Neuropsychologia 234 (2027) 109607

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Testing different metrics for syntactic and morphological complexity in aphasic discourse: A cross-genre comparison in Spanish

Jorge Ricardo Hidalgo Chagoya [a,b,c,d,*,1], Xabi Ansorena [e], Roel Jonkers [a,b,c,f], Maddi Carrera [e], José Ignacio Quemada Ubis [g], Simona Mancini [e,h,**]

[a] Erasmus Mundus Program of European Master's in Clinical Linguistics, University of Groningen, the Netherlands
[b] Erasmus Mundus Program of European Master's in Clinical Linguistics, Ghent University, Belgium
[c] Erasmus Mundus Program of European Master's in Clinical Linguistics, University of Eastern Finland, Finland
[d] Department of Linguistics, Cognitive Science and Semiotics, Aarhus University, Denmark
[e] Neurolinguistics and Aphasia Group, Basque Center on Cognition, Brain and Language, Spain
[f] Center for Language and Cognition Groningen, University of Groningen, the Netherlands
[g] Brain Damage Unit, Hermanas Hospitalarias Aita Menni, Spain
[h] Ikerbasque, Basque Foundation for Science, Spain

ARTICLE INFO

Keywords:
Discourse
Syntax
Morphology
Complexity
Aphasia
Syntactic complexity
Morphological complexity
Spanish

ABSTRACT

Purpose: The aim of this study is to determine which of four tasks testing different discourse genres (descriptive, procedural, conversational, and narrative) is the most sensitive in detecting syntactic and morphological complexity deficits in Spanish aphasia. Additionally, we compare two approaches for measuring syntax in discourse to identify the most accurate set of metrics for discriminating group membership.
Method: In this comparative cross-sectional study, we analyzed the discourse of an aphasia group (N = 20) and a non-brain-damaged group (N = 14). Data was preprocessed using manual and automated procedures. We measured syntactic complexity through proxies (mean length of utterance, noun-verb ratio, and verbs per utterance), in addition to a competing 'syntactic structures' metric. The latter includes variables of subordination, left branching, relative clauses, and word integration, while morphological complexity was measured through an inflectional index.
Results: The descriptive and narrative tasks highlighted the most significant differences across syntactic complexity metrics between the groups. Additionally, nested binomial logistic regression models showed that the 'syntactic structures' metric predicted groups better than the proxies alone.
Conclusions: A comprehensive assessment of syntactic complexity can be obtained through the administration of both descriptive and narrative tasks. The measurement of 'syntactic structures' in discourse proved a fitter and finer-grained approach for group discrimination than the incorporation of syntactic proxies alone. These findings provide guidance for a precise discourse assessment in Spanish aphasia.

1. Introduction

Syntax and morphology — the rules governing sentence structure and word formation and inflection, respectively — are key language components that are often impaired in aphasia following damage to fronto-temporal regions of the left hemisphere and the white-matter fibers that connect them. Syntactic complexity involves the use of recursive structures through clausal subordination and embedding (Hunt, 1970), while morphological complexity concerns how morphemes combine within words as well as how words relate to one another through inflection and derivation (Stump and Finkel, 2015). Deficits in both dimensions typically manifest in language use above the sentence level, that is, in discourse production (Armstrong, 2000).

Discourse, or connected speech, emerges from the interaction of phonological, lexico-semantic, and morphosyntactic processes (Prins and Bastiaanse, 2004), and is further shaped by pragmatic, cognitive

* Corresponding author. Erasmus Mundus Program of European Master's in Clinical Linguistics, University of Groningen, the Netherlands.
** Corresponding author. Neurolinguistics and Aphasia Group Basque Center on Cognition, Brain and Language, Spain.
E-mail addresses: hidalgo@cc.au.dk (J.R. Hidalgo Chagoya), s.mancini@bcbl.eu (S. Mancini).
[1] Present address: Department of Linguistics, Cognitive Science and Semiotics, Aarhus University (DK).

https://doi.org/10.1016/j.neuropsychologia.2026.109607
Received 4 April 2025; Received in revised form 11 September 2026; Accepted 8 October 2026
Available online 9 October 2026
0028-3932/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

and social factors (Linnik et al., 2015; Stark and Fukuyama, 2020; see Frau et al., 2025 for a review on pragmatic impairments).[2] Given this complexity, discourse assessment is essential for both aphasia evaluation and research (Bryant et al., 2016). Crucially, discourse analysis can uncover deficits that more constrained language tasks tend to miss (Beeke et al., 2003; Kemper et al., 2006; Marini et al., 2011), owing to its reliance on naturalistic and ecologically valid linguistic samples (Davidson et al., 2003; Doedens and Meteyard, 2020; World Health Organization, 2018). As components of the micro-linguistic dimension of discourse analysis (Glosser and Deser, 1991; Kintsch, 1994), syntax and morphology are a fundamental aspect of the evaluation of connected speech.

Despite its potential to inform the assessment of post-stroke syntactic deficits, cross-study comparison of syntactic complexity measures is challenging due to varying metrics and elicitation methods (Linnik et al., 2015; Stark et al., 2021a). Moreover, research on Spanish morphological deficits in aphasia discourse is particularly limited (Martínez-Ferreiro et al., 2017). To fill these gaps and guide researchers and clinicians toward a more efficient assessment of morphosyntactic deficits in aphasic discourse, this study investigates the simplification of syntactic and morphological complexity associated with aphasia in Spanish speakers, employing two metric approaches across different discourse genres to identify the most sensitive tasks and metrics.

1.1. Measuring syntactic complexity in aphasic discourse

Simplification of syntactic complexity, mainly manifested in a reduction of clause length and lower frequency of embedded clauses (Ulatowska et al., 1990), is a hallmark of impaired speech in speakers with aphasia (Schnur and Wang, 2024), especially those presenting with agrammatism, a cluster of language symptoms characterized by dysfunction in the production and comprehension of grammatical morphology and sentence structure (Menn et al., 1990). While syntactic simplification has been a common discourse marker in aphasia research, comparing results proves challenging given that discourse metrics measuring syntactic complexity are highly variable as well (Armstrong, 2000; Linnik et al., 2015). Two main approaches for measuring syntactic complexity can be discerned in aphasia studies. The most traditional approach requires parsing each separate utterance, counting of embeddings, and considering the subordination characteristics. Examples akin to direct measures of syntactic complexity include Glosser and Deser's (1991) 'weighted index of subordination', Sherratt's examination (2007) of subordinate clause frequency, clauses length, and the percentages of left branching, right branching, and main clauses; Wright and Capilouto's (2012) 'complexity index', and the automated extraction of 'syntactic accuracy' and 'structural complexity' scores reported by Schnur and Wang (2024). Evaluating the frequency and type of subordination and embeddings provides a direct index of the integrity of speakers' syntactic representations and mechanisms, at the cost of complex and time-consuming discourse processing steps.

Another possible approach to measuring syntactic complexity is implementing syntactic proxies, which involves adopting indirect measures to represent the complexity of a speakers' syntax. For instance, Vermeulen and colleagues (1989) described syntactic simplification in aphasic speech through reduced use of function words and shorter Mean Length of Utterance (MLU). Stark (2019) reported reduced proportion of verbs per utterance, in addition to an augmented noun-verb ratio and open-closed word class ratio. The usage of syntactic proxies has the advantages of saving time and effort, since they allow for the automated extraction of such variables using discourse analysis programs. Yet, metric analysis (e.g., the proportion of verbs per utterance or MLU) can only indirectly inform on finer-grained syntactic aspects, such as the presence of subordination and embeddings.

Relevant for this research, Agmon and colleagues (2024) proposed the 'syntactic structures' multidimensional metric, which consisted of cognitively demanding grammatical variables that are sensitive to aging or dementia. Such metric considers subordination, a homologue to previously used embedding indexes (Bastiaanse et al., 1996a; Gordon, 2020; Schnur and Wang, 2024; Sherratt, 2007; Ulatowska et al., 1990; Wright and Capilouto, 2012); Relative clauses, i.e. a form of subordination where a dependent clause modifies a noun in the independent clause; Left-branching clauses, which refers to a special form of subordination where the embedded clause is placed before or within the main clause; and complex noun phrases and complex adjectival modifications, both representing intraclausal combinatorial syntactic phenomena (see Table 2 for illustrative examples). Agmon et al. (2024) evaluated various syntactic complexity metrics, including the 'syntactic structures' model, and found that it outperformed other measures—e.g. sentence length, the Yngve score (Yngve, 1960), and syntactic frequency (Rezaii et al., 2022)—in detecting discourse differences between young adults and healthy elderlies. They observed that, compared to the younger group, the older group produced fewer subordinate clauses, relative clauses, and center embeddings per utterance.

While the measures described above have proved to be sensitive to the presence of syntactic simplification and agrammatism in aphasic speakers, no study has yet contrasted the relative accuracy, sensitivity and specificity of syntactic proxies and hierarchical syntactic metrics with this population.

1.2. Measuring morphological complexity in aphasic speech

Morphological processes are contingent upon the inflectional and derivational systems inherent to a specific language (Stump and Finkel, 2015). Particularly, inflectional morphology has garnered significant attention in speech and language pathology research, due to the challenges individuals with aphasia face, and particularly those with agrammatism, in producing finite lexical verbs (Bastiaanse et al., 1996a, 1996b, 2002; Martínez-Ferreiro et al., 2017).

While the majority of discourse studies have been conducted in English-speaking countries, the English inflectional paradigm is often regarded as simplistic (Beniamine, 2021) due to its limited number of inflected verb forms. Probably because of such a reduced inflectional paradigm, morphological errors are rare even in aphasic discourse samples (on average < 0.57 errors made per task, as reported in Stark and Fukuyama, 2020). Consequently, morphological complexity as a discourse measure has been largely overlooked (Bryant et al., 2016; Stark et al., 2021a). Notable exceptions in discourse research include using an inflectional index in Dutch (Bastiaanse et al., 1996b), inflection proportion for nouns and verbs in English (Gordon, 2020), morphosyntactic accuracy in French (Sahraoui and Nespoulous, 2012), and finiteness proportion in Spanish (Martínez-Ferreiro et al., 2017).

1.3. Discourse genre sensitivities to syntactic and morphological complexity

An early focus of aphasia discourse studies was on the differential effects that elicitation genres had on linguistic measures. Since the 1972 introduction of the Cookie Theft picture description task in the Boston Diagnostic Aphasia Exam (BDAE; Goodglass and Kaplan, 1972), different discourse elicitation methods have proliferated in both research and clinical assessment. This has spurred increasing interest in whether discourse and associated metrics vary by task, and whether tasks can be meaningfully grouped into broader categories, such as

[2] Discourse is a rich communicative production, which renders it useful for assessing expressive pragmatic impairments regarding coherence, cohesion, informativeness, and appropriateness of information. While relevant for a fine-grained, exhaustive and clinically relevant analysis of spontaneous speech, these factors pertain to the macrolinguistics dimension of discourse (see an introduction to the topic in Garraffa and Smith, 2024), which are beyond the scope of the present study.

discourse genres, based on the patterns they elicit, with potential clinical implications for the selection of assessment tools. Ulatowska and colleagues (1990) were among the first to demonstrate that different discourse genres tax the cognitive system distinctively and consequently elicit content that differs in information and structure (see also Doyle et al., 1995; Olness, 2006; Van Leer and Turkstra, 1999 for similar results). The literature has converged on four core elicitation genres (Ulatowska et al., 1990; for a review, Bryant et al., 2016): expositional/descriptive, procedural/script, narrative, and conversational. In the expositional genre, subjects describe a topic — usually prompted by a single image or picture sequence — whereas the procedural genre elicits explanations of how a particular activity is carried out. The narrative genre involves recounting a factual or fictional story with characters, a setting, and a sequence of events, while the conversational genre entails an interactive dialogue with another speaker. These four genres have been systematically operationalized in AphasiaBank (MacWhinney et al., 2011; MacWhinney, 2019), a computerized database of clinician–patient interviews built around a tightly specified elicitation protocol in which investigators follow a structured script for questioning and narrative elicitation. A key advantage of AphasiaBank is its integration with the CHAT and CLAN programs, which allow a wide range of syntax-related indices to be computed automatically and accurately, facilitating cross-genre comparisons.

Recently, leveraging large-scale data from AphasiaBank and its computational resources, Stark (2019) systematically showed that expositional, narrative and procedural discourse each elicits unique linguistic features in both PWA and non-brain-damaged individuals. Stark and Fukuyama (2020) further highlighted that linguistic production is influenced by task type, aphasia type, and severity, with tasks in the same genre differing significantly from each other, while maintaining comparable linguistic features when compared to other genres. Other researchers (Coelho, 2002; Fergadiotis and Wright, 2011; Stark and Fukuyama, 2020; Wright and Capilouto, 2009) have explored how specific task properties shape linguistic output beyond discourse genre.

Genre effects on discourse have been explained by top-down adjustments in the frame generation phase (Sherratt, 2007) of a multi-layered discourse model (Frederiksen et al., 1990). In Sherratt (2007) model, conceptual frames generated at the early stages of discourse production reflect specific genres and provide the organizing grammar that eventually constrains the selection of syntactic structure during the stage of linguistic formulation. Hence, each genre activates a culturally generic frame structure (Mandler, 1987), while tasks within the same genre activate similar structures, resulting in shared downstream linguistic selection (Stark et al., 2021a).

Comparative studies mostly contrast descriptive and narrative genres (Fergadiotis and Wright, 2011; Olness, 2006; Schnur and Wang, 2024; Ulatowska et al., 1990; Van Leer and Turkstra, 1999). These studies reveal that picture description tasks often simplify syntax, typically eliciting declarative present-tense statements, deictic references, and limited verb production (Garrard and Forsyth, 2010; March et al., 2006; Stark, 2019), which contrasts with the syntactically richer production (e.g., more subordinate clauses) of narrative tasks. Additionally, picture descriptions may not reflect all aspects of real-world discourse and require minimal cognitive organization (Cannizzaro and Coelho, 2013; Olness, 2006). For example, in their study, Schnur and Wang (2024) compared the AphasiaBank's Cinderella story retell task (Goodglass and Kaplan, 1972) against the Picnic Scene description task from the Western Aphasia Battery (Kertesz, 2007) in a group of acute post-stroke participants, in order to determine which elicitation method more effectively captured language capabilities. They concluded that the storytelling task was more sensitive for assessing structural complexity in the discourse of this population.

In contrast, narrative tasks demand greater syntactic complexity, requiring reference to settings, event sequences, and character motives (Stark, 2019). Narratives have been shown to elicit more syntactic errors and complex structures than procedural tasks (Ulatowska et al., 1990), with storytelling being particularly useful for evaluating post-intervention syntactic improvements in PWA (Stark, 2019; Schnur and Wang, 2024). Previous research has concluded that the narrative genre typically involves a blend of past and present tense usage, with a lower frequency of nouns compared to other discourse types (Olness, 2006; Stark and Fukuyama, 2020).

Procedural (Stark, 2019; Ulatowska et al., 1981) and conversational genres seem overall inadequate for assessing syntactic complexity. Although less researched, conversational tasks often result in short, simple utterances with few errors, less conventional grammar, and limited syntactic richness (Beeke et al., 2003; Armstrong et al., 2011). However, high variability in conversational outcomes may be influenced by factors such as aphasia severity and conversational context.

While many aphasia studies have compared outcome measures across two discourse genres (Armstrong et al., 2011; Fergadiotis and Wright, 2011; Leaman and Edmonds, 2023; Olness, 2006; Pritchard et al., 2015; Schnur and Wang, 2024; Ulatowska et al., 1981, 1990; Van Leer and Turkstra, 1999) or all three monologic genres (Stark, 2019; Stark and Fukuyama, 2020), there has not been a single study to our knowledge that encompasses all four identified elicitation genres. Given the task-specific nature of discourse and the heterogeneity of aphasia, a comprehensive study comparing multiple genres within a unified protocol and a same sample is needed. Additionally, employing a direct measure of syntactic complexity in comparison with syntactic proxies, as proposed by Agmon et al. (2024), would provide further insights into the sensitivity of linguistic metrics. Finally, it is essential to study syntactic and morphological complexity in an inflectionally rich language like Spanish (Real Academia Española de la Lengua, 2009), where morphological errors due to inflection challenges (Martínez-Ferreiro et al., 2017) can be more prevalent than in a morphologically poor language like English. To our knowledge, only one study has focused on Spanish post-stroke aphasic discourse (Martínez-Ferreiro et al., 2017), thus a thorough investigation of how morphological and syntactic simplification manifest across discourse tasks bears significant clinical and research relevance.

1.4. The Spanish morpho-syntactic system

Spanish is an inflected and synthetic Indo-European language with a rich morphosyntactic system (Martínez-Ferreiro et al., 2017). In the nominal domain, nouns are inflected for number and gender, and all modifiers within the noun phrase — articles, demonstratives, quantifiers, and adjectives — must agree with the head noun in both features (Real Academia Española de la Lengua, 2009). Grammatical gender is arbitrarily assigned to nouns regardless of real-world referents (*mesa*$_{FEM}$/table, *huerto*$_{MASC}$/garden), though in animate nouns it frequently coincides with the sex of the referent (*hermano*, brother; *hermana*, sister), where the *-a* or *-o* morpheme often, but not always, reflects information about the referent's sex.[3] Both attributive and predicative adjectives canonically follow the head noun (*la casa roja*, 'the red house', *la niña está callada*, 'the girl is quiet').

In the verbal domain, Spanish presents moderate complexity, encompassing 16 complete paradigms: 10 tenses in the indicative mood and 6 in the subjunctive, plus the imperative (Martínez-Ferreiro et al., 2017). Simple finite verbs are built from the verbal root and thematic vowel (*-a, -e, -i*), to which markers of tense, aspect, mood, person, and number are suffixed (Hualde et al., 2012). Aspect is morphologically distinguished in the past tense, which contrasts the preterit and the

[3] As observed by an anonymous reviewer, the correspondence between grammatical gender and the sex of the referent is not systematic. Cases such as *un persona*$_{FEM}$ (a person) referring to a man, or *un puma hembra*$_{MASC}$ (a female puma), illustrate that the sex of the referent does not override the noun's grammatical gender.

imperfect, and periphrastically in the present and future tenses (Real Academia Española de la Lengua, 2009). This articulate inflection paradigm renders subject pronouns pragmatically optional, classifying Spanish as a pro-drop language.

Verb phrases are structured around a lexical verb that may be accompanied by auxiliaries and clitic pronouns (El líbro lo_{clitic} leí ayer, 'The book, (I) read it yesterday'). The auxiliary *haber* constructs compound tenses followed by the past participle (*he comido*, '(I) have eaten'), while transitive verbs allow passive constructions formed with *ser* plus the past participle. Spanish also offers several periphrastic constructions for temporal reference, most productively the present progressive (*estar* + gerund) and future-referring periphrases (*ir a* + infinitive). Temporal adverbs such as *ayer* ('yesterday') and *mañana* ('tomorrow') must be compatible with the verb form used; any mismatch results in ungrammaticality (**ayer vendremos*, 'yesterday we will come').

As for subordination, in Spanish complement clauses and relative clauses are syntactically fixed in post-head position (*Creo que Juan llegó tarde*, 'I think that Juan arrived late'; *el libro que compré ayer*, 'the book that I bought yesterday') (Real Academia Española de la Lengua, 2009). Adverbial clauses, by contrast, are positionally flexible: while they most frequently appear post-verbally (*Me quedé en casa porque llovía*, 'I stayed home because it was raining'), they can be fronted for discourse-pragmatic reasons (*Si llueve, me quedo en casa*, 'If it rains, I'll stay home').

1.5. Morphological and syntactic complexity in Spanish aphasia

Investigations on the morphological complexity of aphasic discourse in Spanish are very scarce. Using semi-structured interviews, Martínez-Ferreiro and colleagues (2017) investigated the production of verbal inflection in Spanish and reported a reduced rate of finite verbs produced by non-fluent people with mixed aphasia (PWA) compared to fluent PWA and the control group. Greater information on finite verb production can be gathered from studies focusing on agrammatic production with structured tasks (e.g., delayed repetition, sentence completion, morphological transformation), which have generally reported selective impairment for tense compared to subject-verb agreement markers (person and number) (Benedet et al., 1998; Diéguez-Vide et al., 2012; Gavarró and Martínez-Ferreiro, 2007; Martínez-Ferreiro, 2003, 2010; Martínez-Ferreiro and Bastiaanse, 2013; Moreno-Torres Sánchez et al., 2005; Stecher et al., 2024; for a review, see Martínez-Ferreiro et al., 2015). This dissociation has been linked to reduced processing resources in PWA compared to brain-undamaged participants (Burchert et al., 2005; for a review, see Druks, 2016; Garraffa and Fyndanis, 2020).

As for nominal inflection, findings from a completion task in Spanish have shown that gender markings are frequently affected in aphasic production, with opaque conditions[4] often resulting in more errors (Piazza et al., 2022).[5] To our knowledge, no study has so far investigated the complexity of nominal inflection in Spanish using discourse tasks.

Available evidence on the syntactic complexity of speech in Spanish aphasia is also scarce. Hernández-Sacristán and Rosell-Clarí (2009) investigated how a group of PWA utilized the Spanish and Catalan "*que*" (that) during natural, undirected conversation. The results revealed not only that PWA used a restricted and simpler range of syntactic configurations compared to healthy controls (e.g., fewer relative clauses), but also that their speech was remarkably free of strictly ungrammatical errors. Similarly, Martínez-Ferreiro et al. (2017) reported that the production of subordinate clauses by fluent and non-fluent PWA in semi-structured interviews was significantly lower compared to a group of unimpaired speakers, although the study does not indicate the type of subordinate clauses produced. This line of evidence converges with findings from sentence elicitation tasks showing that the production of relative clauses (along with wh-questions and object clitics) is significantly impaired in Spanish agrammatism (Martínez-Ferreiro, 2010).

Taken together, these findings from studies testing agrammatic and mixed-aphasia speakers suggest that syntactic and morphological complexity are reduced in Spanish aphasic discourse, with structured task settings confirming significant impairment in terms of inflection and subordinate clause production. However, predominant reliance on conversational genres (semi-structured interviews and undirected conversation) leaves open the question of whether these syntactic and morphological complexity patterns hold across different discourse genres such as narrative retelling or procedural description, which are known to place distinct demands on speakers' linguistic and cognitive resources.

2. The current study

The current study aims to investigate how discourse tasks of different elicitation genres capture syntactic simplification and morphological deficits associated with aphasia in Spanish speakers. Our goals and approach are motivated by the clinical and research relevance of ecologically valid methods for the evaluation of clinical populations (Davidson et al., 2003; World Health Organization, 2018). As reviewed above, spoken language can be elicited through a variety of discourse methods, each placing different demands on speakers that, in turn, can give rise to task-specific degrees of morphological and syntactic complexity in the output. In a similar vein, a range of syntactic complexity measures can be extracted from spontaneous speech, with their relative frequency shaped by the type of task used, presumably due to top-down adjustments in the task-specific conceptual frames generated during the early stages of discourse production (Sherratt, 2007). Comparing outcomes across discourse genres and identifying which metrics and tasks best capture syntactic complexity in Spanish can support the adoption of ecologically valid methods in both research and clinical settings. Critically, by examining discourse production in Spanish, our study also aims to enrich the field of aphasiology by contributing data from an underrepresented language.

Hence, we attempt to answer three main questions: (1) Which discourse task is more sensitive to the detection of syntactic complexity simplification in the discourse of Spanish PWA? (2) Which discourse task is more sensitive to the detection of morphological complexity deficits? (3) Are the 'syntactic structures' metrics (Agmon et al., 2024) a better fit for assessing aphasic features in connected speech than the syntactic proxies? Identifying the most suitable metrics and elicitation tasks to detect syntactic and morphological impairment has the potential to inform therapeutic programs (Leaman and Edmonds, 2023) and supports the crosslinguistic goal of disentangling language-specific impairments in Spanish aphasia (Linnik et al., 2015).

When compared with the expositional and procedural discourse, the narrative genre is frequently identified as the most syntactically and morphologically complex, as reflected in the proportion of embeddings (Schnur and Wang, 2024; Ulatowska et al., 1981, 1990), syntactic complexity proxies (Stark, 2019; Stark and Fukuyama, 2020), and the usage of both past and present tense (Stark and Fukuyama, 2020). Although comparison between narrative and conversational genres are uncommon, the dialogic nature of conversational discourse allows for less conventional grammar to emerge in produced utterances (Armstrong et al., 2011; Beeke et al., 2003; Sajjadi et al., 2012). Thus, we hypothesize that a narrative task will be the most sensitive in

[4] Gender opacity refers to nouns whose endings do not provide a reliable cue to grammatical gender, as opposed to 'transparent' nouns whose endings (typically -o for masculine and -a for feminine) are strongly associated with a particular gender; Sá-Leite et al. (2024).

[5] The cited study differentiates grammatical and semantic gender. As observed by an anonymous reviewer, this differentiation is problematic from a formal linguistic perspective, although it has proven useful in psycholinguistics and aphasiology for distinguishing between gender processing operations with distinct cognitive demands.

detecting syntactic complexity deficits. Considering the possibility of eliciting more verbs with person and tense inflections, (Olness, 2006; Stark and Fukuyama, 2020), we hypothesize that the narrative task will also be the most sensitive in detecting morphological complexity deficits. Additionally, we hypothesize that 'syntactic structures' (Agmon et al., 2024) will be a better metric for measuring the syntactic simplification of aphasic discourse than syntactic proxies, given the former's broader and more comprehensive scope.

3. Method

3.1. Participants

A sample of 20 individuals diagnosed with aphasia by a certified speech-language pathologist and 14 non-brain damaged individuals were employed (see Table 1). All participants were born in Spain and claimed Spanish as their native language, except for two participants who reported Basque as their first language but learned Spanish before the age of 6 years (refer to Supplementary Tables A and B for the individual demographic data of the participants in the aphasia and control group, respectively). Every participant in the aphasia group was diagnosed with aphasia after a cerebrovascular accident, with the exception of P04 (traumatic brain injury) and P14 (herpes meningoencephalitis). They were referred for assessment by clinicians at the Aita Menni Rehabilitation Centers in Bilbao and Donostia-San Sebastián or were otherwise recruited via diffusion efforts. Participants in the control group were individuals listed in the Basque Center for Cognition, Brain and Language (BCBL) database that matched the aphasia group in age, sex and education. This study was approved by the BCBL Ethics Review Board (13022018M) and the Comité de Ética de Investigación de Euskadi (PI2018033).

3.2. Procedure

Before the discourse elicitation tasks, a neuropsychological evaluation (Appendix and Supplementary Tables S3 and S4; Peña-Casanova et al., 2009; Smith Watts et al., 2019; Benton, 1991), along with the administration of selected tasks from the Spanish BETA battery (Cuetos and González-Nosti, 2009) was carried out (Supplementary Tables S1 and S2). Following, discourse elicitation tasks were administered digitally in a Samsung Galaxy S4 Tablet. Data collection was carried out as part of the ongoing validation of a digital language assessment tool (Ansorena et al., 2022), and was conducted by trained neuropsychologists (X.A., M.C.) and one graduate student (J.H.C.) working at the BCBL. Assessment took place in either the behavioral cabins of BCBL or in hospital/rehabilitation facilities.

3.3. Stimuli

All discourse task instructions were presented both visually by text and audibly through a pre-recorded feminine voice and were repeated by the examiner whenever unclear. Participant responses were automatically recorded within the digital platform, with an external recorder as a backup. Building on AphasiaBank elicitation protocol (MacWhinney et al., 2011), participants engaged in four different elicitation genres consisting of the Cookie Theft picture description (Goodglass and Kaplan, 1972), and three newly developed tasks in Spanish (Ansorena et al., 2022)—the Ham and Cheese Sandwich procedural task, a Structured Interview, and the Hare and the Tortoise story retelling. The discourse tasks were consistently applied in the order described above. All four discourse elicitation tasks began with an instruction slide and prompted the participant to start their response with a beep sound.

Additionally, all four tasks were applied during the same assessment session. Each participant responded to every discourse task except participant 15, who skipped the Hare and the Tortoise story retelling (total discourse samples analyzed = 135, missing data = 1). The duration of the discourse output was not restricted in time.

3.3.1. Cookie Theft picture description—descriptive task

The Cookie Theft picture description (Goodglass and Kaplan, 1972) task is the most commonly collected spoken discourse task for aphasia (Stark et al., 2021a). As such, this task was included in the data collection protocol as the gold standard for discourse assessment, but it is not part of the newly-developed digital battery. For its application, the task requires participants to look at a single black and white picture and

Table 1
Comparison of demographic characteristics between the aphasia and control groups.

Group	Sex	Age (in years)	Education (in years)	TPO (in months)
		M (*SD*)	*M* (*SD*)	*M* (*SD*)
		Min–max	Min–max	Min–max
Aphasia	5 F	58.1 (11)	18.9 (3.96)	12.0 (7.2)
$n = 20$	15 M	30–73	10–24	4–36
Control	4 F	55.9 (10.6)	18.9 (8.4)	-
$n = 14$	10 M	29–69	8–27	
	$\chi 2 = 0.05$, $p = 0.81$	$t(32) = 0.57$, $p = 0.58$	$t(31) = 1.30$, $p = 0.20$	-

Note: F = female; M = male; TPO = time post onset.

Table 2
Examples of the measured features as produced by participants with aphasia.

Microlinguistic feature	Participant	Examples as tagged in the main tier of the transcripts
Subordinate clause	21	"&-pues *una vez una tortuga le retó an una liebre de que* &-eh *era capaz* &-eh *de ganarle corriendo* [% ad:7] [% sb]." "Once a tortoise challenged a hare that it could outrun it."
Relative clause	23	"*es* [//] *está el* [/] *el* [/] *el nieto de la casa que se cae del* [/] *del* [/] *del* &+ta *taburete* [% cn:10] [% sb] [% rl:6]." "There is the grandson of the house that is falling from the stool."
Left branching clause	08	&-pues *por las mañanas* [% ad:3], *si viene mi hermano* [% ad:4] [% sb] [% lb:3], *suelo ir a dar una vuelta por el monte* [% ad:3]. "In the mornings, if my brother comes, I usually take a stroll through the hills."
Complex noun phrase	03	"&-a_ver (…) *es el cuento de* [/] *de la liebre y la tortuga* [% cn:8]." "It is the story of the hare and the tortoise."
Complex adjectival/ adverbial modification	18	"*los hermanos están queriendo coger unas galletas de lo alto de un armario* [% ad:6]." "The siblings want to grab some cookies from the top of a cupboard."
Morphological error	19	"*luego la liebre* &-eh *se duerme tranquilamente, porque sé* [*m:p] *que la tortuga está muy lejos* [% ad:9] [% sb]." "Then the hare falls asleep calmly, because I know that the turtle is far away."
Ungrammatical utterance	01	"*los* &-eh *amigos* &-eh *a ir an un leportivo* [: deportivo] [*p:n] *xxx*. [+gra]." "The friends to go to a sports centre."

Note: &- = filled pause; [//] = word revision; [/] = repetitions; &+ = phonological fragment; [*m:p] = morphological error of person-number inflection; [*p:n] = phonological error that leads to a non-word; xxx = unintelligible words; and [+gra] = grammatical error at the utterance level. The above-mentioned transcription conventions are in line with the CHAT guidelines (MacWhinney, 2000), while the 'syntactic structures' measures (Agmon et al., 2024) are tagged in the form of comments ([%]) as specified in the extraction of variables subsection.

to describe what they see.

3.3.2. Ham and Cheese Sandwich—procedural task

The Ham and Cheese Sandwich task was adapted from the Peanut Butter and Jelly task, a widely used procedure developed within the AphasiaBank protocol. Particularly suited for English-speaking populations, the Peanut Butter and Jelly task requires participants to describe the steps involved in making a peanut butter and jelly sandwich (MacWhinney et al., 2011; Stark et al., 2021b). Our modified version, requiring participants to describe how they make a ham and cheese sandwich, was a cultural adaptation to dietary preferences in Spain. Participants described the process while looking at a blank screen with a central cross.

3.3.3. Structured interview—conversational task

The Structured Interview is one of the most employed dialogic tasks to assess conversation (Stark et al., 2021b). The interview script was composed of five consecutive pre-recorded questions that were presented acoustically, while a blank screen with a central cross was visible on the tablet: (1) "What did you do during the last holidays?" ["*¿qué hizo durante sus últimas vacaciones?*"]; (2) "What did you do last weekend?" ["*¿qué hizo el último fin de semana?*"]; (3) "What are you going to do for Christmas?" ["*¿qué va a hacer en Navidades?*"]; (4) "What are you doing this afternoon?" ["*¿qué va a hacer esta tarde?*"]; and (5) "What do you usually do in the mornings?" ["*¿qué hace normalmente por las mañanas?*"]. The course of questions was meant to elicit different tenses, while employing everyday topics. The examiner offered minimal input through verbal and nonverbal cues as needed, offering clarifications when necessary.

3.3.4. The Tortoise and the Hare story retell—narrative task

This sequence of pictures elicits discourse through requiring the participant to tell a well-known fable (Fig. 1). The Tortoise and the Hare fable was selected for its inclusion in the neurolinguistic baattery (Ansorena et al., 2022) since it is widely popular across Spanish-speaking countries. Our narrative task resembled what discourse elicitation protocols such as AphasiaBank and previous studies have called "sequential picture description" (MacWhinney et al., 2011). While the differences in the amount of visual support may confound the comparison of results between the Hare and the Tortoise Story retell used here and the Cinderella story Retell used in other studies, they allow for a clean comparison between the single-picture description and the narrative task used in this study, making it possible to identify the effects on speech of narrating a story, compared to describing a picture without the confounding effects of semantic memory load (Bliss and McCabe, 2006).[6]

3.4. Data preprocessing and extraction of variables

Discourse samples were transcribed and segmented in utterances according to AphasiaBank's CHAT's conventions and coding regulations (MacWhinney, 2000). For the sake of transparency and replicability, the syntactic structures of interest were coded between brackets in the main tier of the transcription. Variables were extracted using the CLAN program available in AphasiaBank (version c for macOS, downloaded on January 12, 2024; MacWhinney, 2000). Unrecognized words were manually added to the lexical libraries. Initially, the *mor* command was used, but issues with the parser and tagger led to the use of the *morphotag* function from the Batchalign Project (Liu et al., 2023).[7]

The computation of syntactic proxies was fully automated using CLAN (see Supplementary Table S5 for detailed information on the definition of the different syntactic proxies metrics). MLU (*mlu* + tPAR), was measured in morphemes. The noun-verb ratio was calculated by determining the frequency of verbs (freq + sm;,|verb + tPAR) and nouns (freq + sm;,|noun + tPAR), and then dividing the noun count by the verb count (it included all verbs constituents in verbal periphrases, following Olbertz, 1998) . Likewise, the rate of verbs per utterance was obtained by dividing the verb frequency by the number of utterances.

The 'syntactic structures' multidimensional metric (see Table 2 for illustrative examples of the measured features produced by PWAs), based on Agmon et al. (2024), was obtained through manual parsing and tagging, followed by automated extraction (*freq + s"\[\% *\]" + r6 + t*PAR*). This metric comprised seven variables, all averaged by the number of utterances: *subordination*, an embedding index calculated by dividing the total number of clauses (subordinate clauses + total utterances) by the number of utterances; *relative clauses*, determined by dividing the count of relative clauses by the number of utterances; *dependency distance in relative clauses*, defined as the interval, measured in words, between the relativized noun and the head verb in the relative clause; *left-branching clauses*, including both initial and center-embedding clauses; *depth of left-branching clauses*, measured in syntactic nodes per clause while excluding the last word; and *complex noun phrases and complex adjectival/adverbial modifications*, both measured by their length in words (see Supplementary Table S5 for additional information on the definition of the different 'syntactic structures' metrics). Importantly, only finite subordinate clauses were tagged and counted in the subordination metric.

Error-centered measures were computed based on manually tagged word- and utterance-level errors. The inflectional index, extending the index of Bastiaanse et al. (1996b) to include gender and number inflections for nouns, determiners, and adjectives (Piazza et al., 2022), was calculated by first extracting the frequency of inflected words (*freq + t% mor + sm&* AND + sm-* + t*PAR* command for both suffix and non-concatenated morpheme markers), deriving the count of morphological errors (*freq + s"\[*m:*\]" + r6 + t*PAR*), and dividing the number of correctly inflected words by the total of inflected words (see Supplementary Table S5 for a detailed definition of the error-centered measures). The proportion of utterances with grammatical violations was determined by counting ungrammatical utterances (*freq + s"\[\ + *\]" + r6 + t*PAR*), following CLAN's definition of ungrammaticality (MacWhinney, 2000), and dividing this count by the total number of utterances.

3.5. Statistical analyses

Statistical analyses were performed using Jamovi (Jamovi Project v 2.3.21.0, 2023), and plots were obtained in R (version 4.2.3, R Core Team, 2023) via the ggplot2 package (v 3.5.1; Wickham, 2016) from RStudio (R Core Team, 2023). For each group, measures of central tendency were first calculated for each task and dependent variable (see Tables 3 and 4). Afterwards, based on the parametricity of the data (obtained via Levene's Test for Homogeneity of Variances), either Kruskall-Wallis tests or Welch's One-Way ANOVAs were executed in order to compare syntactic and morphological variables between both groups in each task. *P*-values resulting from group comparisons for each variable within each task were corrected using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Post hoc analyses were

[6] Participants were also tested on a storytelling task in which they were required to narrate the story described in a sequence of pictures that were not made available during production. As we were interested in comparing picture-based tasks that involved visual support during production, this task was not included in the current study.

[7] Despite this, we continued to use the *mor* command of CLAN for calculating the inflectional index, since the morphological analysis provided by Batchalign marked the grammatical gender for fixed-gender nouns and epicenes.

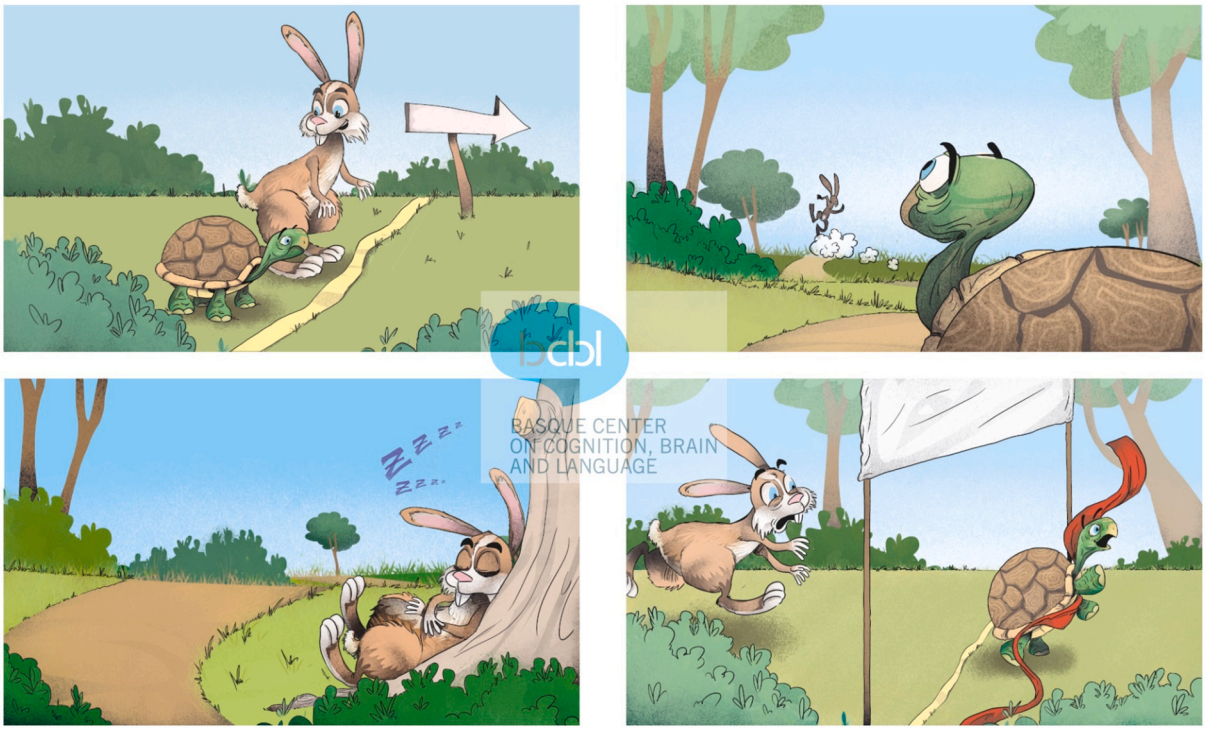


Fig. 1. Sequence of pictures employed in the Hare and the Tortoise story retell task.

Table 3
Descriptive statistics: syntactic proxies.

Task	Group [*n*]	#U *M*(SD)	MLU *M*(SD)	N/V *M*(SD)	V/U *M*(SD)
Cookie Theft	Aphasia [20]	10.6(7.22)	8.97(3.81)	1.12(0.49)	1.14(0.51)
	Control [14]	8.43(4.09)	15(2.87)	1.64(0.49)	1.58(0.40)
Sandwich	Aphasia [20]	4.2(2.33)	8.65(4.14)	2.01(0.97)	0.97(0.33)
	Control [14]	6.14(3.57)	11.5(3.55)	2.00(0.73)	1.15(0.26)
Structured Interview	Aphasia [20]	21.4 (9.13)	7.89(2.82)	1.25(0.5)	1.05(0.35)
	Control [14]	29.2(11.7)	10.6(2.99)	1.13(0.31)	1.39(0.28)
Hare and the Tortoise	Aphasia [19]	9.21(5.11)	9.59(2.69)	1.02(0.42)	1.41(0.42)
	Control [14]	10.4(5.49)	14.5(2.91)	1.41(0.75)	1.81(0.65)

Note: #U = number of produced utterances; MLU = mean length of utterance; N/V = noun-verb ratio; and V/U = verbs per utterance.

conducted after significant main effects of group were observed. A simulation-based post-hoc sensitivity power analysis was conducted in R using the PMCMRplus package (Pohlert, 2024) to estimate the minimum effect size our design had 80% power to detect at $\alpha = 0.05$ (5000 simulations). We assessed the statistical power of each between-group comparison using the same parametric and non-parametric inferential tests as those used in the main analyses. Moreover, to characterize the empirical precision achieved by our study, we calculated effect-size estimates and their 95% confidence intervals (Appendix Table B) using the effect size package in R (Ben-Schachar et al., 2020).

In order to examine whether the syntactic proxies or the 'syntactic structures' metric better predicted group category, four binomial logistic regression models were implemented using all discourse samples combined. Differently from Agmon and colleagues (2024), we used a nested modelling approach, which allowed for an accurate statistical comparison of the amount of variance explained by each model. Predictive performance was assessed using several key measures: accuracy, which refers to the proportion of both correct positive and negative predictions out of total predictions; sensitivity (i.e. the proportion of actual positives correctly identified), specificity (i.e. the proportion of actual negatives correctly identified), and the Area Under the Curve (AUC), which provides an overall assessment of the model's ability to discriminate

Table 4
Descriptive statistics: 'syntactic structures' multidimensional metric and error-centered variables.

Task	Group [*n*]	Sub *M*(SD)	Rl A *M*(SD)	Rl D *M*(SD)	LB A *M*(SD)	LB D *M*(SD)	C NP *M*(SD)	C AM *M*(SD)	Inf I *M*(SD)	Gra V *M*(SD)
Cookie Theft	Aphasia [20]	1.18 (0.19)	0.06 (0.09)	0.25 (0.40)	0.02 (0.03)	0.06 (0.13)	0.82 (0.83)	1.07 (1.13)	0.983 (0.05)	0.14 (0.16)
	Control [14]	1.56 (0.29)	0.21 (0.17)	0.82 (0.75)	0.09 (0.11)	0.32 (0.41)	2.55 (1.86)	2.77 (1.22)	0.996 (0.00)	0.1 (0.26)
Sandwich	Aphasia [20]	1.13 (0.19)	0.01 (0.04)	0.06 (0.27)	0.04 (0.11)	0.27 (0.89)	1.41 (1.65)	0.85 (1.3)	0.992 (0.03)	0.23 (0.34)
	Control [14]	1.24 (0.2)	0.03 (0.07)	0.1 (0.2)	0.05 (0.1)	0.27 (0.69)	2.29 (1.77)	1.86 (1.77)	1.0 (0.00)	0.02 (0.08)
Structured Interview	Aphasia [20]	1.14 (0.13)	0.04 (0.05)	0.14 (0.22)	0.04 (0.05)	0.15 (0.25)	0.62 (0.56)	1.11 (0.82)	0.991 (0.01)	0.12 (0.19)
	Control [14]	1.29 (0.16)	0.06 (0.05)	0.3 (0.31)	0.04 (0.05)	0.19 (0.22)	1.23 (0.90)	2.28 (1.17)	0.999 (0.00)	0.01 (0.03)
Hare and the Tortoise	Aphasia [19]	1.25 (0.19)	0.05 (0.09)	0.3 (0.46)	0.02 (0.04)	0.1 (0.22)	0.57 (0.76)	1.29 (1.14)	0.985 (0.03)	0.19 (0.02)
	Control [14]	1.62 (0.41)	0.07 (0.12)	0.31 (0.72)	0.23 (0.25)	1.1 (1.22)	1.26 (1.55)	4.48 (1.80)	0.999 (0.00)	0.04 (0.09)

Note: Sub = subordination; Rl A = relative clauses; Rl D = dependency distance in relative clauses; LB A = left branching clauses; LB D = depth of the left branching clauses; C NP = complex noun phrases; C AM = complex adjectival/adverbial modifications; Inf I = inflectional index; Gra V = ungrammatical utterances.

between groups.

For this matter, 3 nested models were designed progressively including: A.1) the syntactic proxies, A.2) the measures contained in the 'syntactic structures', and A.3) the error-centered variables. A fourth model (B) was implemented after trimming for the predictors that showed high collinearity (i.e., a Variance Inflation Factors [VIF] greater than 5), and was based on its significant contribution to model fitness, to enhance parsimony.

4. Results

Descriptive statistics (mean and standard deviation) for the two groups across tasks and dependent variables are reported in Tables 3 and 4 Additionally, Fig. 2 illustrates the two groups' performance across tasks and dependent variables.

Across tasks (see Tables 3 and 4), controls consistently produced longer utterances (MLU) and more complex syntactic structures than the aphasia group, with the most pronounced differences observed in the Cookie Theft picture description (Goodglass and Kaplan, 1972) and Hare and the Tortoise story retell. Complex noun phrases and adjectival/adverbial modifications showed the largest between-group differences, with controls scoring substantially higher across all tasks. The inflectional index was close to ceiling for both groups, though with slightly more variability in the aphasia group, while ungrammatical utterances were more frequent in the aphasia group, particularly in the Sandwich task and the Hare and the Tortoise story retell.

4.1. Statistical sensitivity and analyses of variance between groups and across tasks

Statistical sensitivity. The simulation-based sensitivity analysis indicated that the study had 80% power at $\alpha = 0.05$ to detect standardized effects of approximately $|d| = 1.10$ for the Welch tests and $|\delta| = 0.50$ for the Kruskall-Wallis tests. Thus, the study was primarily sensitive to large between-group effects, although effects smaller than these thresholds could also reach statistical significance in the observed sample. A full report of the effect size estimates and their 95% confidence intervals can be found in Appendix B.

Analysis of variance. As illustrated in Table 5 and Fig. 2, control participants significantly outperformed the speakers with aphasia across all syntactic proxies and most syntactic structures variables in their discourse elicited through the Cookie Theft picture description (Goodglass and Kaplan, 1972), with predominantly large effects (ranging from $|d|$: 0.94-1.73; $|\delta|$: 0.5-0.7). In contrast, no significant group differences were observed for the Sandwich task across variables, producing overall small effect sizes ($|d|$: 0.001- 0.72; $|\delta|$: 0.05-0.37). In the Structured Interview, the control group produced significantly longer utterances, more verbs per utterance, a greater rate of subordinate clauses and longer adjectival/adverbial modifications than the aphasia group, with these variables showing the large effect sizes ($|d|$: 0.92 - 1; $|\delta|$: 0.54 - 0.57). Regarding the Hare and the Tortoise story retell, the control group produced significantly longer utterances, with a greater subordination rate, more and longer left branching clauses, longer adjectival/adverbial modifications. These variables also produced medium to large effect sizes ($|d|$: 1.74; $|\delta|$: 0.55- 0.9). In this task, the control group also elicited a lower proportion of ungrammatical

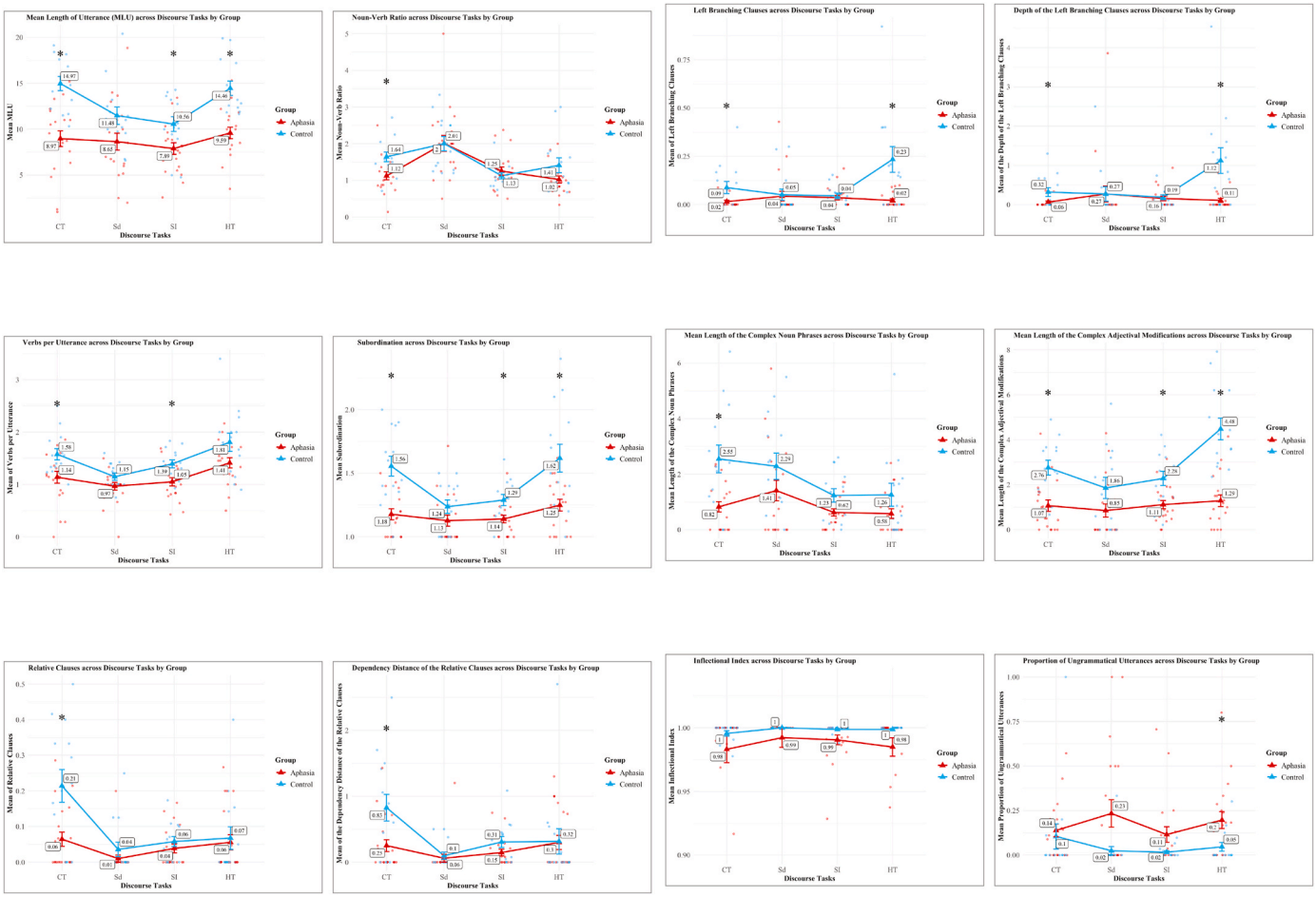


Fig. 2. Group performance across discourse tasks for each linguistic measure.

Table 5
Analyses of variance.

Variables	One-Way ANOVA-/Kruskal-Wallis			
	Cookie Theft	Sandwich	Structured Interview	Hare and the Tortoise
MLU	**$F[1, 31.8] = 27.35, p < 0.001$** **BH $p = 0.004$**	$F[1, 30.6] = 4.56, p = 0.041$ BH $p = 0.228$	**$F[1, 27.1] = 6.84, p = 0.014$** **BH $p = 0.042$**	**$F[1, 26.8] = 23.92, p < 0.001$** **BH $p = 0.003$**
N/V	**$F[1, 28] = 9.08, p = 0.005$** **BH $p = 0.01$**	$F[1, 31] = 0.0005, p = 0.982$ BH $p = 0.982$	$F[1, 31.8] = 0.788, p = 0.381$ BH $p = 0.4572$	$F[1, 19.1] = 3.01, p = 0.099$ BH $p = 0.1485$
V/U	**$F[1, 31.4] = 7.81, p = 0.005$** **BH $p = 0.01$**	$F[1, 31.4] = 3.26, p = 0.081$ BH $p = 0.243$	**$F[1, 31.3] = 9.893, p = 0.004$** **BH $p = 0.03$**	$F[1, 19.1] = 3.88, p = 0.062$ BH $p = 0.1062$
Sub	**$\chi^2[1] = 12.11, p < 0.001$** **BH $p = 0.004$**	$\chi^2[1] = 3.23, p = 0.72$ BH $p = 0.7854$	**$\chi^2[1] = 7.04, p = 0.008$** **BH $p = 0.032$**	**$\chi^2[1] = 7.03, p = 0.008$** **BH $p = 0.0192$**
Rl A	**$\chi^2[1] = 6.65$, p = 0.01** **BH $p = 0.01714$**	$\chi^2[1] = 1.98, p = 0.159$ BH $p = 0.3068$	$\chi^2[1] = 1.64, p = 0.2$ BH $p = 0.2667$	$\chi^2[1] = 0.03, p = 0.845$ BH $p = 0.879$
Rl D	**$\chi^2[1] = 5.31, 0.02$** **BH $p = 0.03$**	$\chi^2[1] = 1.80, p = 0.179$ BH $p = 0.3068$	$\chi^2[1] = 3.08, p = 0.079$ BH $p = 0.1185$	$\chi^2[1] = 0.023, p = 0.879$ BH $p = 0.879$
LB A	**$\chi^2[1] = 4.79, p = 0.029$** **BH $p = 0.0348$**	$\chi^2[1] = 0.2, p = 0.654$ BH $p = 0.7848$	$\chi^2[1] = 0.40, p = 0.525$ BH $p = 0.525$	**$\chi^2[1] = 11.44, p < 0.001$** **BH $p = 0.003$**
LB D	**$\chi^2[1] = 4.79, p = 0.029$** **BH $p = 0.0348$**	$\chi^2[1] = 0.2, p = 0.654$ BH $p = 0.7848$	$\chi^2[1] = 0.617, p = 0.432$ BH $p = 0.4713$	**$\chi^2[1] = 10.64, p = 0.001$** **BH $p = 0.003$**
C NP	**$\chi^2[1] = 8.59, p = 0.003$** **BH $p = 0.009$**	$\chi^2[1] = 2.65, p = 0.103$ BH $p = 0.2472$	$\chi^2[1] = 3.52\, p = 0.06$ BH $p = 0.1029$	$\chi^2[1] = 2.18, p = 0.139$ BH $p = 0.185$
C AM	**$\chi^2[1] = 11.68, p < 0.001$** **BH $p = 0.004$**	$\chi^2[1] = 3.62, p = 0.057$ BH $p = 0.228$	**$\chi^2[1] = 7.838, p = 0.005$** **BH $p = 0.03$**	**$\chi^2[1] = 19.90, p < 0.001$** **BH $p = 0.003$**
Inf I	$\chi^2[1] = 0.002, p = 0.964$ BH $p = 0.964$	$\chi^2[1] = 0.7, p = 0.403$ BH $p = 0.6045$	$\chi^2[1] = 4.08, p = 0.043$ BH $p = 0.1029$	$\chi^2[1] = 1.14, p = 0.285$ BH $p = 0.342$
Gra V	$\chi^2[1] = 2.71, p = 0.1$ BH $p = 0.1090$	$\chi^2[1] = 4.78, p = 0.029$ BH $p = 0.228$	$\chi^2[1] = 3.7, p = 0.054$ BH $p = 0.1029$	**$\chi^2[1] = 6.52, p = 0.011$** **BH $p = 0.022$**

Note: MLU = mean length of utterance; N/V = noun-verb ratio; and V/U = verbs per utterance; Sub = subordination; Rl A = relative clauses; Rl D = dependency distance in relative clauses; LB A = left branching clauses; LB D = depth of left branching clauses; C NP = complex noun phrases; C AM = complex adjectival/adverbial modifications; Inf I = inflectional index; Gra V = ungrammatical utterances. Significant differences after the Benjamini-Hochberg correction (BH *p*) are bolded.

utterances than the aphasia group ($|\delta|$: 0.5).

4.2. Binomial logistic regression models for group prediction

Model A.1 demonstrated a significant association between syntactic proxies and group membership, indicating that the model effectively discriminated between individuals with aphasia and those without brain damage. Inspection of the individual estimates of each predictor obtained via Omnibus Likelihood Ratio tests showed that only MLU ($\chi^2[1] = 7.45$, $p = 0.006$) significantly predicted group membership. Collinearity statistics indicated that all predictors had acceptable VIF levels (<5). Predictive measures showed the model's moderate accuracy in correctly classifying individuals with aphasia, with an acceptable sensitivity and middling specificity. Furthermore, the model possessed good discriminative ability, as reflected by the AUC (see Table 6 for a comparison of fitness and predictive measures of the four models).

Model A.2, which comprised the syntactic proxies together with the 'syntactic structures' as predictors, showed a significant improvement on fitness when compared to the previous model ($\chi^2 = 26.2, p < 0.001$). Individually, dependency distance in relative clauses ($\chi^2[1] = 6.2248$, $p = 0.013$), left branching clauses ($\chi^2[1] = 5.3798, p = 0.020$), depth of left branching clauses ($\chi^2[1] = 7.1651$, $p = 0.007$), complex noun phrases ($\chi^2[1] = 7.1993, p = 0.007$), and complex adjectival/adverbial modifications ($\chi^2[1] = 10.0869$, $p = 0.001$) showed significant associations with group membership. While the model presents good accuracy, sensitivity, specificity, and discriminative ability, high collinearity was found among relative clauses, dependency distance in relative clauses, left branching clauses, and depth of left branching clauses variables.

Model A.3, which added the error-centered variables to the previously included predictors, performed significantly better than the previous model in terms of fitness ($\chi^2 = 13.7, p = 0.001$). The Model A.3 test also confirmed the model's ability to discriminate by group. Dependency distance in relative clauses ($\chi^2[1] = 7.67$, $p = 0.006$), left branching clauses ($\chi^2[1] = 5.87$, $p = 0.015$), depth of left branching clauses ($\chi^2[1] = 10.54$, $p = 0.001$), complex noun phrases ($\chi^2[1] = 4.21$, $p = 0.040$), complex adjectival/adverbial modifications ($\chi^2[1] = 5.95.63, p = 0.015$), inflectional index ($\chi^2[1] = 4.8, p = 0.028$), and proportion of ungrammatical utterances ($\chi^2[1] = 8.56$, $p = 0.003$) showed significant association with group membership. The model A.3 also showed improvement in accuracy, specificity and sensitivity compared to model A2. Yet, similarly to model A.2., A.3. presented high VIF values in relative clauses, dependency distance in relative clauses, left branching clauses, and depth of left branching clauses, in conjunction with MLU and subordination.

Model B results from trimming the MLU, verbs per utterance, relative clauses, and left branching clauses variables from Model A.3 based on collinearity diagnostics and significant associations to group membership. Hence, this trimmed model included the noun-verb ratio and the inflectional index, which did not present significant association to group membership, in addition to subordination ($\chi^2[1] = 5.84$, $p = 0.016$), dependency distance in relative clauses ($\chi^2[1] = 9.67, p = 0.002$), depth of left branching clauses ($\chi^2[1] = 5.39, p = 0.02$), complex noun phrases ($\chi^2[1] = 5.49$, $p = 0.019$), complex adjectival/adverbial modifications ($\chi^2[1] = 8.32$, $p = 0.004$), and proportion of ungrammatical utterances ($\chi^2[1] = 9.86$, $p = 0.002$). Model B also achieves the lowest Bayesian Information Criteria (BIC) from all models, indicating high goodness of fit in relation to a relatively low number of predictors. This final model also presents strong predictive measures, indicating good discriminative ability.

Table 6
Fitness and predictive measures of binomial logistic regression models.

Model	Model fit measures		Predictive measures			
	BIC	χ^2[df](p)	Accuracy	Specificity	Sensitivity	AUC
Model A.1	163	38.0[3] (<0.001)	71.4%	58.9%	80.5%	0.799
Model A.2	171	64.1[10] (<0.001)	78.9%	73.2%	83.1%	0.867
Model A.3	167	77.8[12] (<0.001)	82%	78.6%	84.4%	0.898
Model B	155	70.3[8] (<0.001)	78.2%	80.5%	75%	0.883

Note: BIC = Bayesian Information Criteria; and AUC = Area Under the Curve.

5. Discussion

The current study aimed to identify which elicitation methods, from a set of tasks pertaining to different discourse genres, were most sensitive to capturing simplification of syntactic and morphological complexity in Spanish aphasia, and to determine the best measurement approach for this purpose. PWA and control participants' speech was analyzed to quantify syntactic proxies, such as mean length of utterance, number of verbs per utterance, noun-verb ratio; syntactic structure variables, which included different types of subordinate clauses (relative, left-branching clauses), complex noun phrases and adjectival/adverbial modifications; and finally, error-centered variables (i.e., ungrammatical sentences and morphological errors). The performance of these different metrics was compared to assess their accuracy in classifying participants into either PWA or control participants. Based on previous studies (Olness, 2006; Schnur and Wang, 2024; Stark, 2019; Stark and Fukuyama, 2020; Ulatowska et al., 1981, 1990), we hypothesized that the narrative task would elicit the most significant differences between the aphasia and control group in both syntactic and morphological complexity, and that syntactic structure variables would be more accurate than proxies in determining participants' group membership, in line with Agmon et al. (2024). The results only partially support our hypotheses. Both the descriptive and the narrative tasks were useful to identify patterns of syntactic simplification, although to differing degrees. Moreover, none of the four tasks elicited significant group differences in morphological complexity. Finally, fine-grained syntactic structure metrics proved to be fitter and more accurate than proxies in determining group membership. In the following, we discuss these results and their implications in greater detail.

5.1. Both the narrative and descriptive tasks are useful for comprehensively assessing syntactic complexity

Analyses of variance indicates both the Cookie Theft picture description and the Hare and the Tortoise story retell tasks were sensitive to detecting syntactic simplification in aphasic speech, although the scope of their sensitivity appears to differ. On one hand, the Cookie Theft picture description consistently and reliably elicited group differences in several dependent variables that included both syntactic proxies and syntactic structure, such as MLU, relative and subordinate clauses and complex noun phrases and adjectival modification, but not in error-related measures. On the other hand, the Hare and Tortoise study retell showed reliable group differences in left branching and complex adjectival/adverbial modification. Moreover, our PWA group tended to produce a significantly higher number of ungrammatical sentences compared to the control group (although this effect sits slightly below the minimum detectable effect threshold indicated by our sensitivity analysis, see Appendix B).

Overall, these results are at odds with our research hypothesis, which stated that the narrative genre—Hare and Tortoise story retell—would be the most sensitive to detecting syntactic complexity deficits in PWA. Based on previous findings (Schnur and Wang, 2024), we expected that the narrative task would elicit more syntactically complex discourse than the descriptive task in aphasic populations, which would in turn be reflected in greater inter-group differences. In their study, Schnur and Wang (2024) concluded that storytelling elicits structurally more diverse discourse than picture-description tasks. However, notable methodological differences exist between their study design and the current research, which could account for the differing outcomes. Firstly, Schnur and Wang (2024) employed different variables from the 'syntactic structures' metric (Agmon et al., 2024) that the present study adopted, which makes the comparison difficult and can yield different results. Secondly, Schnur and Wang (2024) utilized different discourse tasks compared to our study, namely the Picnic Scene (Kertesz, 2007) and the Cinderella story retell tasks (Goodglass and Kaplan, 1972). This last point is potentially critical since significantly different outcomes in distinct tasks within the same genre have been reported (Coelho, 2002; Fergadiotis and Wright, 2011; Stark and Fukuyama, 2020; Wright and Capilouto, 2009). Two key differences between the Cinderella and the Hare and Tortoise story retells are that the former provides significantly less visual support, considering participants are asked to narrate the story just after viewing the pictures. In contrast, in the Hare and Tortoise task speakers can inspect the four-picture material while narrating the story. Furthermore, it is possible that the Cinderella story retell elicits syntactically more complex content, as its narrative involves a longer plot, and a greater number of characters and constituent events compared to our task.

Overall, our data show that descriptive and narrative tasks may be sensitive to group differences at different syntactic levels. On the one hand, the Hare and Tortoise story retell task elicits reliable group differences in the number of subordinate clauses, left branching and adjectival/adverbial modifications, all of which relate to speakers' capacity to build hierarchically structured sentences through clause-level modifications. On the other hand, the Cookie Theft picture description task is able to detect group differences in syntactic complexity also at the NP-level, as evidenced by the significant difference between the two groups in relative clauses and complex NPs. Plausibly, picture description tasks require speakers to introduce NPs associated with referents identified in a visually rich scene and to elaborate on them through devices such as relative clauses, while narrating a story demands the creation of a temporal and causal scaffolding that relies more on left branching and adjectival/adverbial modifications.

Visual inspection of the data (Fig. 2) seems to corroborate this interpretation. Specifically, describing a picture and narrating a story may lead the control group to preferentially produce NP- or clause-level syntactic complexity, respectively. Although our narrative task showed less consistency in demonstrating significant group differences across syntactic variables than the Cookie Theft picture description, when comparing both tasks within the control group (see exploratory analysis in the Supplementary Table S6) significant differences (before Benjamini-Hochberg correction) show that the Hare and the Tortoise story retell evoked longer left-branching clauses and more complex adjectival/adverbial modifications than the Cookie Theft description task. Specifically, left branching clauses were frequently used as adjectival/adverbial phrases, since they provided supplementary information pertaining to various aspects such as time specifications, cause-and-effect relationships, and conditional statements. In contrast, a greater number of relative clauses was produced in the Cookie Theft picture description compared to the Hare and Tortoise task. The increased frequency of relative clauses boosted the mean dependency distance in relative clauses and amplified the number of words in complex noun phrases (see for example participant 23 in Table 2). While these patterns should be confirmed by further studies with larger samples, they suggest that speakers within the control group tended to employ a different pattern of syntactic structures and discursive resources in order to respond to the particularities of each task demand, in line with Sherratt (2007) model.

Plausibly, when participants are involved in a single-picture description task, they first retrieve a frame structure that pre-activates the employment of relative clauses, facilitating the specification of nouns leading to a more informative description of the stimuli. Likewise, when participants are required to retell a story, they retrieve a frame structure that eases the availability of left branching clauses, which function as adjectival or adverbial phrases. Adverbial clauses are desirable for a narrative task given they provide an explicit organizing scaffold to a sequence of events, while left branching clauses can operate as a stylistic device that builds up suspense. However, additional research is necessary to corroborate this interpretation and clarify the specific task properties that influence the use of complex syntactic structures.

In line with Ulatowska et al. (1990), the Hare and the Tortoise story retell not only elicited significant differences between groups in MLU

and most 'syntactic structure' variables, but also in the proportion of ungrammatical utterances, unlike the Cookie Theft picture description. Narrating a story involves the embedding of its events in a temporal scaffold and thus it can trigger the formulation of tense sequences and richer syntactic structures, which in turn can increase the probability of producing ungrammatical structures. Eliciting ungrammatical utterances in aphasia assessment has great clinical value, since it can provide qualitative information about which specific language mechanisms are impaired.

The present results also align with prior literature identifying procedural (Stark, 2019; Stark and Fukuyama, 2020; Ulatowska et al., 1981) and conversational genres (Armstrong et al., 2011; Beeke et al., 2003; Sajjadi et al., 2012) as unfit elicitation methods for the assessment of syntax. The Ham and Cheese Sandwich procedural task evoked a high noun-verb ratio given the frequent list-like utterances employed, and the fewest number of utterances (e.g., participant 22: "you need sliced bread, ham, cheese and butter" ["*se necesitan pan de molde, jamón, queso y mantequilla*"]). In contrast, the Structured Interview conversational task elicited the largest number of utterances prompted by five consecutive questions, but the outcome of both groups was often syntactically poor and short in length. This conversational task did not demand extensive answers, as the syntactic information contained in the questions themselves allowed for pragmatically short responses. As an illustration, participant 33 of the control group briefly responded to the question "What are you going to do for Christmas?" ("*¿qué va a hacer en Navidades?*") with "spending time with the family" ("*hacer vida en familia* [% cn:3]"). Although in line with previous literature, our failure to detect syntactic complexity in these two tasks should be interpreted with caution, given their limited power (Appendix B). Beyond the constraints imposed by sample size, the reduced power observed in these two tasks may nevertheless reflect the nature of the speech they elicited: both the list-like formulaic utterances and the pragmatically short responses of the structured-interview task offered very few opportunities for most of the syntactic structure variables under investigation to occur. This restricted range not only limits the number of observable target structures per participant but may also mean that group differences in syntactic capacity are less likely to be elicited by these two tasks, in contrast to the greater variety of syntactic structures available in narrative and descriptive elicitation contexts.

Furthermore, we consider that both pragmatic and contextual factors influenced the performance of participants in the discourse tasks. An important consideration is that participants' informativeness and engagement, as evidenced by the length and number of utterances, may be significantly influenced not only by clinical variables but also by the assessment setting, session rapport, and their inherent personality traits. This is especially relevant for the Structured Interview, where questions delve into the speakers' personal lives and habits. Consequently, in addition to syntactic impairment, the high variability in the number of utterances observed in the Structured Interview (see Table 3) may also be explained by differences in expressive pragmatic skills and/or individual and situational differences (Armstrong et al., 2011). Conversational production is indeed regarded as a useful tool for assessing expressive pragmatic skills (see, for example, the Assessment of Pragmatic Abilities and Cognitive Substrates protocol by Frau et al., 2025). Therefore, it is relevant for future studies to address the relation between pragmatic metrics and the manner in which syntactic structures are employed.

5.2. No discourse task elicited significant group differences in morphological complexity

Morphological complexity was assessed using the inflectional index, which divided the number of correctly inflected words by the total number of inflected words. Overall, morphological errors were uncommon (a frequency similar to that reported by Stark and Fukuyama, 2020), with the aphasia group making three times as many ($n = 30$) as the control group ($n = 9$). The low frequency of errors led to the inflectional index being non-robust, as it neither showed significant group differences in any discourse task in the analysis of variance, nor did it predict group membership in the binomial logistic regression.

However, a trend was observed in the Structured Interview, consistent with findings from Sajjadi et al. (2012), where a task effect in the inflectional index reached significance before correcting for multiple comparisons (Table 5). This result, contrary to our research hypothesis, may be partly explained by the design of the questions, which are intended to elicit verbs in various tense-mood inflections, potentially increasing the opportunities for inflectional errors to surface. However, participants often opted for non-finite verbal forms instead of the target tensed forms, which may have attenuated group differences and could account for why the effect did not survive correction for multiple comparisons. Additionally, this task is cognitively demanding, as it prompts extended discourse through novel questions that require participants to respond via retrieving coherent episodes from long-term memory (Bliss and McCabe, 2006). This trend should be interpreted with caution and considered as hypothesis-generating until replicated in a larger sample, and possibly, with a finer grained morphological metric.

Incorrect person-number agreement of verbs accounted for 50% of the total morphological errors in the aphasia group, while gender agreement errors represented 20% (see Supplementary Table S7 for the classification of morphological errors). In contrast, verb tense errors were rare, amounting to 13%. This may be due to discourse strategies used by individuals with aphasia (Bastiaanse et al., 2002) to reduce linguistic demands, such as avoiding tense and person/number inflection on verbs. For example, a participant with aphasia responded "*en el fin de semana, comprar* (infinitive) *mucha ropa*" ("on the weekend, buying lots of clothes") instead of "*compré* (1st person singular, past tense) *mucha ropa*" ("I bought lots of clothes") to the Structured Interview question "*¿qué hizo el último fin de semana?*" ("What did you do last weekend?").

Finally, although only numerically, the patterns of morphological errors found are not consistent with the tense-agreement dissociation reported in the literature on Spanish aphasia (Benedet et al., 1998; Diéguez-Vide et al., 2012; Gavarró and Martínez-Ferreiro, 2007; Martínez-Ferreiro, 2003, 2010, 2013; Moreno-Torres Sánchez et al., 2005; Stecher et al., 2024; for a review, see Martínez-Ferreiro et al., 2015). This could be due to the different profiles of the PWA tested in this study, very few of whom presented with frank agrammatism, but also to methodological aspects, such as the type of tasks used to elicit inflection (constrained vs. discourse tasks). Past tense is usually the most impaired across tenses (Bastiaanse et al., 2011): It is possible that, with the exception of one specific semi-structured interview question, the four tasks provided limited prompts for the use of the past tense, thereby reducing the number of tense errors observed. However, this interpretation should be confirmed by more robust results with larger group sizes. Also, a combination of discourse analysis and constrained tasks (e.g., sentence completion) might provide a more thorough assessment of the integrity of PWA's inflectional system and highlight potential dissociations between tense and agreement.

5.3. 'Syntactic structures' is more accurate in predicting groups than proxies

The model including 'syntactic structures' (Agmon et al., 2024) variables together with the syntactic proxies, performed significantly better than the model that exclusively used syntactic proxies. The multidimensional metric of 'syntactic structures' was originally designed and tested by Agmon and colleagues (2024) for predicting age groups, with successful results. The present results indicate that 'syntactic structures' provide a model with good fitness and discriminatory power when predicting aphasia versus control groups. Crucially, its fitness and predictive power is not affected when excluding syntactic proxies such as MLU and verbs per utterance, which were found to be

collinear with syntactic structure variables. Notably, the model is significantly enhanced in discriminative ability and fitness when error-centered variables are included. This suggests that including predictors that measure morphological and grammatical errors is relevant, as these variables indicate difficulties in computing morphosyntactic relations. Word- and utterance-level errors, characteristic of the discourse of speakers with aphasia, are frequently assessed in current aphasiology research and clinical practices (Stark et al., 2021b). Thus, it is appropriate that these linguistic phenomena are noted when designing a model that is discriminative between aphasic and non-brain damage discourse.

5.4. Implications for accounts of agrammatism

Recent accounts of morphosyntactic impairments in post-stroke aphasia have advocated for the processing, rather than structural, nature of the deficit. From this perspective, difficulties in the production (and comprehension) of verb inflection and of different types of structures do not stem from the loss of underlying syntactic representations (Tree Pruning Hypothesis, Friedman & Grodzinsky, 1997; Trace Deletion Hypothesis, Grodzinsky, 1995; see Druks, 2016 for a review), but from the damage to specific processing mechanisms (Kolk and Heeschen, 1992; Kok et al., 2007; McNeil et al., 1994; Thompson, 2003, to name a few; see a review in Bastiaanse and Thompson, 2012). Our approach and findings are agnostic to both structural and processing accounts, in that we limit our investigation to metrics that best detect morphosyntactic simplification to predict group membership. However, the different degrees of syntactic simplification that we observe within our PWA group are more easily accounted for under a processing account, since variability in performance is not compatible with the assumption that syntactic representations are unavailable to aphasic speakers.

The lack of online dependent variables, structured stimuli and instructions, and specific experimental manipulations, combined with recruitment criteria that extend beyond agrammatic patients, prevent us from pinpointing the precise syntactic processing mechanisms that may be impaired. Our finding that regression models including hierarchical syntactic measures are more accurate at classifying speakers with and without aphasia could, however, provide some hints about the levels of linguistic analysis at which impaired mechanisms could operate. Specifically, it suggests that in our sample of PWA, the impairment goes beyond the lexical level, reaching processing stages involved in e.g., syntactic structure planning and building. In this respect, two specific patterns are worth highlighting. First, the reduced proportion of left-branching structures, which may require the speaker to hold complex material in memory before reaching the head of the sentence, could reflect impairment in sentence planning mechanisms (arguably mediated by short-term and working memory deficits). Such impairments can be probed with dedicated tasks (Lee et al., 2015; Van Boxtel et al., 2023) and through the analysis of speech dysfluencies, though this falls beyond the scope of the present study. Second, the significant contribution of ungrammatical sentence proportion opens a window onto the integrity of monitoring mechanisms (Botvinick et al., 2001), i.e., the speaker's ability to detect and repair errors. PWA who produce a greater number of ungrammatical sentences may present with impaired monitoring that hinders self-correction, which could be further investigated using tasks tapping attentional mechanisms (e.g., the Attention Network Task; Fan et al., 2005). While theoretically plausible, these implications remain tentative and should be corroborated by data from structured linguistic and cognitive tasks collected in larger samples.

5.5. Discourse, syntax and the brain-language relationship

Despite being limited to behavioral data, this study carries meaningful implications for the brain-language relationship, which future research could further investigate. Specifically, the present results suggest that the heterogeneous brain damage in our sample, which resulted in aphasia, may have affected resources needed to necessary to successfully deploy complex syntactic structures, such as left-branching clauses and relative clauses (see Den Ouden et al., 2008, Kircher et al., 2005; Lorca-Puls et al., 2024 for evidence regarding neuroanatomical correlates of complex syntax production and expressive agrammatism). As noted earlier, these syntactic devices differ not only in their complexity, but also in their locality and function. Relative clauses attach postnominally to an already established head, supporting local, within-sentence referential elaboration. Left-branching clauses, by contrast, are preposed and require the speaker to build and hold open a hierarchical dependency before reaching the main clause. Moreover, they typically encode temporal, causal and conditional relations that make up a discourse model.

Although mapping behavioral impairment to lesioned brain regions is clearly beyond the scope of this study, this functional distinction parallels proposals of a graded cortical hierarchy in which local, sentence-level combinatorial processing is supported by brain regions that operate over intermediate temporal integration windows, such as the superior temporal gyrus (Lerner et al., 2011). In contrast, maintaining and updating a discourse model across sentences involves also regions with longer temporal receptive windows, such as temporo-parietal and prefrontal areas (Lerner et al., 2011; see also Catani and Bambini, 2014), with these two levels interfacing rather than operating as separate modules. This hierarchy is based on narrative comprehension, but evidence of shared and coupled neural systems for the production and comprehension of naturalistic narrative speech (Awad et al., 2007; Giglio et al., 2022; Silbert et al., 2014) suggests that an analogous hierarchy could support narrative production as well. The selective vulnerability of left-branching structures in the narrative task could therefore stem from a bottleneck at this syntax-discourse interface, rather than from a generalized impairment in syntactic production.

5.6. Clinical and research implications

The evaluation of populations with clinical conditions of language should follow ecologically valid approaches such as discourse assessment (Davidson et al., 2003; World Health Organization, 2018). However, most researchers (Stark et al., 2021b) and speech and language therapists (Bryant et al., 2016) use a single elicitation method for evaluating discourse production in aphasia, with the single-picture description most frequently administered in English-speaking countries (Stark et al., 2021b). Under these circumstances, it is possible that aphasiology researchers and speech and language therapists are selecting insensitive discourse elicitation methods (Stark and Fukuyama, 2020). Therefore, and following the rationale of Schnur and Wang (2023), it is relevant to compare discourse elicitation methods in order to inform both empirical research on connected speech production and clinical strategies for identifying the most suitable production task or combination of tasks for the assessment of a specific language impairment. The present research addressed this need in the clinical field of Spanish aphasia by aiming to identify the most sensitive discourse tasks and metrics for eliciting and detecting syntactic and morphological complexity simplification.

Based on our findings, two methodological suggestions can be put forward for researchers and clinicians aiming to comprehensively assess syntactic complexity in Spanish speakers with aphasia. Firstly, when evaluating syntactic complexity, it is pertinent to conjointly administer descriptive and narrative tasks, such as the Cookie Theft picture description (or an equivalent single-picture description task, e.g. the Picnic Scene from the Western Aphasia Battery; Kertesz, 2007) and a narrative task akin to our Hare and the Tortoise story retell. This recommendation is supported by the reliable group differences elicited across several syntactic metrics by the descriptive task, the group difference in ungrammatical utterances elicited by the narrative task (a critical clinical marker of aphasia), and the distinct patterns of syntactic

resources observed across tasks, consistent with Stark and Fukuyama's (2020) argument for the use of varied discourse elicitation methods for aphasia evaluation.

Secondly, from the obtained discourse samples, it is beneficial for researchers to measure a set of fit and noncollinear indicators such as noun-verb ratio, a broad embedding index (e.g., the employed subordination metric), dependency distances between the modifying verb within the relative clause and the relativized noun in the main clause, depth of left branching clauses, length of the complex noun phrases and adjectival/adverbial modifications, and proportion of ungrammatical utterances. The obtained measures could then greatly inform research and clinical settings to guide rehabilitation procedures.[8]

5.7. Limitations and future directions

The most evident limitation of the current study is the small sample size, which limits statistical power such that only large effects can be reliably detected. Moreover, the aphasia group exhibited considerable heterogeneity in symptomatic manifestations, post-onset periods, and comorbidities. Together, these factors prevented us from investigating and discussing genre effects on the morphological and syntactic complexity across different aphasia subtypes, severity and cognitive profiles, thus restricting our comparisons to the PWA vs. control group contrast. Previous aphasiology research has found evidence for the simplification of both syntactic (Bastiaanse et al., 1996a; Glosser and Deser, 1991) and morphological complexity (Bastiaanse et al., 1996b, 2002; Martínez-Ferreiro et al., 2017) in fluent and nonfluent aphasia. A larger, more controlled, sample could allow for testing the sensitivity and accuracy of our metrics across aphasia profiles and exploring possible interactions with tasks, an avenue we intend to pursue in future extensions of this work. Nevertheless, this investigation represents an initial step toward validating the hierarchical syntactic structure metrics as tools for accurately detecting patterns of syntactic simplification in PWA and language-unimpaired speakers.

A second limitation of the present study is the failure to include covariates in the analyses due to heteroscedasticity in most variables, which required non-parametric tests. Increasing the sample size is unlikely to resolve this issue, as the 'syntactic structures' metric showed frequent floor effects and a strongly left-skewed distribution, as anticipated by Agmon et al. (2024). Likewise, the error-centered metrics faced similar challenges. These limitations may be intrinsic to the 'syntactic structures' metric, as it does not account for elements essential to communication, allowing participants to complete tasks without using complex syntactic structures like relative or left-branching clauses.

A third limitation is the lack of reliability and stability estimates for the employed outcome variables. Therefore, as the employed outcome variables lack standardization, it is uncertain to what degree each reflects the linguistic integrity of the participants. Thus, future extensions of this study should investigate which syntactic and morphological discourse metrics are stable over time through test-retest studies with speakers with aphasia (Stark et al., 2021a).

A further limitation of the present study concerns the cross-linguistic generalizability of task-based findings. Our results suggest that the narrative task was not more sensitive than the descriptive task in detecting syntactic complexity deficits in Spanish-speaking PWA, which contrasts with findings reported in English (Schnur and Wang, 2024; Stark, 2019; Stark and Fukuyama, 2020; Ulatowska et al., 1981, 1990). Moreover, the morphological properties of Spanish, such as the bound nature of verb stems, may constrain the type of errors and avoidance strategies available to PWA, which would in turn influence the sensitivity of the inflectional index. However, the present study was not designed to directly compare task-elicited syntactic or morphological complexity across languages and therefore cannot determine whether this discrepancy is attributable to language-specific morphosyntactic properties of Spanish or to other methodological factors discussed above. Future research should directly address these gaps by comparing task-elicited syntactic and morphological complexity across languages using equivalent tasks and metrics, in order to better understand how language-specific properties interact with discourse genre in shaping morphosyntactic production in both typical and aphasic populations.

Finally, the current study only employed an inflectional index (an accuracy indicator based on Bastiaanse et al., 1996b) as a measure of morphological complexity. The results suggest that the inflectional index evokes exceedingly modest differences between the aphasia and control groups. In order to have a better insight of task effects in verb production within aphasic discourse, future research should inspect the distribution of finite and non-finite forms (Martínez-Ferreiro et al., 2017), and implement a finer-grained analysis on the types of inflections, at the interface with other levels of language (Bastiaanse and Prins, 1994).

6. Conclusion

The current study showed that picture-based descriptive and narrative tasks are sensitive tools for the detection of syntactic simplification in the spontaneous speech of Spanish PWA, while the analysis of morphological integrity may require the combination of (dialogic) discourse with constrained tasks. Fine-grained aspects of syntactic production, such as subordination, relative clauses and left-branching depth, emerge as highly accurate predictors of speakers' group membership. In conclusion, this study highlights the combined utility of specific discourse tasks and structured metrics to reveal syntactic simplification in aphasia, underscoring the value of multi-dimensional assessments in comprehensively understanding language impairments.

CRediT authorship contribution statement

Jorge Ricardo Hidalgo Chagoya: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xabi Ansorena:** Software, Methodology, Investigation, Data curation, Conceptualization. **Roel Jonkers:** Writing – review & editing, Visualization, Supervision. **Maddi Carrera:** Investigation, Data curation. **José Ignacio Quemada Ubis:** Resources, Investigation. **Simona Mancini:** Writing – review & editing, Visualization, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization.

Ethics declaration

Written informed consent to take part in the study and to publish the article has been obtained from all participants or their legal representatives. The privacy rights of participants have been observed.

This study was performed in compliance with relevant laws, regulatory frameworks and guidelines where the research took place. This study was approved by the BCBL Ethics Review Board. (Approval No.

[8] Our study broadly shares a multi-metric approach with existing treatment frameworks that tap on morphosyntax, such as the Language Assessment, Remediation and Screening Procedure (LARSP; Crystal, 1979), a point raised by an anonymous reviewer. In brief, LARSP offers a descriptive taxonomy of syntactic production organized across three main levels of analysis: the word level, capturing morphological markers; the phrase level, examining the internal elaboration of noun, verb, and prepositional phrases; and the clause level, investigating how elements combine within the main clause. These structures are analyzed with respect to developmental stages designed to identify the most natural next step in a patient's linguistic growth, thereby informing rehabilitation practices. Although our work shares LARSP's commitment to multi-level structural analysis, the clinical focus on remediation and its stages clearly falls outside our scope. In this sense, our quantitative approach extends the principled structural logic that LARSP established, without seeking to replace its clinical longitudinal insights.

13022018M) This study was also approved by the Comité de Ética de Investigación de Euskadi (Approval No. PI2018033)

Data availability statement

The dataset generated and analyzed during the current study is available from the corresponding authors on reasonable request.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used Claude Sonnet 5, developed by Anthropic, for debugging code and proofreading. The authors reviewed and edited the output as needed and take full responsibility for the content of the published article.

Declaration of competing interests

We have no conflicts of interest to disclose.

Acknowledgements

This research is supported by the Basque Government through the BERC 2022-2025 program and funded by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010/AEI/10.13039/501100011033. S.M. acknowledges funding from the Agencia Estatal de Investigación (project PID2020-113945RB-I00 and project PID2024-159519OB-I00). J.H.C. recognizes funding from the Erasmus Mundus II programme of the European Union.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2026.109607.

Appendix Table A
Mean group raw scores of the neuropsychological and BETA assessment.

	Aphasia *M*(SD)	Control *M*(SD)
Neuropsychological tests (Peña-Casanova et al., 2009; Smith Watts et al., 2019; Benton, 1991)		
Auditory digit span forward	4.77(1.45)	6.31(1.09)
Auditory digit span backwards	3.51(1.52)	4.08(0.84)
Visual digit span forward	4.77(1.83)	6.46(1.16)
Visual digit span backwards	3.91(1.78)	4.85(2.09)
Corsi Blocks span forward	4.59(1.3)	5.08(0.93)
Corsi Blocks span backwards	4.46(1.62)	4.92(1.4)
Trail Making Test A	54.4(25.8)	31.1(19)
Trail Making Test B	154(98.1)	69.4(46)
Benton Visual Retention Test	8.1(1.63)	9(0.79)
BETA subtests (Cuetos and González-Nosti, 2009)		
Lexical decision	30.88(1.07)	31.61(0.65)
Word-picture matching	29.45(0.82)	29.64(0.74)
Pseudoword repetition	27.92(1.85)	28.92(1.2)
Object naming	27.12(2.65)	29.28(1.06)
Action naming	25.93(2.67)	27.57(2.5)
Semantic association	29.52(0.51)	29.3(0.63)
Sentence-to-picture matching	18.33(1.79)	19.07(0.82)

Note: Both scores of the Trail Making Test are expressed in seconds. When individuals were not able to correctly complete the task within the time limit, the maximum score was assigned (marked '101' for Trail A, and '301' for Trail B) according to the scoring convention (Smith Watts et al., 2019).

Appendix Table B
Estimated observed power for each variable in each task, and corresponding 95% confidence intervals.

Cookie Theft	DV	Estimate	CI low	CI high
	MLU	**−1,73**	−2527	−0,92
	N/V	−1,06	−1795	−0,317
	VU	−0,94	−1648	−0,208
	Sub	**−0,7**	−0,855	−0,442
	Rl A	**−0,5**	−0,742	−0,153
	Rl D	−0,45	−0,709	−0,084
	LB A	−0,37	−0,658	0,01
	LB D	−0,37	−0,658	0,01
	C NP	**−0,59**	−0,794	−0,275
	C AM	**−0,7**	−0,851	−0,43
	Gra Vio	0,31	−0,075	0,62
	Inf I	0,01	−0,374	0,386
Sandwich	MLU	−0,72	−1424	−0,013
	N/V	0,01	−0,683	0,698
	VU	−0,6	−1298	0,099
	Sub	−0,34	−0,64	0,043
	Rl A	−0,16	−0,51	0,234
	Rl D	−0,15	−0,504	0,24
	LB A	−0,06	−0,431	0,327
	LB D	−0,06	−0,431	0,327

(*continued on next page*)

Appendix Table B (*continued*)

	C NP	−0,33	−0,63	0,059
	C AM	−0,37	−0,656	0,014
	Gra Vio	0,35	−0,039	0,642
	Inf I	−0,05	−0,422	0,336
Structured interview	MLU	−0,92	−1634	−0,196
	N/V	0,29	−0,403	0,971
	VU	−1,05	−1776	−0,316
	Sub	**−0,54**	−0,763	−0,2
	Rl A	−0,25	−0,575	0,144
	Rl D	−0,34	−0,64	0,043
	LB A	−0,12	−0,479	0,271
	LB D	−0,15	−0,501	0,244
	C NP	−0,38	−0,666	−0,002
	C AM	**−0,57**	−0,782	−0,245
	Gra Vio	0,36	−0,026	0,649
	Inf I	−0,34	−0,64	0,043
Hare and the Tortoise	MLU	**−1,74**	−2549	−0,919
	N/V	−0,66	−1367	0,053
	VU	−0,74	−1449	−0,02
	Sub	**−0,55**	−0,768	−0,204
	Rl A	−0,03	−0,412	0,355
	Rl D	0,03	−0,361	0,406
	LB A	**−0,64**	−0,821	−0,338
	LB D	**−0,62**	−0,809	−0,305
	C NP	−0,29	−0,609	0,102
	C AM	**−0,9**	−0,952	−0,778
	Gra Vio	**0,5**	0,139	0,739
	Inf I	−0,16	−0,511	0,241

Note: Estimates at or above $|d| = 1.1$ and $|\delta| = 0.5$ (MDEs) are bolded. Negative estimates indicate lower counts/accuracy for the aphasia group than the control group. pairwise CIs are unadjusted single-contrast intervals.

References

Agmon, G., Pradhan, S., Ash, S., Nevler, N., Liberman, M., Grossman, M., Cho, S., 2024. Automated measures of syntactic complexity in natural speech production: older and younger adults as a case study. J. Speech Lang. Hear. Res. 1–17. https://doi.org/10.1044/2023_JSLHR-23-00009.

Ansorena, X., Carreiras, M., Hernández, M., Benitez, Baena I., Mancini, S., 2022. Aphasia Cognitive Screening in Spanish (Acs.Esp): *a New Digital Test to Assess Language in Aphasia* [Conference Abstract]. Science of Aphasia (SoA), Bordeaux, France. https://tinyurl.com/4s2ma9z8.

Armstrong, E., 2000. Aphasic discourse analysis: the story so far. Aphasiology 14 (9), 875–892. https://doi.org/10.1080/02687030050127685.

Armstrong, E., Ciccone, N., Godecke, E., Kok, B., 2011. Monologues and dialogues in aphasia: some initial comparisons. Aphasiology 25 (11), 1347–1371. https://doi.org/10.1080/02687038.2011.577204.

Awad, M., Warren, J.E., Scott, S.K., Turkheimer, F.E., Wise, R.J., 2007. A common system for the comprehension and production of narrative speech. J. Neurosci. 27 (43), 11455–11464.

Bastiaanse, R., Bamyaci, E., Hsu, C.J., Lee, J., Duman, T.Y., Thompson, C.K., 2011. Time reference in agrammatic aphasia: a cross-linguistic study. J. Neurolinguistics 24 (6), 652–673. https://doi.org/10.1016/j.jneuroling.2011.07.001.

Bastiaanse, R., Edwards, S., Kiss, D., 1996a. Fluent aphasia in three languages: aspects of spontaneous speech. Aphasiology 10 (6), 561–575. https://doi.org/10.1080/02687039608248437.

Bastiaanse, R., Hugen, J., Kos, M., van Zonneveld, R., 2002. Lexical, morphological, and syntactic aspects of verb production in agrammatic aphasics. Brain Lang. 80 (2), 142–159. https://doi.org/10.1006/brln.2001.2585.

Bastiaanse, R., Jonkers, R., Moltmaker-Osinga, U., 1996b. Aspects of lexical verbs in the spontaneous speech of agrammatic and anomic patients. Lang. Cognit. 5, 13–26.

Bastiaanse, R., Prins, R.S., 1994. Communicative speech therapy in aphasia: what does it mean, can it be effective and how should it be done? Aphasiology 8 (5), 482–488. https://doi.org/10.1080/02687039408248672.

Bastiaanse, R., Thompson, C.K. (Eds.), 2012. Perspectives on Agrammatism. Psychology Press, New York.

Beeke, S., Wilkinson, R., Maxim, J., 2003. Exploring aphasic grammar 2: do language testing and conversation tell a similar story? Clin. Linguist. Phon. 17 (2), 109–134. https://doi.org/10.1080/0269920031000061786.

Ben-Shachar, M.S., Lüdecke, D., Makowski, D., 2020. Effectsize: estimation of effect size indices and standardized parameters. J. Open Source Softw. 5 (56), 2815. https://doi.org/10.21105/joss.02815.

Benedet, M.J., Christiansen, J.A., Goodglass, H., 1998. A cross-linguistic study of grammatical morphology in Spanish-and English-speaking agrammatic patients. Cortex 34 (3), 309–336. https://doi.org/10.1016/S0010-9452(08)70758-5.

Beniamine, S., 2021. One lexeme, many classes: inflection class systems as lattices. In: Crysmann, B., Sailer, M. (Eds.), One-To-Many Relations in Morphology, Syntax and Semantics. Language Science Press, pp. 23–51.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57 (1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Benton, A., 1991. *Benton Visual Retention Test - Fifth Edition* [Database record]. APA PsycTests. https://doi.org/10.1037/t14985-000.

Bliss, L.S., McCabe, A., 2006. Comparison of discourse genres: clinical implications. Contemporary Issues in Commun. Sci. Disord. 33, 126–167 (Fall).

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., Cohen, J.D., 2001. Conflict monitoring and cognitive control. Psychol. Rev. 108 (3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624.

Bryant, L., Ferguson, A., Spencer, E., 2016. Linguistic analysis of discourse in aphasia: a review of the literature. Clin. Linguist. Phon. 30 (7), 489–518. https://doi.org/10.3109/02699206.2016.1145740.

Burchert, F., Swoboda-Moll, M., De Bleser, R., 2005. The left periphery in agrammatic clausal representations: evidence from German. J. Neurolinguistics 18 (1), 67–88. https://doi.org/10.1016/j.jneuroling.2004.11.007.

Cannizzaro, M.S., Coelho, C.A., 2013. Analysis of narrative discourse structure as an ecologically relevant measure of executive function in adults. J. Psycholinguist. Res. 42, 527–549. https://doi.org/10.1007/s10936-012-9231-5.

Catani, M., Bambini, V., 2014. A model for social communication and language evolution and development (SCALED). Curr. Opin. Neurobiol. 28, 165–171.

Coelho, C.A., 2002. Story narratives of adults with closed head injury and non-brain-injured adults. https://doi.org/10.1044/1092-4388(2002/099.

Crystal, D., 1979. Working with LARSP. Edward Arnold, London.

Cuetos, F., González-Nosti, M., 2009. BETA: Batería Para La Evaluación De Los Trastornos Afásicos. EOS, Madrid.

Davidson, B., Worrall, L., Hickson, L., 2003. Identifying the communication activities of older people with aphasia: evidence from naturalistic observation. Aphasiology 17 (3), 243–264. https://doi.org/10.1080/729255457.

den Ouden, D.B., Hoogduin, H., Stowe, L.A., Bastiaanse, R., 2008. Neural correlates of Dutch Verb Second in speech production. Brain Lang. 104 (2), 122–131. https://doi.org/10.1016/j.bandl.2007.05.001.

Diéguez-Vide, F., Gich-Fullà, J., Puig-Alcántara, J., Sánchez-Benavides, G., Peña-Casanova, J., 2012. Chinese–Spanish–Catalan trilingual aphasia: a case study. J. Neurolinguistics 25 (6), 630–641. https://doi.org/10.1016/j.jneuroling.2012.01.002.

Doedens, W.J., Meteyard, L., 2020. Measures of functional, real-world communication for aphasia: a critical review. Aphasiology 34 (4), 492–514. https://doi.org/10.1080/02687038.2019.1702848.

Doyle, P.J., Goda, A.J., Spencer, K.A., 1995. The communicative informativeness and efficiency of connected discourse by adults with aphasia under structured and conversational sampling conditions. Am. J. Speech Lang. Pathol. 4 (4), 130–134. https://doi.org/10.1044/1058-0360.0404.130.

Druks, J., 2016. Contemporary and Emergent Theories of Agrammatism: a Neurolinguistic Approach. Routledge.

Fan, J., McCandliss, B.D., Fossella, J., Flombaum, J.I., Posner, M.I., 2005. The activation of attentional networks. Neuroimage 26 (2), 471–479. https://doi.org/10.1016/j.neuroimage.2005.02.004.

Fergadiotis, G., Wright, H.H., 2011. Lexical diversity for adults with and without aphasia across discourse elicitation tasks. Aphasiology 25 (11), 1414–1430. https://doi.org/10.1080/02687038.2011.603898.
Frau, F., Bosia, M., Bischetti, L., Cappelli, G., Carotenuto, A., Diamanti, L., et al., 2025. Ten years of using the APACS test: a multistudy cross-diagnostic analysis of pragmatic profiles and their relationship with theory of mind. Phil. Trans. B 380 (1932), 20230495. https://doi.org/10.1098/rstb.2023.0495.
Frederiksen, C.H., Bracewell, R.J., Breuleux, A., Renaud, A., 1990. The cognitive representation and processing of discourse: function and dysfunction. In: Discourse Ability and Brain Damage: Theoretical and Empirical Perspectives. Springer New York, New York, NY, pp. 69–110.
Friedmann, N.A., Grodzinsky, Y., 1997. Tense and agreement in agrammatic production: pruning the syntactic tree. Brain Lang. 56 (3), 397–425. https://doi.org/10.1006/brln.1997.1795.
Garraffa, M., Fyndanis, V., 2020. Linguistic theory and aphasia: an overview. Aphasiology 34 (8), 905–926. https://doi.org/10.1080/02687038.2020.1770196.
Garraffa, M., Smith, G., 2024. Linguistic Theory for Aphasia. Palgrave Macmillan.
Garrard, P., Forsyth, R., 2010. Abnormal discourse in semantic dementia: a data-driven approach. Neurocase 16, 373–389. https://doi.org/10.1080/13554791003785901.
Gavarró, A., Martínez-Ferreiro, S., 2007. Tense and agreement impairment in Ibero–Romance. J. Psycholinguist. Res. 36 (1), 25–46. https://doi.org/10.1007/s10936-006-9031-x.
Giglio, L., Ostarek, M., Weber, K., Hagoort, P., 2022. Commonalities and asymmetries in the neurobiological infrastructure for language production and comprehension. Cerebr. Cortex 32 (7), 1405–1418.
Glosser, G., Deser, T., 1991. Patterns of discourse production among neurological patients with fluent language disorders. Brain Lang. 40 (1), 67–88. https://doi.org/10.1016/0093-934X(91)90117-J.
Goodglass, H., Kaplan, E., 1972. The Assessment of Aphasia and Related Disorders, first ed. Lea & Febiger.
Gordon, J.K., 2020. Factor analysis of spontaneous speech in Aphasia. J. Speech Lang. Hear. Res. 63, 4127–4147. https://doi.org/10.1044/2020_JSLHR-20-0034.
Grodzinsky, Y., 1995. Trace deletion, θ-roles, and cognitive strategies. Brain Lang. 51 (3), 469–497. https://doi.org/10.1006/brln.1995.1072.
Hernández-Sacristán, C., Rosell-Clari, V., 2009. Syntax and conversation in aphasia. A strategic restrictive use of Spanish and Catalan connector QUE by aphasic speakers. Clin. Linguist. Phon. 23 (10), 717–741. https://doi.org/10.3109/02699200903063053.
Hualde, J.I., Olarrea, A., O'Rourke, E. (Eds.), 2012. The Handbook of Hispanic Linguistics. Wiley-Blackwell, Oxford.
Hunt, K.W., 1970. Syntactic maturity in schoolchildren and adults. Monogr. Soc. Res. Child Dev. 35 (1), III–67. https://doi.org/10.2307/1165818.
Jamovi Project v 2.3.21.0, 2023. Jamovi [Computer software]. Retrieved from. https://www.jamovi.org.
Kemper, S., McDowd, J., Pohl, P., Herman, R., Jackson, S., 2006. Revealing language deficits following stroke: the cost of doing two things at once. Aging Neuropsychol. Cognit. 13 (1), 115–139.
Kertesz, A., 2007. Western Aphasia Battery-Revised (WAB-R). APA PsycTests. https://doi.org/10.1037/t15168-000.
Kintsch, W., 1994. Text comprehension, memory, and learning. Am. Psychol. 49 (4), 294.
Kircher, T.T., Oh, T.M., Brammer, M.J., McGuire, P.K., 2005. Neural correlates of syntax production in schizophrenia. Br. J. Psychiatr. 186 (3), 209–214. https://doi.org/10.1192/bjp.186.3.209.
Kok, P., van Doorn, A., Kolk, H., 2007. Inflection and computational load in agrammatic speech. Brain Lang. 102 (3), 273–283. https://doi.org/10.1016/j.bandl.2007.03.001.
Kolk, H., Heeschen, C., 1992. Agrammatism, paragrammatism and the management of language. Lang. Cognit. Process. 7 (2), 89–129. https://doi.org/10.1080/01690969208409381.
Leaman, M.C., Edmonds, L.A., 2023. Analyzing language in the picnic scene picture and in conversation: the type of discourse sample we choose influences findings in people with aphasia. Am. J. Speech Lang. Pathol. 1–18. https://doi.org/10.1044/2023_AJSLP-22-00279.
Lee, J., Yoshida, M., Thompson, C.K., 2015. Grammatical planning units during real-time sentence production in speakers with agrammatic aphasia and healthy speakers. J. Speech Lang. Hear. Res. 58 (4), 1182–1194. https://doi.org/10.1044/2015_JSLHR-L-14-0250.
Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31 (8), 2906–2915.
Linnik, A., Bastiaanse, R., Höhle, B., 2015. Discourse production in aphasia: a current review of theoretical and methodological challenges. Aphasiology 30 (7), 765–800. https://doi.org/10.1080/02687038.2015.1113489.
Liu, H., MacWhinney, B., Fromm, D., Lanzi, A., 2023. Automation of language sample analysis. J. Speech Lang. Hear. Res. 66 (7), 2421–2433. https://doi.org/10.1044/2023_JSLHR-22-00642.
Lorca-Puls, D.L., Gajardo-Vidal, A., Mandelli, M.L., Illán-Gala, I., Ezzes, Z., Wauters, L.D., et al., 2024. Neural basis of speech and grammar symptoms in non-fluent variant primary progressive aphasia spectrum. Brain 147 (2), 607–626. https://doi.org/10.1093/brain/awad327.
MacWhinney, B., 2000. The CHILDES Project: Tools for Analyzing Talk, third ed. Lawrence Erlbaum Associates. https://doi.org/10.21415/T5B97X.
MacWhinney, B., 2019. Understanding spoken language through TalkBank. Behav. Res. Methods 51 (4), 1919–1927. https://doi.org/10.3758/s13428-018-1174-9.
MacWhinney, B., Fromm, D., Forbes, M., Holland, A., 2011. AphasiaBank: methods for studying discourse. Aphasiology 25 (11), 1286–1307. https://doi.org/10.1080/02687038.2011.589893.
Mandler, J.M., 1987. On the psychological reality of story structure. Discourse Process. 10 (1), 1–29. https://doi.org/10.1080/01638538709544656.
March, E.G., Wales, R., Pattison, P., 2006. The uses of nouns and deixis in discourse production in Alzheimer's disease. J. Neurolinguistics 19, 311–340. https://doi.org/10.1016/j.jneuroling.2006.01.001.
Marini, A., Andreetta, S., del Tin, S., Carlomagno, S., 2011. A multi-level approach to the analysis of narrative language in aphasia. Aphasiology 25 (11), 1372–1392. https://doi.org/10.1080/02687038.2011.584690.
Martínez-Ferreiro, S., 2003. Verbal Inflectional Morphology in Broca's Aphasia. Unpublished MA Thesis. Universitat Autònoma de Barcelona.
Martínez-Ferreiro, S., 2010. Towards a Characterization of Agrammatism in Ibero-Romance. Universitat Autònoma de Barcelona. Doctoral dissertation.
Martínez-Ferreiro, S., Bastiaanse, R., 2013. Time reference in Spanish and Catalan non-fluent aphasia. Lingua 137, 88–105. https://doi.org/10.1016/j.lingua.2013.09.003.
Martínez-Ferreiro, S., de Aguiar, V., Rofes, A., 2015. Non-fluent aphasia in Ibero-Romance: a review of morphosyntactic deficits. Aphasiology 29 (1), 101–126. https://doi.org/10.1080/02687038.2014.958915.
Martínez-Ferreiro, S., González, E.V., Clari, V.R., Bastiaanse, R., 2017. Quantifying connected discourse in spanish-speaking individuals with aphasia: the case of mixed aphasias. J. Neurolinguistics 44, 38–53. https://doi.org/10.1016/j.jneuroling.2017.03.001.
McNeil, M.R., Hashi, M., Southwood, H., 1994. Acoustically derived perceptual evidence for coarticulatory errors in apraxic and conduction aphasic speech production. Clinical Aphasiology 22, 203–218.
Menn, L., Obler, L.K., Miceli, G. (Eds.), 1990. Agrammatic Aphasia: a cross-language Narrative Sourcebook, 2. John Benjamins Publishing.
Moreno-Torres Sánchez, I., 2005. Agrammatism as feature degradation. Retrieved from. http://www.fe2.uma.es/personal/imoreno/imoreno.htm.
Olbertz, H., 1998. Verbal Periphrases in a Functional Grammar of Spanish, 22. Gruyter.
Olness, G.S., 2006. Genre, verb, and coherence in picture-elicited discourse of adults with aphasia. Aphasiology 20 (02–04), 175–187. https://doi.org/10.1080/02687030500472710.
Peña-Casanova, J., Quiñones-Úbeda, S., Quintana-Aparicio, M., Aguilar, M., Badenes, D., Molinuevo, J.L., et al., 2009. Spanish multicenter normative studies (NEURONORMA Project): norms for verbal span, visuospatial span, letter and number sequencing, trail making test, and symbol digit modalities test. Arch. Clin. Neuropsychol. 24 (4), 321–341. https://doi.org/10.1093/arclin/acp038.
Piazza, G., Calabria, M., Semenza, C., Poletto, C., 2022. Processing of semantic and grammatical gender in Spanish speakers with aphasia. Aphasiology 36 (8), 940–961. https://doi.org/10.1080/02687038.2021.1924355.
Pohlert, T., 2024. PMCMRplus: calculate pairwise multiple comparisons of mean rank sums extended. R package version 1.9.12. https://CRAN.R-project.org/package=PMCMRplus.
Prins, R., Bastiaanse, R., 2004. Analyzing the spontaneous speech of aphasic speakers. Aphasiology 18 (12), 1075–1091. https://doi.org/10.1080/02687030444000534.
Pritchard, M., Dipper, L., Morgan, G., Cocks, N., 2015. Language and iconic gesture use in procedural discourse by speakers with aphasia. Aphasiology 29 (7), 826–844. https://doi.org/10.1080/02687038.2014.993912.
R Core Team, 2023. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Austria. Retrieved from. https://www.R-project.org/.
Real Academia Española de la Lengua, 2009. Nueva Gramática de la Lengua Española. Espasa Calpe. https://www.rae.es/gram%C3%A1tica/.
Rezaii, N., Mahowald, K., Ryskin, R., Dickerson, B., Gibson, E., 2022. A syntax–lexicon trade-off in language production. Proc. Natl. Acad. Sci. 119 (25), e2120203119. https://doi.org/10.1073/pnas.2120203119.
Sá-Leite, A.R., Simpson, I.C., Fraga, I., Comesaña, M., 2024. A standardized approach to measuring gender transparency in languages. Acta Psychol. 246, 104236. https://doi.org/10.1016/j.actpsy.2024.104236.
Sahraoui, H., Nespoulous, J.L., 2012. Across-task variability in agrammatic performance. Aphasiology 26 (6), 785–810. https://doi.org/10.1080/02687038.2011.650625.
Sajjadi, S.A., Patterson, K., Tomek, M., Nestor, P.J., 2012. Abnormalities of connected speech in the non-semantic variants of primary progressive aphasia. Aphasiology 26, 1219–1237. https://doi.org/10.1080/02687038.2012.710318.
Schnur, T.T., Wang, S., 2024. Differences in connected speech outcomes across elicitation methods. Aphasiology 1–22. https://doi.org/10.1080/02687038.2023.2239509.
Sherratt, S., 2007. Multi-level discourse analysis: a feasible approach. Aphasiology 21 (3–4), 375–393. https://doi.org/10.1080/02687030600911435.
Silbert, L.J., Honey, C.J., Simony, E., Poeppel, D., Hasson, U., 2014. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proc. Natl. Acad. Sci. 111 (43), E4687–E4696.
Smith Watts, A.K., Ahern, D.C., Jones, J.D., Farrer, T.J., Correia, S., 2019. Trail-making test part B: evaluation of the efficiency score for assessing floor-level change in veterans. Arch. Clin. Neuropsychol. 34 (2), 243–253. https://doi.org/10.1093/arclin/acy025.
Stark, B.C., 2019. A comparison of three discourse elicitation methods in aphasia and age-matched adults: implications for language assessment and outcome. Am. J. Speech Lang. Pathol. 28 (3), 1067–1083. https://doi.org/10.1044/2019_AJSLP-18-0265.
Stark, B.C., Dutta, M., Murray, L., Bryant, L., Fromm, D., MacWhinney, B., Ramage, A.E., Roberts, A., den Ouden, D.B., Brock, K., McKinney-Bock, K., Paek, E.J., Harmon, T. G., Yoon, S.O., Themistocleous, C., Yoo, H., Aveni, K., Gutierrez, S., Sharma, S., 2021a. Standardizing assessment of spoken discourse in aphasia: a working group

with deliverables. Am. J. Speech Lang. Pathol. 30 (1S), 491–502. https://doi.org/10.1044/2020_AJSLP-19-00093.
Stark, B.C., Dutta, M., Murray, L.L., Fromm, D., Bryant, L., Harmon, T.G., Ramage, A.E., Roberts, A.C., 2021b. Spoken discourse assessment and analysis in Aphasia: an international survey of current practices. J. Speech Lang. Hear. Res.: JSLHR (J. Speech Lang. Hear. Res.) 64 (11), 4366–4389. https://doi.org/10.1044/2021_JSLHR-20-00708.
Stark, B.C., Fukuyama, J., 2020. Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. Language, Cognition and Neuroscience 36 (5), 562–585. https://doi.org/10.1080/23273798.2020.1862258.
Stecher, C., Sánchez, M.E., Carden, J.R., Jaichenco, V.I., 2024. Tense and agreement processing in native Spanish speakers with aphasia. Ment. Lex. 19 (1), 25–36. https://doi.org/10.1075/ml.24019.ste.
Stump, G., Finkel, R.A., 2015. The complexity of inflectional systems. Linguistics Vanguard 1 (1), 101–117. https://doi.org/10.1515/lingvan-2014-1007.
Thompson, C.K., 2003. Unaccusative verb production in agrammatic aphasia: the argument structure complexity hypothesis. J. Neurolinguistics 16 (2–3), 151–167. https://doi.org/10.1016/S0911-6044(02)00014-3.
Ulatowska, H.K., Allard, L., Chapman, S.B., 1990. Narrative and procedural discourse in aphasia. In: Joanette, Y., Brownwell, H.H. (Eds.), Discourse Ability and Brain Damage: Theoretical and Empirical Perspectives. Springer-Verlag, pp. 180–198.
Ulatowska, H.K., North, A.J., Macaluso-Haynes, S., 1981. Production of narrative and procedural discourse in aphasia. Brain Lang. 13 (2), 345–371. https://doi.org/10.1016/0093-934X(81)90100-0.
Van Boxtel, W.S., Cox, B.N., Keen, A., Lee, J., 2023. Planning sentence production in aphasia: evidence from structural priming and eye-tracking. Front. Lang. Sci. 2, 1175579. https://doi.org/10.3389/flang.2023.1175579.
Van Leer, E., Turkstra, L., 1999. The effect of elicitation task on discourse coherence and cohesion in adolescents with brain injury. J. Commun. Disord. 32 (5), 327–349. https://doi.org/10.1016/S0021-9924(99)00008-8.
Vermeulen, J., Bastiaanse, R., Van Wageningen, B., 1989. Spontaneous speech in aphasia: a correlational study. Brain Lang. 36 (2), 252–274. https://doi.org/10.1016/0093-934X(89)90064-3.
Wickham, H., 2016. *ggplot2: Elegant Graphics for Data Analysis* [R package]. Springer-Verlag, New York, ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.
World Health Organization, 2018. International classification of functioning, disability and health. http://www.who.int/classifications/icf/en/.
Wright, H., Capilouto, G.J., 2009. Manipulating task instructions to change narrative discourse performance. Aphasiology 23 (10), 1295–1308. https://doi.org/10.1080/02687030902826844.
Wright, H., Capilouto, G.J., 2012. Considering a multi-level approach to understanding maintenance of global coherence in adults with aphasia. Aphasiology 26 (5), 656–672. https://doi.org/10.1080/02687038.2012.676855.
Yngve, V.H., 1960. A model and an hypothesis for language structure. Proc. Am. Phil. Soc. 104 (5), 444–466. http://www.jstor.org/stable/985230.