificatio, subtyping, and prognosis (Fraser et al., 2014; König et al., 2018; Themistocleous et al., 2021).

Modern automated AI tools typically generate two distinct classes of linguistic measures, each serving a unique function in clinical research. The first class consists of interpretable biomarkers with direct physical or clinical correlates. These measures are associated with specific pathologies; for instance, deficits in function word ratios can indicate agrammatism, while a reduced noun-to-verb ratio may signal anomia (Themistocleous et al., 2020). Because of their overt clinical interpretation, these metrics are easily integrated into diagnostic assessments and used to define specific therapeutic targets.

The second class comprises latent representations, such as high-dimensional word and sentence embeddings (Bengio and Heigold; Mikolov et al., 2013). While these measures—often derived from large-scale transformer models—lack immediate transparency for clinicians, they serve as highly robust predictors in supervised and unsupervised classification tasks. Both types of measures are essential: while interpretable features provide the "why" for clinical intervention, latent embeddings often provide superior accuracy for automated screening and condition subtyping.

2.3. From Extraction to Interpretation

Despite the proliferation of tools designed to *extract* linguistic data from discourse, a critical gap remains in the *interpretation* of the interpretable language biomarkers within a clinical time-frame. Although existing software provide raw counts or ratios, the burden of statistical comparison often falls on the clinician, who must manually reference published norms or reference datasets. Our application, ALBA (`https://openbrainai.com/measures`), is integrated directly within the *Open Brain AI* ecosystem and addresses this limitation by providing a robust statistical infrastructure that bridges the gap between computational measures and clinical decision-making.

First, it utilizes large-scale normative measures for mitigating the impact of individual linguistic idiosyncrasies (Stark and Fukuyama, 2021). Historically, the use of standardized corpora has been foundational to clinical and psychological research. Early efforts relied on general-purpose datasets, such as the *Brown Corpus* in English and the Språkbanken Text to establish baseline word frequencies and lexical expectations (Francis and Kucera, 1982; Forsberg et al., 2025). Large-scale collections are critical because they define the boundaries of "normal" linguistic variation, accounting for the vast diversity in human speech influenced by age, education, and cognitive health. Without large, validated datasets, researchers and clinicians are unable to determine if a patient's performance represents a pathological deficit or simply a point within the tail of typical variation. Language measures from large-scale normative datasets are thus essential for mitigating the impact of individual linguistic idiosyncrasies serving as benchmarking data for the scores elicited in the clinic.

ALBA aggregates measures derived from standardized corpora, offering an interactive environment where practitioners can empirically determine the degree of divergence between their specific patient data and established population norms. Specifically, by benchmarking against HCs clinicians can quantitatively assess whether a patient's linguistic profile falls within a normative range. Furthermore, it allows for differential comparisons across a spectrum of neurological conditions, including LHD, RHD, TBI, MCI, and dementia.

Secondly, ALBA's integration with Open Brain AI ensures methodological consistency by enabling a direct comparison between clinician-elicited measures and standardized benchmarks, both of which are processed through the same computational pipeline. This homogeneity eliminates cross-platform variance and enhances interpretative accuracy substantially.

3. The ALBA User Interface

The interface of the *(ALBA)* application shown in Figure 1 is designed for clinical intuition and research rigor. The architecture is divided into two primary functional zones:

The *Configuration Sidebar (Left Panel)* facilitates the parameterization of the analysis. Users can select the specific elicitation *task* (e.g., the Cinderella story-retell), the *category* of linguistic measures (lexical, phonological, morphological, syntactic, or

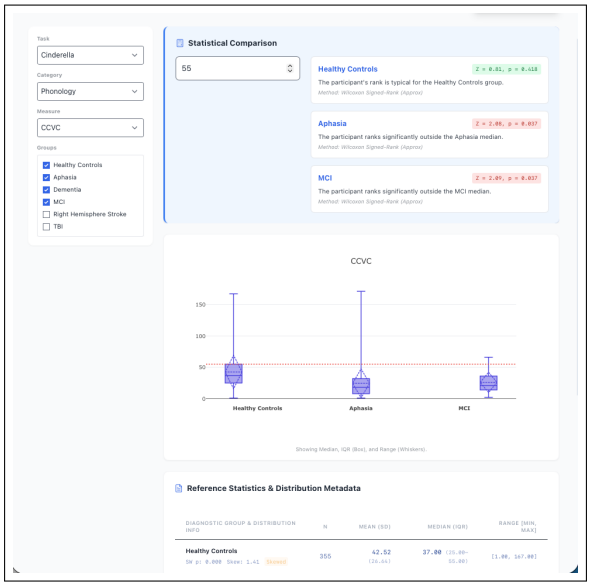

Figure 1: Interface of the Linguistic Biomarkers Application (ALBA). The left hand area allows the selection of the corpus, the category of linguistic measures (e.g., lexical, phonological, morphological, syntactic, semantic) and the language group (healthy controls, patients with different conditions). The main area includes the statistical interface that provides a score with a p-value on whether the measure provided by clinicians is the same or different from the one in the visible corpora based on statistical analysis (see Statistical Methodology). Users can selected different visualization options and view descriptive statistics as a table.

semantic), and the target *clinical cohorts* for comparison (e.g., Healthy Controls vs. various Aphasia subtypes).

The *Analytical Dashboard (Main Panel)* constitutes the central workspace features, an interactive statistical comparison tool. Upon entering a patient's score, the system dynamically calculates and reports a p-value and a formatted APA-style narrative indicating whether the individual's performance significantly diverges from the selected reference corpora (see *Statistical Methodology*).

To support diverse interpretative needs, the main panel offers multiple visualization modalities, including box plots, bar charts with standard deviation error bars, and density-style violin plots. Below the visual representations, a comprehensive Reference Statistics Table provides metadata—including sample sizes, medians, Interquartile Ranges (IQR), and pre-computed normality indicators—ensuring full transparency of the underlying normative data.

4. Methodology

4.1. Data

A flattened relational schema designed for rapid client-side parsing generated from an earlier study (Themistocleous and Stark, 2026) using Open Brain AI from TalkBank data (MacWhinney) is utilized as the primary dataset.

4.2. Statistical Methodology

To maintain high system performance and low latency, ALBA utilizes a hybrid decision engine for statistical test selection. Rather than performing resource-intensive computations on raw transcripts ($N > 600$) at the client level, the system utilizes precomputed distributional metadata. These data correspond to a unique combination of *Task*, *Diagnosis*, *Linguistic Variable* with eleven foundational metrics. *Central Tendency and Dispersion* (i.e., Mean, Standard Deviation (SD), and Median), *Distribution Geometry* (i.e., Min, Max, and the Interquartile Range (Q_1, Q_3)), *Inferential Metadata* (Precomputed Shapiro-Wilk p-values ($P\_Shapiro$) and Fisher-Pearson skewness coefficients ($Skewness$), which inform the automated `IsNormal` flag).

4.3. Test Selection Criteria

The application evaluates each linguistic variable against two primary assumptions:

1. *Parametric Path:* If the reference distribution is pre-validated as normal via the Shapiro-Wilk test ($p > .05$) or meets the Central Limit Theorem criteria ($n \geq 30$ with low skewness), a *one-sample t-test* is performed.

2. *Non-Parametric Path:* If the distribution exhibits significant skewness ($|\mu - M| > 0.1\mu$) or fails normality testing, the system automatically employs a *one-sample Wilcoxon signed-rank* approximation based on the median and Interquartile Range (IQR).

4.4. Parametric Comparison

When the normality criteria are met, the system calculates a one-sample t-test to determine if the user value significantly differs from the reference mean (μ). The t-statistic is calculated as:

$$t = \frac{\mu - x_{observed\ individual\ score}}{SE} \quad (1)$$

where the Standard Error (SE) is defined as:

$$SE = \frac{\sigma}{\sqrt{n}} \quad (2)$$

The p-value is then derived from the t-distribution with $df = n - 1$ degrees of freedom.

4.5. Non-Parametric Comparison

For the non-normally distributed pathological speech data in the standardized datasets, the system employs a non-parametric approach utilizing the Wilcoxon Signed-Rank test. For a single user-provided observation ($x_{observed\ individual\ score}$), we evaluate the probability of observing a value at least as extreme as $x_{observed\ individual\ score}$ given the reference median (M) and the distribution of ranks. The test statistic W is calculated as:

$$W = \sum_{i=1}^{n} \text{sgn}(x_i - x_{observed\ individual\ score}) \quad (3)$$

In our implementation, the system approximates the p-value by determining the percentile rank of $x_{observed\ individual\ score}$ within the reconstructed distribution of the reference group.

The application automatically generates a clinical explanation of the output following the *Publication Manual of the American Psychological Association* (7th ed.) standards.

4.6. Implementation and Visualization

The statistical engine is implemented using the `jStat` library, allowing for real-time, client-side computation. This ensures low latency and enhances user privacy, as the comparative value x_{user} remains local to the user's browser.

To provide immediate clinical intuition, the system overlays the user's value onto the population distribution using a dynamic reference line across three visualization modes Box Plots, Bar Charts with *Significance Feedback*. The user interface provides a "Significant" (red) or "Similar" (green) status based on an alpha level of $\alpha = 0.05$.

5. Conclusion

In this paper, we presented the Automated Language Biomarker Application (ALBA), an integrated statistical interface designed to bridge the gap between automated linguistic feature extraction and clinical interpretation. By situating ALBA within the Open Brain AI ecosystem, we have created an accessible workflow that converts isolated language discourse measures into actionable "Language Biomarkers." The contribution of this work is three-fold. First, we demonstrate how modern AI pipelines can be coupled with clinical resources like *TalkBank* to automate discourse analysis. Second, ALBA provides quick and easy statistical comparisons of clinical measures, distinguishing between typical linguistic variation and significant pathological deficits and within different conditions based on the available data. Third, by providing real-time, dynamic visualizations and automated significance testing, the application lowers the barrier to entry for evidence-based biomarker screening in busy clinical environments.

Future work will focus on integrating cross-linguistic norms to support the transdiagnostic assessment of multilingual populations. Ultimately, ALBA represents a step toward a more objective, reproducible, and computationally-informed approach to speech and language pathology.

6. Data Availability

The raw linguistic transcripts used to derive the normative benchmarks in ALBA are sourced from the *AphasiaBank* and other clinical repositories within the *TalkBank* system (MacWhinney et al., 2011). Access to these raw data is governed by the TalkBank clinical data-sharing agreement, which requires researcher registration to protect patient confidentiality. Access to the aggregated reference measures (means, standard deviations, and deciles) for all 290 linguistic biomarkers across healthy and clinical cohort in the supplementary material as a machine-readable csv file and a stable URL to the ALBA interface are provided in `https://openbrainai.com/measures` clinical use.

7. References

S. Bengio and G. Heigold. Word embeddings for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1053–1057. Export Date: 20 February 2015.

Lucy Bryant, Alison Ferguson, and Elizabeth Spencer. 2016. Linguistic analysis of discourse in aphasia: A review of the literature. *Clinical linguistics phonetics*, 30(7):489–518.

Lucy Bryant, Alison Ferguson, Megan Valentine, and Elizabeth Spencer. 2018. Implementation of discourse analysis in aphasia: investigating the feasibility of a knowledge-to-action intervention. *Aphasiology*, 33(1):31–57.

Madeline Cruice, Nicola Botting, Jane Marshall, Mary Boyle, Deborah Hersh, Madeleine Pritchard, and Lucy Dipper. 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.

K. T. Cunningham and K. L. Haley. 2020. Measuring lexical diversity for discourse analysis in

aphasia: Moving-average type-token ratio and word information measure. *J Speech Lang Hear Res*, 63(3):710–721.

Gerasimos Fergadiotis. 2011. *Modeling lexical diversity across language sampling and estimation techniques*. Arizona State University.

Markus Forsberg, Dana Dannélls, Lars Borin, and Aleksandrs Berdicevskis. 2025. Background: Språkbanken text. In *Sixty years of Swedish computational lexicography / Dana Dannélls, Kristian Blensenius and Lars Borin (eds.)*, page 161–173. De Gruyter, Berlin.

W. N. Francis and H. Kucera. 1982. *Frequency analysis of English usage*. Houghton-Mifflin Company.

K. C. Fraser, J. A. Meltzer, N. L. Graham, C. Leonard, G. Hirst, S. E. Black, and E. Rochon. 2014. Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex*, 55:43–60. Fraser, Kathleen C Meltzer, Jed A Graham, Naida L Leonard, Carol Hirst, Graeme Black, Sandra E Rochon, Elizabeth eng MOP-82744/Canadian Institutes of Health Research/Canada Research Support, Non-U.S. Gov't Italy Cortex. 2014 Jun;55:43-60. doi: 10.1016/j.cortex.2012.12.006. Epub 2012 Dec 21.

T. E. A. Hansen, J. Praestegaard, T. Tjornhoj-Thomsen, M. Andresen, and B. Norgaard. 2022. Dementia-friendliness in danish and international contexts: A critical discourse analysis. *Gerontologist*, 62(1):130–141. Hansen, Tania E A Praestegaard, Jeanette Tjornhoj-Thomsen, Tine Andresen, Mette Norgaard, Birgitte eng FF2-R69-A1566/Danish Occupational Therapists Association ALZ/Alzheimer's Association/ 2021/05/18 Gerontologist. 2022 Jan 14;62(1):130-141. doi: 10.1093/geront/gnab056.

Swathi Kiran, L. Meier Erin, and P. Johnson Jeffrey. 2019. Neuroplasticity in aphasia: A proposed framework of language recovery. *Journal of Speech, Language, and Hearing Research*, 62(11):3973–3985. Doi: $10.1044/2019_{JSLHR-L-RSNP-19-0054}$.

Alexandra König, Aharon Satt, Alexander Sorin, et al. 2018. Automatic speech analysis for the assessment of patients with predementia and alzheimer's disease. *Alzheimer's Dementia: Diagnosis, Assessment Disease Monitoring*, 1(1):112–124.

Houjun Liu, Brian MacWhinney, Davida Fromm, and Alyssa Lanzi. 2023. Automation of language sample analysis. *Journal of Speech, Language, and Hearing Research*, 66(7):2421–2433.

Brian MacWhinney. The childes project: Tools for analyzing talk: Transcription format and programs, vol. 1, 3rd ed.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307. Doi: 10.1080/02687038.2011.589893.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Brian Paltridge. 2006. *Discourse analysis : an introduction*. Continuum, London. GBA670244 bnb Brian Paltridge. Continuum discourse Includes bibliographical references and index. Formerly CIP. Uk.

Brielle C. Stark, Lucy Bryant, Charalambos Themistocleous, Dirk-Bart den Ouden, and Angela C. Roberts. 2023. Best practice guidelines for reporting spoken discourse in aphasia and neurogenic communication disorders. *Aphasiology*, 37(5):761–784. Doi: 10.1080/02687038.2022.2039372.

Brielle C. Stark and Julia Fukuyama. 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. Doi: 10.1080/23273798.2020.1862258.

Charalambos Themistocleous. 2024. Open brain ai and language assessment. *Frontiers in Human Neuroscience*, 18.

Charalambos Themistocleous, Bronte Ficek, Kimberly Webster, Dirk-Bart den Ouden, Argye E. Hillis, and Kyrana Tsapkini. 2021. Automatic subtyping of individuals with primary progressive aphasia. *Journal of Alzheimer's Disease*, 79:1185–1194.

Charalambos Themistocleous and Brielle C. Stark. 2026. Language biomarker screening using ai: a transdiagnostic approach to brain. *Scientific Reports*.

Charalambos Themistocleous, Kimberly Webster, Alexandros Afthinos, and Kyrana Tsapkini. 2020. Part of speech production in patients with primary progressive aphasia: An analysis based on natural language processing. *American Journal of Speech-Language Pathology*, pages 1–15.

Donna C. Tippett, Katelyn Surrao, Kyriaki Neophytou, Hana Kim, Jessica Gallegos, Charalambos Themistocleous, Brenda Rapp, Argye E. Hillis, and Kyrana Tsapkini. 2025. Written picture descriptions distinguish variants of primary progressive aphasia. *Journal of Alzheimer's Disease*, page 13872877251376381. Doi: 10.1177/13872877251376381.

Maria Varkanitsa, Swathi Kiran, and Klaus Willmes. 2023. Measures of language and communication. *GJ Boyle, Y. Stern, DJ Stein, BJ Sahakian, CJ Golden*, pages 121–144.