Received: 28 August 2021 | Revised: 4 November 2021 | Accepted: 15 November 2021 | Published online: 14 January 2022

DOI: 10.1002/dad2.12276

Alzheimer's & Dementia
Diagnosis, Assessment & Disease Monitoring

RESEARCH ARTICLE

Automated text-level semantic markers of Alzheimer's disease

Camila Sanz[1] | **Facundo Carrillo**[2] | **Andrea Slachevsky**[3,4,5,6,7] | **Gonzalo Forno**[5,8,9] | **Maria Luisa Gorno Tempini**[10] | **Roque Villagra**[4,7] | **Agustín Ibáñez**[11,12,13,14] | **Enzo Tagliazucchi**[1,11] | **Adolfo M. García**[12,13,14,15]

[1] Departamento de Física, Universidad de Buenos Aires and Instituto de Física de Buenos Aires (IFIBA-CONICET)Pabellón ICiudad Universitaria (1428)CABA, Buenos Aires, Argentina

[2] Applied Artificial Intelligence Lab (ICC-CONICET), Pabellón ICiudad Universitaria (1428)CABA, Buenos Aires, Argentina

[3] Memory and Neuropsychiatric Clinic, Neurology Department, Hospital del Salvador (7500000), SSMO & Faculty of Medicine (8380000), University of Chile, Santiago, Chile

[4] Center for Brain Health and Metabolism (GERO) (7500922), Santiago, Chile

[5] Neuropsychology and Clinical Neuroscience Laboratory (LANNEC), Physiopathology Department, Institute of Biomedical Sciences (ICBM), Neuroscience and East Neuroscience Departments, Faculty of Medicine, University of Chile (7500922), University of Chile, Santiago, Chile

[6] Servicio de Neurología, Departamento de Medicina, Clínica Alemana-Universidad del Desarrollo (7550000), Santiago, Chile

[7] East Neuroscience Department, Faculty of Medicine (7650567), University of Chile, Santiago, Chile

[8] School of Psychology, Universidad de los Andes (7550000), Santiago, Chile

[9] Alzheimer's and other cognitive disorders group, Institute of Neurosciences (08035), University of Barcelona, Barcelona, Spain

[10] Memory and Aging Center, Department of Neurology (94143), University of California, San Francisco, California, USA

[11] Latin American Brain Health Institute (BrainLat) (7550000), Universidad Adolfo Ibáñez, Santiago, Chile

[12] Cognitive Neuroscience Center (1644), Universidad de San Andrés, Buenos Aires, Argentina

[13] National Scientific and Technical Research Council (1425), Buenos Aires, Argentina

[14] Global Brain Health Institute (94143), University of California-San Francisco, San Francisco, California, USA; and Trinity College Dublin (D02), Dublin, Ireland

[15] Departamento de Lingüística y Literatura, Facultad de Humanidades (9160000), Universidad de Santiago de Chile, Santiago, Chile

Correspondence
Adolfo M. García PhD, Global Brain Health Institute, University of California, San Francisco; 505 Parnassus Ave, San Francisco, CA 94143, USA.
Email: adolfo.garcia@gbhi.org

Abstract

Introduction: Automated speech analysis has emerged as a scalable, cost-effective tool to identify persons with Alzheimer's disease dementia (ADD). Yet, most research is undermined by low interpretability and specificity.

Methods: Combining statistical and machine learning analyses of natural speech data, we aimed to discriminate ADD patients from healthy controls (HCs) based on automated measures of domains typically affected in ADD: semantic granularity (coarseness of concepts) and ongoing semantic variability (conceptual closeness of successive words). To test for specificity, we replicated the analyses on Parkinson's disease (PD) patients.

Results: Relative to controls, ADD (but not PD) patients exhibited significant differences in both measures. Also, these features robustly discriminated between ADD

patients and HC, while yielding near-chance classification between PD patients and HCs.

Discussion: Automated discourse-level semantic analyses can reveal objective, interpretable, and specific markers of ADD, bridging well-established neuropsychological targets with digital assessment tools.

KEYWORDS
Alzheimer's disease dementia, automated speech analysis, Parkinson's disease, semantic granularity, semantic variability

1 | INTRODUCTION

Over 43 million individuals are affected by Alzheimer's disease (AD), a disorder characterized by progressive temporo-parieto-hippocampal atrophy alongside semantic and episodic memory impairments.[1–3] Given its high disability and mortality rate, its growing economic burden, and its expert-dependent diagnosis,[1,4] a call has been raised for objective, scalable, low-cost approaches favoring disease identification and characterization.[5] Prominent among these is automated speech analysis (ASA).[6] Participants are simply required to speak, yielding diverse features that can be automatically extracted and analyzed to detect persons with and without AD dementia (ADD). Yet, most research is undermined by low interpretability and specificity, often targeting features unrelated to the disorder's core neuropsychological profile while lacking a disease control group.[7] This may cast doubt on the clinical utility of ensuing findings. Here, leveraging ASA tools with ADD patients, healthy controls (HCs), and Parkinson's disease (PD) patients, we examined whether ADD-specific markers can be captured through measures of semantic granularity and ongoing semantic variability, two domains that are systematically disrupted in standard assessments.[8–10]

ASA has proven useful for discriminating between AD patients and HCs,[6] predicting dementia onset,[6,11] and differentiating among autopsy-proven disease subtypes.[12] Yet, most studies have examined heterogeneous ad hoc domains, revealing patterns that are not readily interpretable against core neuropsychological outcomes. For instance, inconsistent accuracy rates are obtained upon targeting mixed articulatory[13] and syntactic[14] dimensions that are typically spared in early testing.[15–17] Also, few studies have included a neurodegenerative control group, prompting questions about the specificity of findings. Moreover, several reports have used unmatched and imbalanced groups, restricted tasks eliciting little data, and suboptimal machine learning approaches.[7]

The present study tackles these issues. We investigated disruptions of semantic granularity and ongoing semantic variability, two well-established manifestations of ADD (Figure 1A). AD patients are typified by coarse (ie, general) conceptual choices, evincing a propensity to use hypernyms (eg, "animal," "fruit") and few hyponyms (eg, "cat," "berry").[9,10] Also, they exhibit sudden changes in speech flow, as their discourse becomes progressively digressive, with frequent interruptions and inquiries (eg, "What was I saying?") causing conceptual discontinuity.[8,18] Our approach captures these phenomena automatically. We employed the WordNet taxonomy[19] to quantify word-by-word semantic granularity (Figure 1B) and FastText embedding to measure ongoing semantic variability across successive word pairs[20] (Figure 1C). Furthermore, to test whether such domains are distinctively affected in ADD, we included persons with PD, a neurodegenerative disease with early semantic alterations restricted to particular domains—mainly, action-related concepts.[21,22] Finally, we circumvented key caveats in the literature.[7] First, we formed strictly matched groups with similar sample sizes. Second, we combined several speech tasks in an integrative analysis, capturing various language behaviors and avoiding inflated results based on unduly small datasets (a key requisite for testing novel metrics). Third, we employed robust machine learning methods for patient identification.

Briefly, we performed the first automated assessment of semantic granularity and variability on ADD patients, relative to HCs and PD patients. We integrated statistical (analysis of variance [ANOVA]) and machine learning (Gradient Boosting) analyses on a rich, diverse set of language tasks. We hypothesized that automated measures of semantic granularity and ongoing semantic variability would yield (1) significant differences, and (2) high classification accuracy between ADD patients and HCs, but (3) not between PD patients and HCs. With this approach, we seek to better test the sensitivity and clinical utility of ASA for dementia assessments.

2 | METHODS

2.1 | Participants

We recruited 55 native Spanish speakers, with normal or corrected-to-normal hearing, from the Memory and Neuropsychiatry Clinic, hosted by Universidad de Chile and Hospital del Salvador, Chile. The sample comprised 21 ADD patients, 18 PD patients, and 16 HCs, reaching adequate power (Appendix A). Patients were diagnosed by expert neurologists following the National Institute of Neurological and Communicative Diseases and Stroke-Alzheimer's Disease and Related Disorders Association clinical criteria for AD, and the United Kingdom Parkinson's Disease Society Brain Bank standards for PD.[23] As in

RESEARCH IN CONTEXT

1. Systematic review: Through a thorough PubMed search, we reviewed the strengths and limitations of automated speech analysis (ASA) research on Alzheimer's disease dementia (ADD). Crucially, most studies targeted features unrelated to the disorder's core neuropsychological profile and lacked disease control groups.
2. Interpretation: Our findings show that ASA can capture interpretable condition-specific markers of ADD. Compared with controls, ADD (but not Parkinson's disease [PD]) patients exhibited significant reductions of semantic granularity and increased semantic variability across speech tasks. Machine learning analyses yielded robust classification of ADD patients (receiver operating characteristic, area under the curve [AUC] = 0.8), but not PD patients (AUC = 0.65), relative to controls. Thus, ASA emerges as an affordable and scalable method to support ADD diagnosis.
3. Future directions: These proposed markers should be examined in larger cohorts (to test their systematicity), in longitudinal designs (to assess their sensitivity to disease progression), and in cross-linguistic studies (to favor more global validations of ASA).

HIGHLIGHTS

- We examined markers of Alzheimer's disease (AD) via automated speech analysis.
- We targeted semantic granularity and variability, two clinically sensitive domains.
- Relative to controls, AD patients were impaired in and classified by both measures.
- These results were not replicated in PD patients.
- Our approach can reveal scalable, interpretable, condition-specific markers of AD.

previous works,[22,24–26] diagnoses were supported by extensive neurological, neuropsychological, and neuroimaging examinations. No patient reported a history of other neurological disorders, psychiatric conditions, primary language deficits, or substance abuse.

Mean scores on the Montreal Cognitive Assessment fell below the cutoffs for dementia in the ADD group and for mild cognitive impairment in the PD group.[27] ADD patients presented executive dysfunction, as established through the INECO Frontal Screening battery.[28] PD patients had no symptoms of Parkinson-plus and were assessed in the "on" phase of medication. HCs were cognitively preserved, functionally autonomous, and had no background of neuropsychiatric disease or drug abuse. All groups were matched for sex, age, and education. For demographic and neuropsychological details, see Table 1.

All participants provided written informed consent pursuant to the Declaration of Helsinki. The study was approved by the institutional ethics committee of the Memory and Neuropsychiatric Clinic, Neurology Department, Hospital del Salvador (7500000), SSMO & Faculty of Medicine, University of Chile.

2.2 | Speech elicitation protocol

Participants performed seven naturalistic language tasks covering varied communicative behaviors. Four were spontaneous speech tasks, requiring participants to describe (1) their daily routine and (2) main interests, and to narrate (3) a pleasant and (4) an unpleasant memory. In these, discourse is driven by personal experience, allowing for varied linguistic patterns.[29] The remaining three were semi-spontaneous speech tasks, involving descriptions of (5) the modified Picnic Scene of the Western Aphasia Battery[30] and (6) a picture of a family working in an unsafe kitchen, as well as (7) immediate recall and narration of a one-minute silent animated film. These tasks elicit diverse and partly predictable linguistic patterns.[29]

Recordings were obtained in a quiet room on laptop computers with noise-cancelling microphones, and saved as .wav files (44100 Hz, 16 bits) via Cool Edit Pro 2.0. Normal pace and volume were encouraged. Recordings were transcribed via an automatic speech-to-text service and manually revised. The rare occurrences of unintelligible words were discarded.

2.3 | Speech data pre-processing

Transcriptions were pre-processed on Python's TreeTagger library with the AnCora Spanish corpus (http://clic.ub.edu/corpus/es/ancora). We converted all characters to lowercase and remove all punctuation marks and symbols.[31,32] Each text was split into individual words. These were assigned part-of-speech tags and lemmatized (ie, converted to their base form). To maximize statistical power and feature diversity while capturing multiple linguistic scenarios, analyses were performed collapsing all tasks. Mean lemmatized word counts did not differ significantly ($F[2,52] = 0.64$, $P = .53$, $\eta_p^2 = 0.02$) among ADD patients (1,051; $SD = 112$), HCs (1,239; $SD = 124$), and PD patients (1,193; $SD = 140$).

2.4 | measures

2.4.1 | Semantic granularity

Granularity scores were computed via Python's NLTK library (https://www.nltk.org/) as interface to access WordNet's lexical database in English (https://wordnet.princeton.edu/). WordNet includes over

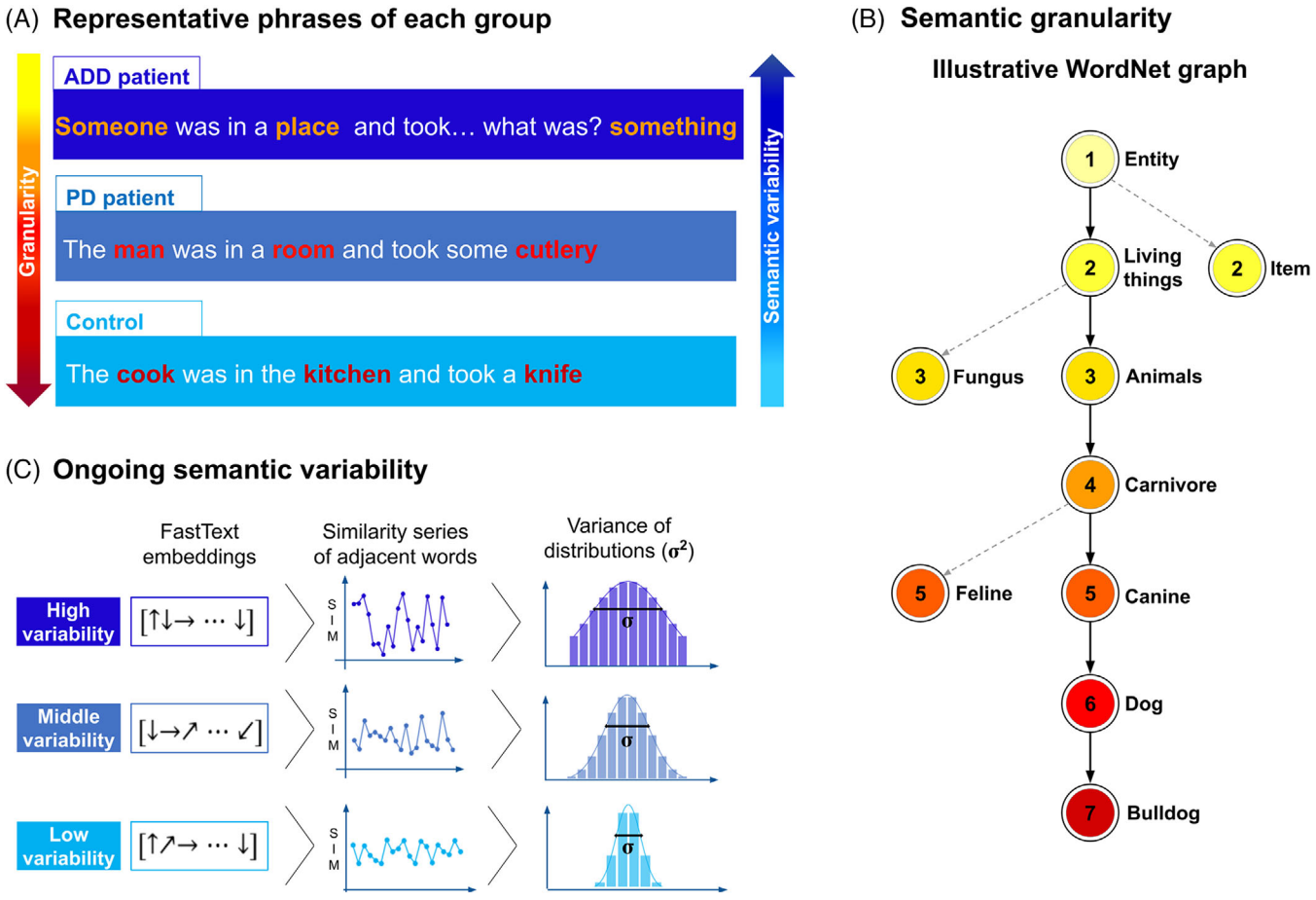

FIGURE 1 Illustration of target measures. **(A)** Representative phrases of ADD patients, PD patients, and healthy controls, showing the predicted gradient of semantic granularity (red scale) and ongoing semantic variability (blue scale). **(B)** Segment of the WordNet network showing hierarchical relations from the least granular node ("entity") to progressively more granular nodes (down to "bulldog"). Granularity values are marked by color and number. Nodes serving as starting points of dotted lines show network bifurcations that do not lead to the "bulldog" node. Multiple relevant and intermediate nodes are omitted for brevity. **(C)** Schemes for the computation of ongoing semantic variability. The diagrams show FastText embeddings, adjacent-word-pair similarity series, and distributions for texts presenting high variability (top row), middle variability (middle row), and low variability (bottom row). Abbreviations: ADD, Alzheimer's disease dementia; PD, Parkinson's disease

155,000 words organized in synonym sets called "synsets." Roughly 80,000 correspond to nouns. These are grouped into a taxonomy that can be visualized as a hierarchical (direct, acyclic, non-weighted) graph spanning hypernyms from above (eg, "animal") and hyponyms from below (eg, "dog").[19] The highest hypernym is "entity," with progressively less coarse terms appearing downstream (Figure 1B). A noun's granularity can be defined as the number of nodes separating it from "entity." Accordingly, general terms like "food" or "animal" have lower granularity scores than more precise terms such as "carrot" or "bulldog."

Nouns were automatically identified with TreeTagger (Section 2.3), manually checked to avoid erroneous tagging, and automatically translated into English using WordNet. Granularity scores were assigned to each noun by considering its shortest path to "entity" (ie, the "synset" with fewer nodes to "entity"). Nouns not included in WordNet's corpus (~ 5.68% across texts) were discarded (rejected nouns did not differ significantly among groups, $P = .93$). For subsequent analyses, scores were stored in lists and converted to histograms using bins of increasing granularity, from 2 to 12 (bins 2, 3, and 4 reflect the number of nouns with granularity scores 2, 3, and 4, respectively, and so on). Bin 1 was not considered, since the word "entity" was not present in any text. Bin 12 included all words with granularity score 12 and the very few words with higher granularity (~0.18% across texts). To avoid verbosity-related confounds, bins were normalized by the total number of nouns.

2.4.2 | Ongoing semantic variability

Ongoing semantic variability was analyzed with a FastText model (https://fasttext.cc/) pre-trained with over 2,000,000 unique Spanish words from Common Crawl and Wikipedia corpora.[33] The FastText model assigns a vector to each unique word in the vocabulary and is trained to map similar concepts to vectors that are close within

TABLE 1 Participants' demographic and neuropsychological information

	ADD ($n = 21$)	PD ($n = 18$)	Controls ($n = 16$)	Statistics (all groups)	Pairwise comparisons		
					Groups	MSE	*P*-value
Demographic data							
Sex (F:M)	13:8	10:8	13:3	$\chi^2 = 4.86$ $P = .1$[a]	––	––	––
Age	77.24 (6.47)	76.50 (6.40)	75.94 (4.35)	$F = 0.21$ $P = .81$[b]	––	––	––
Years of education	11.24 (3.78)	9.39 (5.11)	12.94 (4.28)	$F = 2.62$ $P = .08$[b]	––	––	––
Neuropsychological data							
MoCA	13.90 (4.34)	20.33 (4.68)	25.07 (3.43)	F = 29.01 $P < .001$[b]	ADD vs HCs PD vs HCs ADD vs PD	12.75 29.39 23.27	$< .001$[c] .006[c] $< .001$[c]
IFS battery	11.07 (4.48)	17.08 (4.86)	18.90 (4.26)	$F = 14.30$ $P < .001$[b]	ADD vs HCs PD vs HCs ADD vs PD	13.85 57.72 18.98	$< .001$[c] .51[c] $< .001$[c]

Abbreviations: ADD, Alzheimer's disease dementia; PD, Parkinson's disease; MoCA, Montreal Cognitive Assessment; IFS, INECO Frontal Screening battery. Data presented as mean (*SD*), with the exception of sex.
[a]*P*-values calculated via chi-squared test (χ^2).
[b]*P*-values calculated via independent measures ANOVA.
[c]*P*-values calculated via Tukey's HSD post hoc tests.

the embedding. The distance between words can be quantified with the cosine of the angle between their assigned vectors: $d(u, v) = 1 - cos\cos(u, v) = 1 - \frac{u \cdot v}{\|u\| \|v\|}$, for two vectors u and v.

As in previous works,[20,32] the vector embedding was used to compute each text's ongoing semantic variability (Figure 1C). First, each pre-processed text was represented as a series of vectors, $[v_1, v_2, \ldots]$, preserving the words' sequential order. Second, the distances between adjacent vectors, $d_i = d(v_i, v_{i+1})$ were stored into a time series. Third, ongoing semantic variability was computed as the variance of the joint time series across speech tasks: $\frac{1}{n-1} \sum_{i}^{n-1} (d_i - \mu)^2$, with μ representing the mean of all d_i. Thus, when adjacent words referred to concepts far apart in the embedding space, a text was typified by high semantic variability, reflecting discontinuous discourse. To avoid biases driven by disfluencies, hesitations, or word-finding strategies, consecutive repeated words were omitted before the second step (a text consisting of a single repeated word would feature null variability). Ultimately, each participant's semantic variability across tasks was used for ANOVA and as a feature for machine learning analyses.

2.5 | Statistical analysis

Between-group comparisons were performed via one-way ANOVAs, with Tukey's HSD tests for post hoc contrasts. Alpha levels were set at $P < .05$. Effect sizes were computed via partial eta squared (η_p^2) for ANOVAs and with Cohen's *d* for pairwise comparisons. Given their different distributions and variances, each of the 12 measures (the 11 granularity bins, and the global measure of semantic variability) was framed as a separate dependent variable. No participant was detected as an outlier in any measure. Analyses were performed with Pingouin Python library (https://pingouin-stats.org/).

2.6 | Machine learning analysis

We implemented machine learning classifiers between ADD patients and HCs (to reveal candidate ADD markers) and between PD patients and HCs (to test whether predicted markers proved specific to ADD). A single model was trained for each contrast using the corresponding histograms of granularity and variability scores as input features. Analyses were based on a Gradient Boosting classifier, which surpasses the robustness of other algorithms.[34,35] Scikit-learn (https://scikit-learn.org/) was used to implement the classifiers with 5000 independent estimators, a learning rate of 0.01, and a maximum of two features per split.

For each iteration, data were randomly divided into three folds preserving the proportion of labels (stratified cross-validation). Two folds were used for training and the other for testing, so that all folds were used once to test the classifier. Univariate feature selection was applied to the training set within each fold (the top five features were selected according to their ANOVA *F*-value between groups). This process was repeated 1000 times with and without shuffling the target labels, and a *P*-value was constructed by counting the number of times the area under the receiving operator characteristic curve (ROC AUC) value of the classifier with shuffled labels exceeded that obtained without shuffling, normalized by the total number of iterations. A feature importance score was constructed by counting the number of times a feature was selected based on its *F*-value (Appendix B), divided by the number of folds multiplied by the number of iterations.[36] Importantly, the

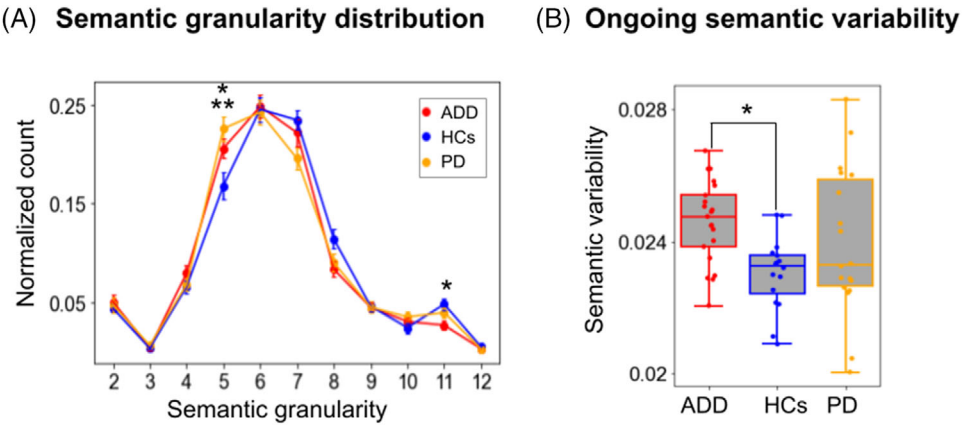

FIGURE 2 Statistical differences in semantic granularity and ongoing semantic variability across diverse speech tasks. **(A)** Normalized values of semantic granularity for each bin. Relative to controls, ADD patients exhibited higher values in a low granularity bin (5) and lower values in a high granularity bin (11), suggesting greater reliance on hypernyms and reduced reliance on hyponyms. **(B)** Boxplot representation of ongoing semantic variability. Successive semantic choices proved significantly more variable in ADD patients than in HCs. Significant pairwise differences ($P < .05$) are indicated with a single asterisk (*) for the contrast between ADD patients and HCs, and with a double asterisk (**) for the contrast between PD patients and HCs. Abbreviations: ADD, Alzheimer's disease dementia; HCs, healthy controls; PD, Parkinson's disease

number of features per participant ($n = 12$) was more than four times smaller than the number of participants ($n = 55$), and the feature selection procedure further reduced the number of features to five. This feature-to-sample ratio, combined with the stratified cross-validation procedure, contributed to alleviate potential overfitting issues. Classifier performance is reported as the mean and *SD* (extent of the shaded region) of the ROC curve across all 1000 iterations, both for shuffled and unshuffled labels, and as confusion matrices showing the proportion of correct/incorrect classifications in each class.

3 | RESULTS

3.1 | Statistical results

ADD patients exhibited lower semantic granularity scores than HCs and PD patients in most of the largest bins (8-12), indicating scarcer use of hyponyms (Figure 2A). Significant group differences were found for bins 5 ($F[2,52] = 5.43$, $P = .007$, $\eta_p^2 = 0.17$) and 11 ($F[2,52] = 4.71$, $P = .013$, $\eta_p^2 = 0.15$). Post hoc analyses, via Tukey's HSD tests, revealed that ADD patients scored significantly higher than HCs in bin 5, a low granularity bin ($P = .072$, $d = 0.73$); and significantly lower than HCs in bin 11, a high granularity bin ($P = .008$, $d = 1$). Bin 5 also yielded significantly higher scores for PD patients than HCs ($P = .003$, $d = 1.11$). The remaining pairwise comparisons yielded non-significant differences (all *P*-values > .05).

Ongoing semantic variability results (Figure 2B) yielded a significant group effect ($F[2,52] = 4.24$, $P = .02$, $\eta_p^2 = 0.14$), with post hoc comparisons revealing significantly higher scores for ADD patients than HCs ($P = .011$, $d = 0.97$), alongside non-significant differences for the remaining pairwise comparisons (HCs vs PD patients: $P = .21$, $d = 0.58$; ADD vs PD patients: $P = .45$, $d = 0.39$).

3.2 | Machine learning results

Collapsing both measures, classification between ADD patients and HCs (Figure 3A) reached an AUC of .80 ± .06 (accuracy: .71 ± .11; sensitivity: .80 ± .15; precision: .73 ± .12). This AUC value was significantly higher ($P = .022$) than that obtained upon shuffling participants' labels, which yielded chance levels (0.49 ± .13) and lower scores across measures (accuracy: .52 ± .16; sensitivity: .64 ± .21; precision: .57 ± .18).

Conversely, classification between PD patients and HCs (Figure 3B) yielded an AUC of .65 ± .08 (accuracy: .60 ± .12; sensitivity: .61 ± .20; precision: .64 ± .15). This AUC value did not differ significantly ($P = .16$) from that obtained upon shuffling participants' labels, which yielded chance values (.50 ± .13) and chance-level results in other measures (accuracy: .50 ± .15; sensitivity: .56 ± .23; precision: .53 ± .20).

4 | DISCUSSION

We examined potential markers of ADD via automated measures of semantic granularity and variability. Both measures discriminated ADD patients from HCs (based on ANOVAs) and allowed identifying them robustly on a subject-level basis (based on machine learning). No such differentiations were present for PD patients relative to HCs. Below we discuss these findings.

Relative to HCs, ADD patients used more coarse and fewer precise concepts. This indicates reduced semantic granularity, a phenomenon observed in controlled tasks (eg, picture naming, category fluency) through standard measures (eg, correct responses).[9] Our study suggests that increased reliance on hypernyms in ADD also typifies the patients' natural speech. In this sense, reduced granularity has been proposed as a marker of diseases with primary semantic memory impairments.[37] Indeed, abnormally coarse-grained abstractions are

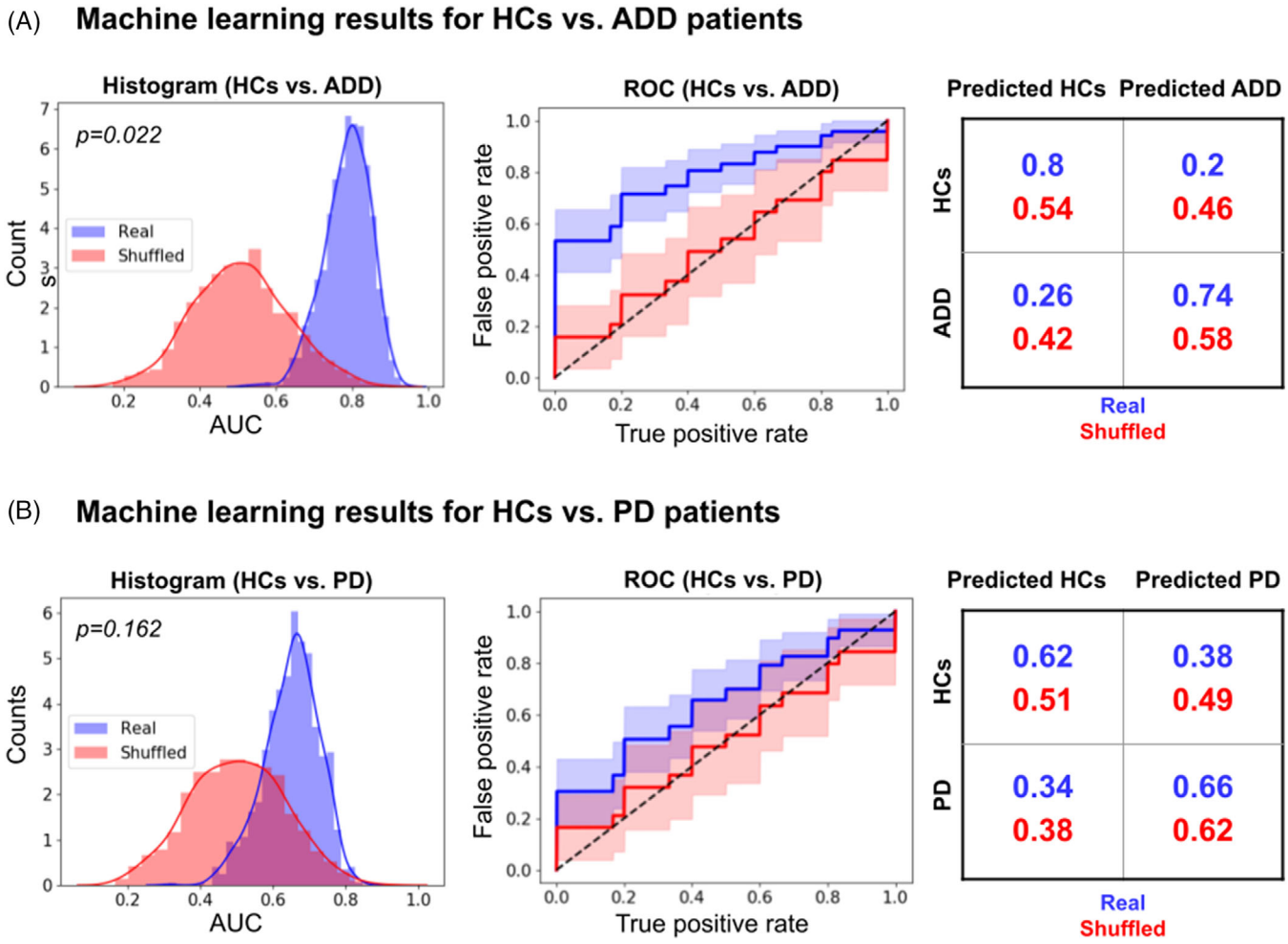

FIGURE 3 Classifications between patients and controls combining semantic granularity and ongoing semantic variability features across diverse speech tasks. The Gradient Boosting classifier successfully distinguished **(A)** ADD patients from HCs, but not **(B)** PD patients from HCs. The panels show normalized AUC histograms (left inset), average ROC curves (middle inset), and confusion matrices normalized by row and averaged across iterations (right inset). Real results are shown in blue, while results obtained upon shuffling participants' labels are shown in red. Abbreviations: ADD, Alzheimer's disease dementia; AUC, area under the curve; HCs, healthy controls; PD, Parkinson's disease; ROC, receiver operating characteristic

also typical in semantic dementia patients,[37] some of whose core atrophy regions (eg, hippocampus, temporal lobes) are also affected in AD.[38] Accordingly, although several granularity bins showed substantial overlap between ADD patients and HCs, our automated granularity measure might capture subtle but informative disruptions.

ADD patients also presented greater semantic variability than HCs, indicating more discontinuous speech (eg, see Appendix C). Previous studies have reported reduced cohesion and coherence in AD,[18,39] for example, by counting digressive utterances (or words) or unrelated adjacent utterances.[8,18] Similar patterns are observed in persons with mild cognitive impairment, at increased risk for AD.[40] Our study shows that dissimilar semantic relations also emerge across word-to-word relations. Specifically, the patients' discourse abounded in interruptions and gap fillers via ready-made phrases (eg, "I don't know," "I forget the name," "I don't remember"), in line with evidence that this population may overuse formulaic language.[41] Here, the FastText word-vectorial representations revealed that such phrases deviate from their adjacent semantic choices, revealing further neuropsychological aspects of ADD.

The robustness of both measures was corroborated by machine learning results. Joint analysis of semantic granularity and variability features yielded an AUC of 80%, correctly identifying 80% of HCs and 74% of ADD patients. These results surpass those from previous ASA studies targeting domains that are not markedly affected in AD, such as articulation or syntax.[15–17] Importantly, classification results were near chance upon shuffling participants' labels, indicating that these features do capture distinguishing properties of ADD rather than fortuitous differences between random samples. Briefly, semantic granularity and variability measures may contribute to revealing clinically relevant differences between ADD patients and HCs.

Importantly, the above results were partly specific to ADD. Except for one granularity bin, the features affected in ADD were preserved in PD. Likewise, classification between PD patients and HCs was near chance and non-significantly different from that obtained via random

groupings. This is a non-trivial finding, since other verbal domains more systematically assessed in AD, such as semantic and phonemic fluency,[42] are also frequently compromised in PD,[43] limiting their use for disease differentiation. Yet, while we targeted PD patients on levodopa, as in previous works,[25] semantic alterations in this disease are sensitive to medication status.[44] New studies should explore whether the deficits observed in ADD remain specific when considering PD patients with varying levels of dopamine bioavailability. Still, our findings suggest that theoretically informed semantic measures may prove useful not only to identify specific brain diseases, but also to discriminate among them.

Previous ASA studies have often assessed unmotivated, heterogeneous domains in combination with feature importance techniques that favor classification outcomes over interpretability.[7] While often successful in terms of classifier performance, this approach fails to capture features that can be readily aligned with mainstream clinical knowledge. In fact, diverse constellations of phonological and syntactic features might contribute to patient identification[13,14] while challenging straightforward neuropsychological interpretation. Moreover, this evidence is hard to reconcile with abundant neuropsychological literature attesting to the preservation of such domains in AD.[15–17] In contrast, we first identified linguistic domains consistently affected by the disease and then developed a pipeline to track them in natural discourse. By bridging the gap between well-established deficits and cutting-edge automated tools, our approach paves the way for more clinically relevant uses of ASA.

Moreover, our design overcomes key limitations of previous ASA research on AD and related diseases. Frequently, these studies are undermined by unbalanced samples[45,46] and by poor or null control of sociodemographic confounds, such as sex, age, and education.[45,47] Our strict group-matching protocol circumvented major alternative explanations of our results (ie, higher education levels could entail richer vocabulary, potentially increasing semantic granularity). Moreover, while most previous works used isolated tasks or narrow combinations therefrom, we used a range of spontaneous (autobiographical) and semi-spontaneous (stimulus-based) tasks,[29] covering a rich repertoire of daily linguistic behaviors. Critically, this approach increases data quantity and variability across groups, avoiding over-optimistic results from brief discourse samples. While we obtained similar results even upon considering a single task—the one of longest duration (Appendix, section D)—the present approach avoids important caveats while maximizing the representativeness of ASA.

This study attests to the usefulness of ASA as a complement for mainstream AD assessments. Standard evaluations of neurodegenerative conditions may prove expensive, yield examiner-driven scores,[48] and overlook spontaneous behavior.[22] Conversely, ASA entails minimal costs, generating objective naturalistic data.[5,49] Furthermore, speech tasks can be administered remotely, maximizing accessibility and equity for persons with reduced mobility or capacity to afford transportation costs. These possibilities open exciting avenues to further test our measures.

Yet, our work is not without limitations. First, although groups were balanced and in keeping with the field's typical *Ns*,[8] their sizes were small. While this is a common hurdle in studies pursuing standardized, good-quality speech samples, it would be important to replicate our work with more participants. Second, while the use of several speech tasks allows capturing diverse linguistic behaviors, it also increases test duration. This can be attenuated by having participants record themselves remotely, which could be especially promising for longitudinal assessments. Third, our study focused exclusively on Spanish speakers. Given that different languages may become differently affected by the same disease,[50] cross-linguistic studies would be critical towards more global approaches to ASA.

In sum, this study shows that ASA can be leveraged to yield differential and interpretable markers of ADD across diverse linguistic behaviors. ADD patients seem typified by reduced semantic granularity and higher ongoing semantic variability, both patterns being absent in PD patients. By further targeting well-established linguistic aspects of ADD through customized methods, ASA may boost the development of digital markers of dementia.

ACKNOWLEDGEMENT

Andrea Slachevsky has received support from ANID/FONDAP/15150012, ANID/FONDEF/18I10113, ANID/Fondecyt/1191726, 1210176, and 1210195; and MULTI-PARTNER CONSORTIUM TO EXPAND DEMENTIA RESEARCH IN LATIN AMERICA (ReDLat, supported by National Institutes of Health, National Institutes of Aging [R01 AG057234], Alzheimer's Association [SG-20-725707], Tau Consortium, and Global Brain Health Institute) and Alzheimer's Association GBHI ALZ UK-20-639295. In the past 36 months, Andrea Slachevsky has received grants from ANID/FONDAP/15150012, ANID/Fondecyt Regular/1191726, and the MULTI-PARTNER CONSORTIUM TO EXPAND DEMENTIA RESEARCH IN LATIN AMERICA (ReDLat, supported by National Institutes of Health, National Institutes of Aging [R01 AG057234], Alzheimer's Association [SG-20-725707], Tau Consortium, and Global Brain Health Institute) and Alzheimer's Association GBHI ALZ UK-20-639295. In the past 36 months, Andrea Slachevsky has served as Board Director for the Global Brain Health Institute and BrainLat, Member of the Scientific Program Committee of the Alzheimer's Association International Congress AAIC, and Vice President of the non-profit organization COPRAD (Corporacion Profesional de Alzheimer y Otras demencia). Facundo Carrillo has stocks of a company that makes EHR for mental health professionals (Sigmind: https://www.sigmind.net). In the past 36 months, Facundo Carrillo had a fellowship with a travel grant. In the past 36 months, Gonzalo Forno has served as Associate Professor at Universidad de Los Andes, Santiago, Chile. In the past 36 months, Maria Luisa Gorno Tempini has been supported by grants from the National Institutes of Health (NINDS R01 NS050915, NIDCD K24 DC015544; NIA U01 AG052943). Roque Villagra has received support from ANID/FONDAP/15150012. In the past 36 months, Agustín Ibáñez has been partially supported by grants from CONICET; ANID/FONDECYT Regular (1170010); FONCYT-PICT 2017-1820; ANID/FONDAP/15150012; Takeda CW2680521; Sistema General de Regalías (BPIN2018000100059), Universidad del Valle (CI 5316) Alzheimer's Association GBHI ALZ UK-20-639295; and the

MULTIPARTNER CONSORTIUM TO EXPAND DEMENTIA RESEARCH IN LATIN AMERICA (ReDLat, supported by National Institutes of Health, National Institutes of Aging [R01 AG057234], Alzheimer's Association [SG-20-725707], Rainwater Charitable foundation - Tau Consortium, and Global Brain Health Institute). The contents of this publication are solely the responsibility of the authors and do not represent the official views of these Institutions. Adolfo García is an Atlantic Fellow at the Global Brain Health Institute (GBHI) and is supported with funding from GBHI, Alzheimer's Association, and Alzheimer's Society (GBHI ALZ UK-22-865742); CONICET; FONCYT-PICT (grant number 2017-1818); ANID, FONDECYT Regular (grant numbers 1210176 and 1210195); and Programa Interdisciplinario de Investigación Experimental en Comunicación y Cognición (PIIECC), Facultad de Humanidades. In the past 36 months, Adolfo García has received grants from the GBHI, the Alzheimer's Association, the Alzheimer's Society, and ANID (FONDECYT Regular 1210176). He has also served as advisory board member for the BrainLat Institute (Chile). No payments are involved in this appointment. In the past 36 months no author has received any royalties, licenses, consulting fees; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing, or educational events, expert testimony; support for attending meetings and/or travel; equipment, materials, drugs, medical writing, gifts, or other services. In the past 36 months, no author has had any patents planned, issued, or pending; nor any financial or non-financial interests. Camila Sanz, Roque Villagra, and Enzo Tagliazucchi have nothing to disclose. We express our deep gratitude to all participants as well as the patients' caregivers for contributing their valuable time to this study.

ORCID

Adolfo M. García https://orcid.org/0000-0002-6936-0114

REFERENCES

1. Nichols E, Szoeke CE, Vollset SE, et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol*. 2019;18(1):88-106.
2. Ozel-Kizil ET, Bastug G, Kirici S. Semantic and episodic memory performances of patients with Alzheimer's disease and minor neurocognitive disorder. *Alzheimers Dement*. 2020;16(S6):e039310.
3. Scheltens P, De Strooper B, Kivipelto M, et al. Alzheimer's disease. *Lancet*. 2021;397(10284):1577-1590.
4. Nakamura AE, Opaleye D, Tani G, Ferri CP. Dementia underdiagnosis in Brazil. *Lancet*. 2015;385(9966):418-419.
5. Laske C, Sohrabi HR, Frost SM, et al. Innovative diagnostic tools for early detection of Alzheimer's disease. *Alzheimers Dement*. 2015;11(5):561-578.
6. König A, Satt A, Sorin A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimers Dement (Amst)*. 2015;1(1):112-124.
7. de La Fuente Garcia S, Ritchie CW, Luz S. Artificial intelligence, speech, and language processing approaches to monitoring Alzheimer's Disease: a systematic review. *J Alzheimers Dis*. 2020;78(4):1547-1574.
8. Dijkstra K, Bourgeois MS, Allen RS, Burgio LD. Conversational coherence: discourse analysis of older adults with and without dementia. *J Neurolinguistics*. 2004;17(4):263-283.
9. Giffard B, Desgranges B, Nore-Mary F, et al. The nature of semantic memory deficits in Alzheimer's disease: new insights from hyperpriming effects. *Brain*. 2001;124(8):1522-1532.
10. Hodges JR, Salmon DP, Butters N. Semantic memory impairment in Alzheimer's disease: failure of access or degraded knowledge? *Neuropsychologia*. 1992;30(4):301-314.
11. Eyigoz E, Mathur S, Santamaria M, Cecchi G, Naylor M. Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*. 2020;28:100583.
12. Rentoumi V, Raoufian L, Ahmed S, de Jager CA, Garrard P. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *J Alzheimers Dis*. 2014;42(s3):S3-S17.
13. Fraser KC, Meltzer JA, Rudzicz F. Linguistic features identify Alzheimer's disease in narrative speech. *J Alzheimers Dis*. 2016;49(2):407-422.
14. Orimaye SO, Wong JS, Golden KJ. Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In: Resnik P, Resnik R, Mitchell M, eds. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Stroudsburg, PA: Association for Computational Linguistics; 2014:78-87.
15. Ahmed S, Haigh AM, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. *Brain*. 2013;136(12):3727-3737. https://doi.org/10.1093/brain/awt269.
16. Kaprinis S, Stavrakaki S. Morphological and syntactic abilities in patients with Alzheimer's disease. *Brain Lang*. 2007;103(1-2):59-60.
17. Weiner MF, Neubecker KE, Bret ME, Hynan LS. Language in Alzheimer's disease. *J Clin Psychiatry*. 2008;69(8):1223-1227.
18. Pistono A, Jucla M, Bézy C, Lemesle B, Le Men J, Pariente J. Discourse macrolinguistic impairment as a marker of linguistic and extralinguistic functions decline in early Alzheimer's disease. *Int J Lang Commun Disord*. 2019;54(3):390-400.
19. Bird S, Klein E, Loper E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly; 2009.
20. Corcoran CM, Carrillo F, Fernández-Slezak D, et al. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*. 2018;17(1):67-75.
21. Birba A, García-Cordero I, Kozono G, et al. Losing ground: frontostriatal atrophy disrupts language embodiment in Parkinson's and Huntington's disease. *Neurosci Biobehav Rev*. 2017;80:673-687.
22. García AM, Bocanegra Y, Herrera E, et al. Parkinson's disease compromises the appraisal of action meanings evoked by naturalistic texts. *Cortex*. 2018;100:111-126.
23. Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J Neurol Neurosurg Psychiatry*. 1992;55(3):181-184.
24. Eyigoz E, Courson M, Sedeño L, et al. From discourse to pathology: automatic identification of Parkinson's disease patients via morphological measures across three languages. *Cortex*. 2020;132:191-205.
25. García AM, Carrillo F, Orozco-Arroyave JR, et al. How language flows when movements don't: an automated analysis of spontaneous discourse in Parkinson's disease. *Brain Lang*. 2016;162:19-28.
26. Santamaría-García H, Baez S, Reyes P, et al. A lesion model of envy and Schadenfreude: legal, deservingness and moral dimensions as revealed by neurodegeneration. *Brain*. 2017;140(12):3357-3377.
27. Delgado C, Araneda A, Behrens MI. Validación del instrumento Montreal Cognitive Assessment en español en adultos mayores de 60 años. *Neurología*. 2019:34;376-385.
28. Torralva T, Roca M, Gleichgerrcht E, López P, Manes F. INECO Frontal Screening (IFS): a brief, sensitive, and specific tool to assess executive functions in dementia. *J Int Neuropsychol Soc*. 2009;15(5):777-786.

29. Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected speech in neurodegenerative language disorders: a review. *Front Psychol*. 2017;8:269.
30. Kreutzer J S, DeLuca J, and Caplan B, eds. *Encyclopedia of Clinical Neuropsychology*. Springer-Verlag; 2011. Accessed April 17, 2021. https://www.springer.com/gp/book/9780387799476
31. Sanz C, Zamberlan F, Erowid E, Erowid F, Tagliazucchi E. The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports. *Front Neurosci*. 2018;12:7.
32. Sanz C, Pallavicini C, Carrillo F, et al. The entropic tongue: disorganization of natural language under LSD. *Conscious Cogn*. 2021;87:103070.
33. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T. Learning word vectors for 157 languages. Preprint. Posted online February 19, 2018. Updated March 28, 2018. ArXiv1802.06893v2. http://arxiv.org/abs/1802.06893. [Format follows that for preprints in AMA *Manual of Style*, 11th ed., Section 3.11.4.1.]
34. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367-378.
35. Moguilner S, Birba A, Fino D, et al. Multimodal neurocognitive markers of frontal lobe epilepsy: insights from ecological text processing. *NeuroImage*. 2021;235:117998.
36. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507-2517.
37. Patterson K, Nestor PJ, Rogers TT. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat Rev Neurosci*. 2007;8(12):976-987.
38. Chapleau M, Aldebert J, Montembeault M, Brambati SM. Atrophy in Alzheimer's disease and semantic dementia: an ALE meta-analysis of voxel-based morphometry studies. *J Alzheimers Dis*. 2016;54(3):941-955.
39. Ash S, Moore P, Vesely L, Grossman M. The decline of narrative discourse in Alzheimer's disease. *Brain Lang*. 2007;103(1-2):181-182.
40. Drummond C, Coutinho G, Fonseca RP, et al. Deficits in narrative discourse elicited by visual stimuli are already present in patients with mild cognitive impairment. *Front. Aging Neurosci*. 2015;7:96.
41. Zimmerer VC, Wibrow M, Varley RA. Formulaic language in people with probable Alzheimer's disease: a frequency-based approach. *J Alzheimers Dis*. 2016;53(3):1145-1160.
42. Laws KR, Duncan A, Gale TM. 'Normal' semantic-phonemic fluency discrepancy in Alzheimer's disease? A meta-analytic study. *Cortex*. 2010;46(5):595-601.
43. Henry JD, Crawford JR. Verbal fluency deficits in Parkinson's disease: a meta-analysis. *J Int Neuropsychol Soc*. 2004;10(4):608-622.
44. Norel R, Agurto C, Heisig S, et al. Speech-based characterization of dopamine replacement therapy in people with Parkinson's disease. *NPJ Parkinsons Dis*. 2020;6(1):12.
45. Luz S. Longitudinal monitoring and detection of Alzheimer's type dementia from spontaneous speech data. In: Panagiotis D, Bamidis PD, Konstantinidis ST, Rodrigues PP, eds. *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. Piscataway, NJ: IEEE Computer Society; 2017:45-46.
46. Weiner J, Herff C, Schultz T. Speech-based detection of Alzheimer's disease in conversational German. In: International Speech Communication Association, ed. Interspeech 2016, 17th Annual Conference of the International Speech Communication Association. Red Hook, NY: Curran Associates; 2016:1938-1940.
47. Prud'hommeaux ET, Roark B. Alignment of spoken narratives for automated neuropsychological assessment. In: Institute of Electrical and Electronics Engineers, ed. *2011 IEEE Workshop on Automatic Speech Recognition Understanding*. Piscataway, NJ: IEEE Computer Society; 2011:484-489. https://doi.org/10.1109/ASRU.2011.6163979.
48. Orimaye SO, Wong JS, Wong CP. Deep language space neural network for classifying mild cognitive impairment and Alzheimer-type dementia. *PLoS One*. 2018;13(11):e0205636.
49. García AM, Arias-Vergara T, Vasquez-Correa J, et al. Cognitive determinants of dysarthria in Parkinson's disease: an automated machine learning approach. *Mov Disord*. 2021;36(12):2862-2873.
50. Canu E, Agosta F, Battistella G, et al. Speech production differences in English and Italian speakers with nonfluent variant PPA. *Neurology*. 2020;94(10):e1062-e1072.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Sanz C, Carrillo F, Slachevsky A, et al. Automated text-level semantic markers of Alzheimer's disease. *Alzheimer's Dement*. 2022;14:e12276. https://doi.org/10.1002/dad2.12276