Impaired language in Alzheimer's disease: A comparison between English and Persian implicates content-word frequency rather than the noun-verb distinction

Mahya Sanati¹, Sabereh Bayat², Mehrdad Mohammad Panahi³, Amirhossein Khodadadi⁴, Sahar Rezaee⁵, Mahdieh Ghasimi⁵, Sara Besharat⁵, Zahra Mahboubi Fooladi⁵, Mostafa Almasi Dooghaee, Morteza Sanei Taheri⁵, Bradford C Dickerson⁶, Adele Goldberg⁷, Neguine Rezaei⁶


NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

This study challenges the conventional psycholinguistic view that the distinction between nouns and verbs is pivotal in understanding language impairments in neurological disorders. Traditional views link frontal brain region damage with verb processing deficits and posterior temporoparietal damage with noun difficulties. However, this perspective is contested by findings from patients with Alzheimer’s disease (pwAD), who show impairments in both word classes despite their typical temporoparietal atrophy. Notably, pwAD tend to use semantically lighter verbs in their speech than healthy individuals. By examining English-speaking pwAD and comparing them with Persian-speaking pwAD, this research aims to demonstrate that language impairments in Alzheimer’s disease (AD) stem from the distributional properties of words within a language rather than distinct neural processing networks for nouns and verbs. We propose that the primary deficit in AD language production is an overreliance on high-frequency words. English has a set of particularly high-frequency verbs that surpass most nouns in usage frequency. Since pwAD tend to use high-frequency words, the byproduct of this word distribution in the English language would be an over-usage of high-frequency verbs. In contrast, Persian features complex verbs with an overall distribution lacking extremely high-frequency verbs like those found in English. As a result, we hypothesize that Persian-speaking pwAD would not have a bias toward the overuse of high-frequency verbs.

We analyzed language samples from 95 English-speaking pwAD and 91 healthy controls, along with 27 Persian-speaking pwAD and 27 healthy controls. Employing uniform automated natural language processing methods, we measured the usage rates of nouns, verbs, and word frequencies across both cohorts.

Our findings showed that English-speaking pwAD use higher-frequency verbs than healthy individuals, a pattern not mirrored by Persian-speaking pwAD. Crucially, we found a significant interaction between the frequencies of verbs used by English and Persian speakers with and without AD. Moreover, regression models that treated noun and verb frequencies as separate predictors did not outperform models that considered overall word frequency alone in classifying AD.

In conclusion, this study suggests that language abnormalities among English-speaking pwAD reflect the unique distributional properties of words in English rather than a universal noun-verb class distinction. Beyond offering a new understanding of language abnormalities in AD, the study highlights the critical need for further investigation across diverse languages to deepen our insight into the mechanisms of language impairments in neurological disorders.
Introduction

The distinction between nouns and verbs is considered central to language. While a few languages may lack distinct classes of adjectives, prepositions, or adverbs, it is unclear that any language entirely abolishes the distinction between nouns and verbs. A noun-verb distinction may be inevitable for functional reasons. Effective communication typically requires a subject to talk about (usually denoted by nouns) and some comment about that subject (often conveyed through verbs). Owing to its centrality, many language impairments have been examined through the lens of the noun and verb distinction. This approach has led to the conventional perspective that patients with frontal lobe lesions produce fewer verbs, while those with posterior brain lesions produce fewer nouns. This pattern of impairment is often attributed to the assumed neurobiological functions of these word classes, with verbs processed in areas associated with actions near the motor cortex, and nouns linked to the posterior sensory association areas. However, more detailed psycholinguistic analysis of the language produced by people with brain damage indicates this conventional understanding may be oversimplified. For instance, although patients with frontal lobe damage typically use fewer verbs, the verbs they produce are often semantically richer or “heavier” than unimpaired controls.

Another complex pattern is evident among people with Alzheimer’s disease (pwAD), who typically have predominant neurodegeneration in temporoparietal regions. The conventional view predicts pwAD should show greater deficits with nouns than verbs. However, pwAD have deficits in production and comprehension of both nouns and verbs compared to healthy individuals, and they display worse performance on verbs relative to nouns at the single-word level. In connected discourse, pwAD tend to rely on nonspecific and semantically light verbs during sentence production. In explaining why pwAD over-rely on nonspecific verbs, Kim et al. proposed that the deficit stems from the semantic complexity associated with heavier verbs. In particular, more specific verbs, such as “walk” or “ski”, may well be associated with more specific nouns, such as “shoes” or “boots,” compared to less specific verbs, such as “go”. As the specific noun associates degrade in pwAD, verb retrieval becomes more challenging. Therefore, according to this view, access to verbs is indirectly affected by the loss of specific nouns rather than through a direct impairment in verb access.

In the current work, we test a different hypothesis: the language spoken by pwAD is not directly influenced by the distinction between nouns and verbs but is instead a byproduct of content word frequency in a particular language. A condition such as AD, which makes lexical access more difficult, is recognized to involve an overreliance on high-frequency words compared to controls, who are better able to access words that specifically suit their intended messages. We propose that English-speaking AD patients over-rely on certain verbs, an observation replicated here, because of the distributional properties of verbs in English. In particular, English has a set of very high-frequency main verbs (e.g., be, do, have, know, go, get, think, come, see, take), which are more frequent than nearly any noun (e.g., people, time, or thing) (Figure 1). English also has a large set of low-frequency verbs, partly due to the existence of Latinate synonyms for common verb meanings (e.g., exist, act, acquire, understand, possess, consider, arrive, visualize) and partly because English verbs often encode manner (e.g., waltz, shuffle, saunter) or a resultant state (redden, flatten, amuse). Therefore, examining verb impairment only in
English makes it challenging to disentangle various contributing factors—such as word class and word frequency—that may underlie the observed language features in AD. In this context, an investigation of AD speech in a language with different distributional properties than English is used to distinguish the role of grammatical category (nouns vs. verbs) and frequency.

**Figure 1.** Distribution of verb and noun lemmas in English based on the spoken modality of The Corpus of Contemporary American English (COCA)\(^{27}\). Most of the highest frequency words are semantically light verbs.

For these reasons, we conducted this study to compare the production of verbs and nouns by English speakers with that of Persian speakers. Like English, Persian (also known as Farsi) has light verbs that can appear as main verbs, but Persian’s light verbs are commonly used as part of complex verbs. Complex verbs consist of a non-verb element, the host, followed by a light verb.\(^{28-30}\) The nature of Persian verbs results in a wide variety of complex verbs instead of a collection of high-frequency main verbs, as in English. Thus, a comparative analysis of Persian and English provides a unique vantage point that allows us to ask whether English-speaking pwAD may over-rely on certain verbs because they are particularly frequent rather than because of a difference between nouns and verbs. If the noun-verb distinction is an explanatory factor in language impairment among pwAD, we would expect consistent challenges with verbs (or nouns) to be evident across both languages. We hypothesized that the data would not support this possibility. Instead, the core problem in pwAD is access to low-frequency, semantically richer words, which would impact lexical classes differently in languages with different distributional properties of word classes.

**Methods**

Participants
The English cohort. We obtained English samples from DementiaBank, a component of the TalkBank project. The dataset was collected at the University of Pittsburgh as part of the Alzheimer Research Program between 1983 and 1988 during a 5-year-follow up (comprehensive information available in Becker et al. 1994). Inclusion criteria consisted of being older than 44 years old, having at least seven years of education, not having previous neurologic disorders, not taking neuroleptic drugs, having at least a score of 10 in Mini-Mental State Exam (MMSE) and being able to give informed consent. Participants were assessed using comprehensive examinations, including neuropsychiatric batteries, laboratory data, CT scans, and EEGs. From this dataset, we included 95 patients with probable AD and 91 age-matched healthy individuals and included each person’s first language sample if more than one was collected. We selected the subset of pwAD from the DementiaBank in such a way that they match the MMSE and age of pwAD in the Persian cohort as closely as possible (Table 1).

The Persian cohort. Twenty-seven Persian-speaking pwAD and 27 age-matched healthy individuals were recruited from the Brain and Cognition Clinic in Tehran, Iran. A complete clinical history was taken from patients and their caregivers. Demographic features, clinical presentations, medical, and family histories were included in the interviews. A complete clinical examination was performed, emphasizing the assessment of motor features such as parkinsonism, praxis, language and speech, gait, and balance. The cognitive examination included Addenbrook’s Cognitive Examination-Persian version (ACE) and Mini-Mental Status Exam (MMSE). Blood tests and the brain MRI of all patients were reviewed to confirm the diagnosis and rule out other medical conditions. The blood tests included complete blood count, biochemistry, renal and liver function tests, vitamin B12, 25-OH-D3, thyroid function test, syphilis, and HIV serologic test. Mild or Major Neurocognitive Disorders and Alzheimer’s disease were diagnosed by DSM-5 and NINCDS-ADRDA criteria, respectively. In addition to obtaining a clinical history for all healthy individuals, the neurotypicality was confirmed using MMSE, brain MRI, or both. The research section of the Persian cohort was approved by the ethics committee of Iran University of Medical Sciences, which governs human subjects research in accordance with their guidelines. All participants provided written informed consent to participate in this study.

Table 1. Demographic and clinical characteristics of pwAD and healthy speakers of English and Persian

<table>
<thead>
<tr>
<th></th>
<th>AD</th>
<th>Healthy individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English cohort</strong></td>
<td>N = 95</td>
<td>N = 91</td>
</tr>
<tr>
<td>MMSE (mean ± SD)</td>
<td>18.99 (4.35)</td>
<td>28.99 (1.10)</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>67.11 (6.01)</td>
<td>65.02 (6.81)</td>
</tr>
<tr>
<td>Handedness (percent right-handed)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sex (percent female)</td>
<td>68.42%</td>
<td>59.34%</td>
</tr>
<tr>
<td><strong>Persian cohort</strong></td>
<td>N = 27</td>
<td>N = 27</td>
</tr>
<tr>
<td>MMSE (mean ± SD)</td>
<td>18.15 (4.89)</td>
<td>28.28 (1.23)</td>
</tr>
<tr>
<td>Age (mean ± SD)</td>
<td>76.07 (5.49)</td>
<td>68.48 (9.38)</td>
</tr>
<tr>
<td>Handedness (percent right-handed)</td>
<td>100%</td>
<td>96.30%</td>
</tr>
<tr>
<td>Sex (percent female)</td>
<td>74.07%</td>
<td>51.58%</td>
</tr>
</tbody>
</table>
Language samples. Connected speech samples were obtained for both languages by asking participants to describe the Cookie Theft picture, a component of the Boston Diagnostic Aphasia Examination. The recorded samples were transcribed into text for further analysis.

Language features. We used Stanza, an open-source Python natural language processing toolkit that supports 66 human languages, including English and Persian, to extract part-of-speech (POS) tags such as verbs and nouns. The dependency relation "compound:lvc" in Stanza denotes complex verbs in Persian.

Word frequency. Frequencies of both nouns and verbs were calculated for each language on a corpus of word counts extracted from Wikipedia, comprising 4 million words in each language. We opted for Wikipedia for its cross-linguistic consistency. The process of extracting Wikipedia pages adhered to an algorithm that encompassed similar major topics across both languages. The algorithm initiated the extraction of text from eleven distinct topics shared by both languages and extended to include all text accessible through hyperlinks on the respective pages until it reached the desired word count. We used the lemmatized form of words to determine word frequency, excluding the verb "be" and auxiliary verbs from our analysis to be conservative since the effects described below would be even more pronounced if these verbs had been included. Log frequencies are used in the models described below.

Results

Analysis of the Wikipedia corpus in English and Persian

The distributions of verbs and nouns in the 4 million word samples in each language are shown in Figure 2. The highest-frequency verbs in English are higher frequency than those in Persian and far higher than other verbs in either language. For instance, the cumulative count of the top 20 verbs in English (shaded area) reaches 1,187,341, whereas in Persian, the top 20 verbs accumulate a total of 697,890 counts. On the other hand, the distribution of nouns in the two languages is more similar. The cumulative count of the top 20 nouns is 2,702,463 in English and 3,196,274 in Persian.

Figure 2. The distributions of verbs (A) and nouns (B) in English and Persian.
Analysis of language samples of pwAD and healthy controls across English and Persian

First, we measure the noun-verb ratios in both languages to test for a disproportionate deficit in one or the other word class among pwAD compared to healthy individuals. We used the following formula to obtain the noun-verb ratio in each language sample: total number of nouns/(total number of nouns + total number of verbs).

Results show no significant difference in the ratio of noun-verb use between pwAD and healthy individuals in either language: English (pwAD, $M = 0.54$; healthy $M = 0.57$; $t = 1.53, p = 0.128$); Persian (pwAD, $M = 0.70$; healthy $M = 0.71$; $t = 0.324, p = 0.75$).

There was a significant interaction between the frequencies of verbs used by English and Persian speakers with and without AD ($\beta = -0.638, p = 0.015$), so we compared speakers with and without AD in each language. Consistent with prior literature, English-speaking pwAD favored higher (log) frequency verbs (Mean = 7.03) in comparison to healthy individuals (M = 6.76), ($t = -2.595, p = 0.010$). Critically, the log frequency of verbs used by Persian-speaking pwAD (M = 7.22) was not higher than verbs used by healthy Persian speakers (M = 7.60) ($t = 1.108, p = 0.27$).

We also compared the frequencies of nouns by English and Persian speakers with and without AD. Here we find an interaction showing the opposite pattern: Persian speakers’ production of nouns was more influenced by AD than English speakers’ production of nouns ($\beta = 0.448, p = 0.009$). However, direct comparisons between noun frequencies used by pwAD and healthy individuals did not reach significance in either language: in English speakers: pwAD ($M = 5.13$) and healthy individuals ($M = 5.26$) $t = 1.715, p = 0.088$; in Persian speakers pwAD ($M = 6.19$) and healthy individuals ($M = 5.88$) $t = 1.925, p = 0.060$.

We next sought to evaluate whether the distinction of verb frequency and noun frequency as two separate variables offers an advantage in classifying AD compared to content word frequency as a single variable. For this, we generated two models. The first model consisted of verb frequency and noun frequency as two separate predictor variables to classify AD, while the second model consisted of content word frequency as a single predictor variable. We analyzed the ROC curves for each model, and the AUC was used as a measure of discriminative ability. Results indicated no evidence to claim that the distinction between nouns and verbs offers an advantage over content word frequency in predicting AD. Specifically, we used the DeLong test to statistically compare two AUCs from correlated ROC curves. In the English cohort, the DeLong test yielded a Z-score of 1.312 ($p = 0.190$). The AUC for content word frequency was 0.589, and the AUC for the two-variable model, which consisted of verb frequency and noun frequency, was 0.524. Similarly, in the Persian cohort, the DeLong test yielded a Z-score of 1.876 ($p = 0.060$). The AUC for content word frequency was 0.653, marginally higher than the AUC for the two-variable model consisting of verb frequency and noun frequency, which was 0.498.
**Figure 3.** A comparison of ROC curves for predicting AD based on word frequency versus distinguishing the frequency of nouns and verbs in English (A) and Persian (B) shows no advantage for the separation of word classes.

**Discussion**

In this study, we explored language production in pwAD to address discrepancies between emerging findings on noun and verb processing in this population and the longstanding views on the brain’s segregated processing networks for the two word classes. According to the conventional view, grammatical class information is crucial in language production and comprehension. The formulation of these views has evolved and consolidated in tandem with observations from patients with various forms of aphasia. Particularly noteworthy were findings viewed as a double dissociation between word class and aphasia type. Aphasic patients with temporoparietal damage were found to exhibit more pronounced difficulties in producing nouns than verbs, while patients with Broca’s aphasia with lesions in left frontal regions showed the opposite pattern. These observations have been construed as compelling evidence for the claim that lexical knowledge in the brain is organized by grammatical class.

This conceptualization posits a division within neural networks, with left temporal regions dedicated to noun representation and left frontal areas specialized for verb representation.

However, this view has proven overly simplistic as it fails to account for several emerging observations from patients with various neurological deficits. For example, while nonfluent aphasic patients with left frontal lobe damage exhibit a lower ratio of verbs to nouns in language production, they often produce semantically complex or heavy verbs. The verbs that nonfluent patients produce can be
even heavier than those produced by healthy individuals, indicating that a simple verb class cannot provide a comprehensive account of the linguistic characteristics observed in the disease. Another notable discrepancy is language production in pwAD. The traditional understanding of word processing across separate neural networks predicts that pwAD would struggle more with noun production than verb production, given the predominance of posterior temporal and parietal neurodegeneration and the relative sparing of frontal areas, particularly in the early stages of the disease. Reflecting the prevailing notions, early studies primarily focused on nouns in this patient population. However, accumulating findings challenge this prediction, with some outcomes being completely contrary, such as the one we focus on here. Specifically, English-speaking pwAD use verbs of higher frequencies than controls, and this frequency effect in verb naming has been shown to worsen as the disease progresses.

In this work, we adopt a cross-linguistic approach to test an alternative hypothesis about language changes in pwAD, which is that the fundamental difficulty is in accessing less frequent (therefore more specific and informative) words of any lexical class. This hypothesis posits that word class effects are simply a reflection of the distributional properties of words in a given language. Our investigation suggests that the well-documented overreliance on high-frequency verbs in English-speaking pwAD compared to healthy controls stems from their particularly skewed distribution in English. English features a group of especially high-frequency verbs, such as “take” and “get,” which may emerge as a distinct subset within the broader category of verbs upon cluster analyses based on frequency. These high-frequency verbs possess abstract and flexible meanings as illustrated by phrases like “take a walk,” “take a break,” and “take a bath,” which convey very distinct actions. Conversely, in Persian, light verbs are more commonly integrated into complex verb structures rather than combined with independent noun phrases as in the English examples just mentioned. As predicted, if frequency distinguishes verbs and nouns more strongly in English than in Persian, the observation of increased use of high-frequency verbs by English-speaking pwAD did not hold for Persian-speaking pwAD, whose preference for higher-frequency words did not disproportionately affect verbs. Our analysis confirmed a significant interaction between verb frequencies among English and Persian speakers with and without AD. The interaction undermines the notion of segregated processing for nouns and verbs as the explanatory factor and instead indicates that English-speaking patients’ overreliance on high-frequency verbs stems from English verbs’ distributional properties. We further showed that a regression model that incorporates verb frequency and noun frequency as two distinct predictors offers no advantage over a model with content word frequency as a single predictor in classifying people as AD or controls.

What emerges as a common finding across both languages is the reliance of pwAD on higher-frequency words, with differing consequences dependent on the distributional properties of the language spoken. Basic principles of information theory confirm that higher-frequency words tend to offer less information due to their higher predictability within context. This reliance on high-frequency words seems to be invariant to word class, with word class effects being a byproduct of how nouns and verbs are distributed within a given language. It is well-established that the language of individuals with AD often exhibits vagueness, attributed to the use of uninformative high-frequency words. The direction of causation still remains unclear: Either individuals with AD struggle to access low-frequency words,
resulting in uninformative language, or they experience vague thoughts that hinder the selection of specific and informative words.

The new insight into language abnormalities of AD – that implicates word frequency rather than word class distinction-- is consistent with our recent theoretical framework about nonfluent aphasia, known as the Information Restoration Paradigm.62 This framework also argued that language impairments could be explained in terms of frequency, except that nonfluent aphasic patients favor lower-frequency words over higher-frequency words.47,48,62–64 We analyzed the language of nonfluent patients—with both Broca’s and the nonfluent variant of primary progressive aphasia—who, consistent with existing literature, produce higher proportions of content to function words, heavy verbs to light verbs, and nouns to verbs. Rather than interpreting these findings as isolated symptoms, we identified a common thread: low frequency. Across all three cases of word class dissociation, we observed that patients with nonfluent aphasia consistently opted for the lower-frequency word type. We interpreted this pattern as reflecting the patients’ economic behavior in language production. Confronted with challenges in generating long and complex sentences, patients gravitated towards using low-frequency words as a compensatory strategy since low-frequency words inherently offer more informational value than high-frequency words. Consequently, these patients tend to avoid high-frequency verbs and rely on nouns, along with low-frequency verbs, due to their informational advantage. This pattern of language production in nonfluent aphasia presents a mirror image of the phenomenon observed in AD, where patients struggle with producing informative messages.

Therefore, our interpretation, grounded in word frequency rather than word class distinction, offers a theoretical framework capable of reconciling the two incongruent findings present in the literature on the pattern of word use in AD and nonfluent aphasia. This interpretation is consistent with views that consider nouns and verbs to be processed through the same shared neural networks.65,66 These views, initially introduced by Sapir (1921)1 and later refined through functionalist approaches to language processing,67 suggest that grammatical class emerges as a consequence of distribution and semantics.66 Even small recurrent networks can predict the grammatical properties of words and categorize them based on their grammatical class in a probabilistic way on the basis of their distribution.68 In this framework, lexical categories include prototypical and less prototypical exemplars which self-organize based on distribution, semantics, and pragmatics.66,69–71

Future studies are needed to investigate diverse languages with varying distributional properties within their lexicons to better understand the root psycholinguistic causes of impairments in pwAD. For instance, unlike findings in which English-speaking pwAD exhibited better performance in single-word production of nouns compared to verbs, a study of Chinese speakers with AD showed similar performance across both nouns and verbs.72 Adopting cross-linguistic approaches is crucial to avoid presumptions about certain linguistic features being language universals or biologically determined when, in reality, they may be contingent upon the properties of specific languages. The comparative methodology further enhances our understanding of the nuances in diverse languages and, critically, allows for new insights into the causes of impaired language in different groups of people.
Availability of data and materials

The English dataset was obtained from the DementiaBank repository. The Persian dataset obtained for this study, in addition to the computational codes in Python, can be accessed upon contacting the senior author at nrezaii@mgh.harvard.edu

Competing interests

The authors declare no competing interests to declare.

Funding

This work was supported by Alzheimer's Association Clinician Scientist Fellowship (AACSF) 2022A015154 and MGH Screening Technologies in Primary Care Innovation Fund (PCIF) 2023A063002, as well as National Institutes of Health grants R21 DC019567, R21 AG073744, and R01 NS131395.

Authors’ contributions

SB, MS, MMP, AK, MG, SR, ZM, and MST participated in the data collection. SB, MS, MMP, AK, MG, SR, and MAD contributed to patients' clinical and neuropsychological assessment. NR and AG conceptualized the idea and were major contributors to writing the manuscript. NR and AG analyzed data. NR, AG, and BCD edited the manuscript. All authors contributed to the initial writing and read and approved the final manuscript.

Acknowledgment

We thank Harvard Catalyst, The Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health Award UL1 TR002541) for its biostatistician consultation service. We also thank Mahan Rezaei for transcriptional services in Persian.
References


29. Karimi-Doostan, G. Light verbs and Structural case.


63. Rezaii, N., Ren, B., Quimby, M., Hochberg, D. & Dickerson, B. Less is more in language production: Shorter sentences contain more informative words. 2022.06.02.22275938 Preprint at https://doi.org/10.1101/2022.06.02.22275938 (2022).


