Journal of Neurolinguistics 79 (2026) 101343

Contents lists available at ScienceDirect

Journal of Neurolinguistics

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

Research paper

Standardization of the MultiLevel discourse Analysis (MLA) and identification of data-driven age bands for research on language in healthy aging

A. Marini *, F. Petriglia, G. Gasparotto, S. D'Ortenzio, S. Andreetta, M. Gobbo

Cognitive Neuroscience Laboratory, University of Udine, Via Margreth, 3, 33100, Udine, Italy

ARTICLE INFO

Keywords:
Discourse production
Aging
Multilevel procedure of discourse analysis
Cognition

ABSTRACT

Examining specific linguistic aspects in isolation, traditional language tests fail to capture the dynamic interactions that support discourse production abilities. The accurate assessment of discourse production is therefore crucial for identifying language difficulties and procedures of discourse analysis have emerged as a valid methodological solution. Despite this, heterogeneity in discourse measures limits comparability across studies, and the lack of normative data across the adult lifespan complicates the differentiation between healthy and pathological aging. The present study addresses this issue by providing the first standardization of linguistic measures extracted using a MultiLevel procedure of discourse Analysis (MLA). Narrative samples from 717 healthy Italian-speaking adults (aged 20 - 94) were elicited through a picture description task using two single images and three vignettes. Speech samples were transcribed and analyzed using a semi-automatic pipeline. Normative data, adjusted for age and education, and data-driven age bands were calculated for linguistic measures assessing productivity, lexical difficulties, grammatical construction, macrolinguistic difficulties, and lexical informativeness. Results provide standardized norms across multiple linguistic measures and reveal distinct age-related shifts in performance. Together, these findings offer the most comprehensive adult lifespan framework to date for narrative discourse production and highlight the importance of data-driven age bands for research and clinical assessment in healthy aging. Furthermore, they show that healthy aging disproportionately affects higher-order integrative discourse mechanisms rather than core lexical and morphosyntactic encoding processes.

1. Introduction

Language processing relies on a widespread neural network (Indefrey, 2012) and manifests through both micro- and macrolinguistic processes (Marini et al., 2005). Microlinguistic processes support the organization of phonemes into morphemes and words (lexical processing), and their integration into sentences (grammatical processing). Macrolinguistic processes, in turn, enable the formulation and interpretation of communicative intentions and the contextualization of meanings derived from microlinguistic operations (pragmatic processing). They also encompass discourse processing. In this study, discourse is defined as a meaningful unit of language that extends beyond the boundaries of a single sentence and is central to everyday interactions (Dipper et al., 2021).

* Corresponding author., Cognitive Neuroscience Laboratory, University of Udine, Via Margreth, 3, 33100, Udine, Italy.
E-mail address: andrea.marini@uniud.it (A. Marini).

https://doi.org/10.1016/j.jneuroling.2026.101343
Received 28 December 2025; Received in revised form 27 February 2026; Accepted 8 April 2026
Available online 14 April 2026

Therefore, discourse processing refers to the ability to connect propositions across sentences using cohesive and coherent links, thereby constructing a mental model of a narrative (Johnson-Laird, 1983) or building a scenario through scene construction (Buckner & Carroll, 2007).

Accurately assessing language production to characterize the linguistic difficulties of patients with acquired brain lesions is of major clinical importance. Traditional assessments typically evaluate discrete language components in isolation and therefore fail to capture the dynamic interplay between micro- and macrolinguistic abilities involved in the various stages of language production (Dipper et al., 2021; Indefrey & Levelt, 2000). For example, naming tasks provide insight into lexical selection and the retrieval of a word's morphological form, syllabic structure, and articulatory plan, yet they do not allow clinicians to pinpoint difficulties within specific stages of this process (e.g., Harry & Crowe, 2014). Verbal fluency tasks, in turn, are assumed to engage executive functions (Aita et al., 2019), lexical knowledge (Kavé & Yafé, 2014), and mechanisms of lexical access and selection (Gordon et al., 2018; Pekkala, 2012). However, much like naming tasks, fluency measures do not enable the identification of impairments at distinct levels of lexical production (such as lemma retrieval, morphological and syllabic encoding, articulation, or self-monitoring; Weiss et al., 2006) and yield results with limited ecological validity. Similar limitations apply to sentence completion and generation tasks (Cupit et al., 2016). Another drawback of traditional approaches is that they provide limited information about a speaker's informativeness (i.e., the ability to produce contextually meaningful content) and their capacity to structure discourse conceptually. Moreover, these tasks are usually administered in highly artificial testing contexts: naming pictures or completing sentences bears little resemblance to how language is used in everyday communication. Consequently, this fragmented approach reduces ecological validity and fails to reflect patients' actual communicative difficulties. For instance, individuals with traumatic brain injury may score within the normal range on naming tasks yet still exhibit anomia or slowed lexical retrieval during spontaneous discourse, where linguistic and cognitive demands are substantially greater (Marini et al., 2017).

Over the past two decades, discourse analysis has emerged as a valuable method for assessing the communicative and linguistic abilities of patients (see Marini, 2022 for an overview). Compared to traditional language tasks, analyses of spoken discourse offer a more ecologically valid means of evaluating language skills in both healthy individuals and those with linguistic impairments (Bryant et al., 2016; Kong et al., 2025; Marcotte et al., 2024; Prins & Bastiaanse, 2004; Stark et al., 2021). In their review Bryant et al. (2016) reported that, across 156 studies employing discourse analysis, researchers extracted 536 distinct measures to characterize speakers' linguistic performance. These measures can be broadly grouped into three domains: language productivity (e.g., fluency, lexical retrieval), informational content (e.g., informativeness, cohesion), and grammatical complexity (e.g., morphological and syntactic structures). The substantial variability across studies highlights the lack of uniformity and standardization within the literature using discourse analysis, an issue that limits comparability across findings and reduces generalizability across research contexts and clinical groups. Discourse itself can be elicited using a variety of tasks, including story retelling (Saffran et al., 1989), procedural descriptions (Ulatowska et al., 1983), personal event recounts (Glosser & Deser, 1990), and single-picture or cartoon-based narrative tasks (Nicholas & Brookshire, 1993). Bryant et al. (2016) noted that single-picture descriptions are the most used elicitation method. However, they tend to prompt descriptive rather than truly narrative samples, and the resulting productions are often brief. This poses a limitation, as Brookshire and Nicholas (1994) recommend that discourse samples consist of at least 300 words to obtain reliable analyses. One way to address this issue is to combine multiple picture stimuli, provided that they are carefully selected to evoke comparable forms of narrative discourse. Useful options include single images designed to generate narrative content rather than simple description (e.g., the *Cookie Theft* picture from the Boston Diagnostic Aphasia Examination; Goodglass et al., 2001) as well as cartoon-picture stories like the *Flowerpot* (Huber & Gleber, 1982) and the *Quarrel* (Nicholas & Brookshire, 1993). Such discourse-based assessments offer rich insight into the interplay between micro- and macrolinguistic processes and how these interact with cognitive functions across the various stages of message production.

Given its usefulness, a growing body of research has focused on developing quantitative procedures for narrative discourse analysis (e.g., Berube et al., 2019; Boucher et al., 2022; Kong et al., 2025; Marini, Andreetta, et al., 2011). Nonetheless, normative data for narrative discourse production across the adult lifespan remain limited. Existing norms are often based on a narrow set of measures, small samples, and broad or poorly specified age ranges. As a result, clinicians and researchers may identify anomalous discourse patterns without a solid quantitative basis for distinguishing age-appropriate changes from early signs of pathology. This gap is particularly problematic in the context of healthy aging. A growing literature suggests that aging has a heterogeneous impact on language, with narrative discourse remaining relatively stable through most of adulthood but showing progressive decline in later life (Juncos-Rabadán et al., 2005; Marini & Andreetta, 2016; Marini et al., 2005; Pistono et al., 2017; Thornton & Light, 2006). Recent studies applying the MultiLevel procedure of discourse Analysis (MLA; Marini, Andreetta, et al., 2011) to narrative discourse samples produced by adults aged 20 to 90 have revealed specific age-related effects on narrative production. These findings indicate both linear and non-linear trends across different aspects of discourse: older adults produced narratives containing more semantic errors and demonstrating lower coherence and informativeness (Hilviu et al., 2025; Marini et al., 2025), reduced grammatical organization (D'Ortenzio et al., 2025), and diminished lexical variety (Petriglia et al., 2025). Moreover, individuals in their 70s and 80s showed a reduced capacity to express personal viewpoints and emotional content during narrative production (Gallo et al., 2025). Without robust normative benchmarks, however, it remains difficult to determine when such changes reflect typical age-related reorganization and when they signal emerging neurodegenerative disease or other clinical conditions. The same challenge applies to the correct characterization of discourse production difficulties often observed in individuals with acquired brain injuries. The MLA has proven effective in delineating the linguistic characteristics of individuals with hyperacute and chronic post-stroke aphasia, both for diagnostic assessment (Andreetta et al., 2012, 2025) and for tracking treatment outcomes (Marangolo et al., 2013). It has also been successfully employed to characterize the linguistic impairments of patients with focal right-hemisphere lesions (Marini, 2012), traumatic brain injury (Marini, Galetto, et al., 2011), mild cognitive impairment (Mazzon et al., 2019), and schizophrenia (Marini

et al., 2008), highlighting potential interconnections between micro- and macrolinguistic difficulties.

A further source of uncertainty arises from the way age is operationalized in discourse research. Many studies treat age as a continuous variable or rely on broad, sometimes arbitrary, groupings (e.g., "young" vs. "elderly"). Others employ decade-based bands (20s, 30s, 40s, etc.) but lack the sample sizes needed to generate reliable norms for each group or to examine non-linear trajectories. Yet available evidence indicates that age-related change in discourse is rarely linear (Marini et al., 2025): some measures remain stable well into late midlife and show decline only in the 70s or 80s, whereas others exhibit earlier or more gradual shifts. Collapsing across wide age spans or imposing *a priori* cut-offs may therefore obscure meaningful patterns and reduce the sensitivity of discourse measures in both research and clinical contexts. What is missing, therefore, is a comprehensive data-driven characterization of narrative discourse across the adult lifespan based on a) a large sample, b) multiple indices assessing both micro- and macrolinguistic dimensions of processing, and c) analytical methods that determine the shape of age effects empirically rather than assuming it. In particular, there is a need for: 1) Normative values for key components of narrative discourse production, such as productivity (e.g., total words, speech rate), lexical skills, syntactic accuracy, macro-linguistic organization (e.g., cohesion and coherence), and informativeness, covering healthy adults from early adulthood to very old age; 2) Education-adjusted norms, given that years of schooling modulate language and discourse performance and may act as a form of cognitive reserve; 3) Data-driven age bands derived from statistical modelling (e.g., segmented regression) that identify points at which the relationship between age and performance changes slope. Such empirically derived bands can complement traditional decade-based groupings by highlighting functionally distinct phases (e.g., preserved performance, early decline, marked decline) that are particularly relevant for future studies of healthy and pathological aging. Beyond its normative relevance, a comprehensive dataset that includes several measures of discourse production offers an opportunity to address theoretical questions concerning the architecture of language production in aging. Contemporary models of discourse production assume that language output emerges from the interaction of multiple processing levels, ranging from lexical retrieval and morpho-syntactic encoding to higher-order semantic integration, executive control, and situation model construction (Indefrey & Levelt, 2000; Marini et al., 2005). However, it remains unclear whether age-related changes affect these levels uniformly or whether higher-level integrative mechanisms are more vulnerable than lower-level structural processes. Previous research has suggested that word retrieval difficulties in older adults may diminish in connected speech due to contextual support (Kavé & Goral, 2017), whereas reduced discourse coherence may reflect impairments in semantic selection and executive regulation rather than purely linguistic decline (Hoffman et al., 2018). A comprehensive dataset covering the entire adult lifespan that examines productivity, lexical errors, grammatical structure, coherence, and informativeness simultaneously provides a unique opportunity to test whether aging primarily impacts lower-level encoding mechanisms or disproportionately affects higher-level integrative and control-dependent processes.

The present study was designed to address these gaps. We elicited narrative discourse samples administering a picture description task made of two single images and three cartoon-picture stories, to a cohort of more than 700 Italian-speaking adults stratified by age (20 to 94 years) and education (from middle school to PhD level). Speech samples were analyzed using the MLA, yielding measures of productivity, lexical difficulties, grammatical construction, macrolinguistic difficulties, and lexical informativeness. For each measure, we first adopted a normative approach: we removed outliers, formed seven decade-based age groups (20s through 80–90s), and examined the effects of age and education on the linguistic variables. This allowed us to generate age- and education-referenced normative values, as well as education-corrected scores based on regression residuals. In a second step, we treated age as a continuous predictor and applied segmented (piecewise) regression to identify data-driven breakpoints in the age–performance relationship. This approach makes it possible to detect non-linear trajectories and to derive empirically grounded age bands that reflect genuine shifts in discourse production, rather than arbitrary chronological cut-offs. We then used these bands to compute descriptive statistics within each segment and to characterize the trajectory of each discourse component during adulthood. To the best of our knowledge, this study represents both the first Italian standardization of a picture description task and the first to implement such standardization using a sample of this size, even in comparison with similar work conducted in other languages (Berube et al., 2019; Boucher et al., 2022; Kong et al., 2025).

2. Materials and methods

2.1. Participants

Seven hundred and seventeen Italian-speaking healthy adults were recruited and divided into seven age-groups: N = 121 aged 20-29; N = 105 aged 30-39; N = 89 aged 40-49; N = 101 aged 50-59; N = 127 aged 60-69; N = 90 aged 70-79; and N = 84 aged 80-94 (see

Table 1
– General characteristics of the participants. Data are presented as means, with standard deviations in parentheses, for each age group. For sex, the percentage of female participants is reported. Legend: MoCA = Montreal Cognitive Assessment; AAT = Aachen Aphasia Test.

	20-29	30-39	40-49	50-59	60-69	70-79	80-94
Age	23.88 (2.54)	33.97 (2.59)	44.38 (3.07)	54.59 (2.83)	63.66 (2.86)	74.22 (2.86)	83.50 (3.41)
Education	15.65 (2.14)	16.35 (3.28)	15.19 (3.92)	14.94 (3.68)	12.91 (4.51)	11.73 (4.40)	8.06 (3.96)
Sex	F = 69 (57%)	F = 61 (58%)	F = 50 (56%)	F = 71 (70%)	F = 70 (55%)	F = 52 (58%)	F = 53 (63%)
MoCA	28.05 (1.66)	28.57 (1.33)	28.10 (1.91)	27.15 (1.84)	27.47 (1.93)	25.35 (2.44)	24.55 (2.71)
Token	4.40 (.51)	5.0 (.00)	4.96 (.19)	4.99 (.08)	4.91 (.54)	4.93 (.17)	4.96 (.17)
Naming_AAT	117.20 (2.64)	119.31 (1.35)	119.10 (2.02)	118.30 (2.03)	118.27 (2.29)	115.88 (3.42)	114.76 (4.89)

Table 1 for demographic and general characteristics of the groups). Participants were recruited from different areas of Italy. The majority (N = 484, 68%) were from Northern regions, 162 (23%) from Central regions, and 64 (9%) from Southern regions. Seven participants lived abroad. Of the total sample, 499 participants (69.6%) reported being monolingual, 142 (19.8%) reported being bilingual, and 76 (10.6%) did not provide information regarding language background. All participants were recruited in a project aimed at standardizing the MLA for adults. Parts of this database were used in previous studies assessing the effects of aging on discourse-production skills (e.g., Marini et al., 2025). Participants were recruited through Lifelong Learning Institutes, local social or sport clubs, voluntary and charitable associations, and personal contacts, using informational flyers and advertisements posted on social networks. Inclusion criteria required the absence of neurological or neuropsychiatric disorders and performance above the cutoff on the Italian version of the Montreal Cognitive Assessment (MOCA; Conti et al., 2015), the Naming task of the Aachener Aphasie Test (Luzzatti et al., 1996), and the short version of the Token Test (De Renzi & Vignolo, 1962).

The age-groups did not differ in gender distribution with a small effect size measured by Cramérs V [X^2 (6, N = 717) = 7.523, p = .298, Cramérs V = .101]. A group-related difference was found for years of education, F (6, 710) = 55.248, $p < .001$, $\eta^2 = .318$. Tukey's post-hoc analyses confirmed that persons in their 60s, 70s, and 80s had fewer years of education than all younger groups ($p < .001$) and those in their 70s and 80s had fewer years of education than those in their 60s ($p < .001$).

Across the seven age-groups, participants produced an average of 521.63 words with a standard deviation of 284.80. As shown in Table 2, all groups exceeded the 300-word threshold considered necessary for high test-retest stability in clinical settings (Brookshire & Nicholas, 1994).

This study was not preregistered. The Ethical Committee of the University of Udine approved the study (Protocol CGPER-2024-02-27-01). All participants signed an informed consent form and provided written permission to participate in this study.

3. Data availability

The datasets generated during and/or analyzed during the current study are not publicly available for privacy or ethical restrictions; however, they will be available from the corresponding author upon reasonable request.

3.1. Assessment of narrative discourse production

Narrative assessment was carried out on speech samples elicited with five picture stimuli: two single-picture scenes (the *Picnic* scene from the Western Aphasia Battery by Kertesz [1982] and the *Cookie Theft* scene by Goodglass and Kaplan [1972]), and three cartoon-picture sequences (the *Flowerpot* by Huber and Gleber [1982], the *Quarrel* by Nicholas and Brookshire [1993], and the *Nest Story* by Paradis and Libben [1987]). The stimuli were colored and presented in random order on a laptop, with the screen facing the participant to prevent referent sharing with the examiner. Participants read the following on-screen instructions: "You will see stories in the form of images. The images are randomly selected by the program, so I don't know them. As soon as the image appears, describe it to me. There is no time limit and no right or wrong way to describe them. You can speak as much or as little as you like. However, avoid using words like 'here,' 'this,' etc., and be as clear as possible." All narratives were audio-recorded.

Speech samples were transcribed using a semi-automatic pipeline developed by one of the authors (F.P.). To streamline transcription and subsequent analysis, an automated Python workflow was implemented and hosted on Google Colab. The Whisper large-v3 model (Radford et al., 2023) was used to generate initial text transcriptions from the narrative audio recordings. Although Whisper is a state-of-the-art speech-to-text system, it tends to omit repetitions, reformulations, and false starts. The raw output was therefore processed with the Spacy Python library (Honnibal et al., 2020) to assign Part-of-Speech labels to every word in the transcript. Following the automatic labeling step, omitted material such as false starts, repetitions, and reformulations was manually reintegrated by the operator through direct comparison with the original audio recordings. The resulting files were plain text (.txt) documents structured in a format similar to the CHAT transcription conventions (MacWhinney, 2000). A sample of 10 analyzed transcripts is available here: https://osf.io/29kva/overview?view_only=e81cccc3ab254e9ca4fe043cc429e9bb.

Final transcripts included the duration (in seconds) of each story, as well as annotations for phonological fillers, pauses, false starts, phonological errors, and neologisms. For each story, utterance segmentation followed the criteria detailed in Marini, Andreetta, et al. (2011). Because no single criterion is sufficient to reliably segment spontaneous speech, we jointly applied acoustic, semantic, grammatical, and phonological criteria. According to the **acoustic criterion**, utterances are separated by clearly perceptible pauses, whether silent (e.g., silence) or filled (e.g., emissions such as "ehm" or fillers such as "I think" or "I don't know"). For example, in the sequence "this is a … [3-s silent pause] man," the clear pause between "this is a" and "man" justifies segmentation into two utterances:/This is a … (5 s)/child/. When pauses are not evident, the **semantic criterion** is used, defining an utterance as a conceptually coherent proposition, such that boundaries are marked when one proposition ends (or is left suspended) and another begins. For instance, "The dog is on the sidewalk with the man. A flower pot falls on the man's head" is segmented into two utterances:/The dog is on the sidewalk with the man/A flower pot falls on the man's head/, as the second clause introduces a new proposition. Similarly, in "A man is walking on/He is running on the sidewalk," the first proposition is incomplete, and the second segment reformulates it; therefore, two distinct utterances are identified. The **grammatical criterion** identifies an utterance as a grammatically complete sentence, including subordinate clauses (e.g.,/The man is walking on the sidewalk with a dog that looks very nice/), while coordinated clauses may be segmented if they constitute distinct syntactic units. (/The man is walking on the sidewalk/and a dog is following him/). Finally, under the **phonological criterion**, false starts (i.e., interrupted words) signal the end of an utterance, as these interruptions mark a disruption in speech flow as in the following sequence:/and she is ca-/stroking his d-/his d-/the dog of the man/. Following utterance segmentation, the narratives were analyzed for productivity (i.e., number of words produced and speech rate),

Table 2
– Global narrative performance of the seven groups of participants. Data are presented as means, with standard deviations in parentheses, for the total of the five stories for each age group.

	20-29	30-39	40-49	50-59	60-69	70-79	80-94
Words (total)	554.9 (250.11)	577.35 (216.92)	581.28 (287.30)	514.68 (243.01)	463.29 (226.88)	462.21 (246.92)	340.8 (154.46)
Speech rate	136.37 (22.86)	136.1 (22.58)	134.1 (24.28)	131.86 (24.47)	129.93 (24.53)	123.63 (20.73)	118.83 (21.91)
% Phonological errors	.67 (.54)	.65 (.55)	.70 (.61)	.73 (.58)	.83 (.71)	1.07 (.74)	1.44 (.89)
% Semantic errors	.73 (.84)	.74 (.95)	.90 (1.02)	.84 (.98)	1.38 (1.74)	2.13 (1.99)	4.20 (2.98)
% Morphological errors	1.42 (1.80)	1.96 (1.88)	2.35 (2.58)	2.24 (2.28)	2.80 (3.36)	4.16 (3.34)	4.32 (3.04)
% Complete sentences	73.50 (13.00)	73.90 (11.04)	73.50 (12.05)	72.30 (11.88)	69.28 (13.29)	64.49 (13.59)	59.62 (14.20)
% Cohesion errors	19.35 (9.28)	18.83 (8.61)	18.08 (8.56)	19.26 (9.12)	19.75 (9.76)	21.13 (9.40)	22.43 (9.29)
% Local coherence errors	19.37 (9.57)	17.47 (10.85)	18.99 (9.77)	20.27 (8.81)	26.04 (12.68)	33.68 (15.78)	44.58 (20.20)
% Global coherence errors	5.78 (4.41)	6.94 (4.89)	8.02 (6.08)	7.42 (5.60)	7.5 (6.29)	11.23 (6.78)	12.99 (7.12)
% Lexical informativeness	85.99 (5.80)	85.36 (6.18)	83.69 (6.88)	85.09 (6.27)	84.64 (7.07)	77.51 (9.29)	74.46 (8.92)

microlinguistic abilities (i.e., lexical and grammatical accuracy), and macrolinguistic efficiency (i.e., discourse organization and informativeness) (Marini, Andreetta, et al., 2011). The linguistic measures were selected based on the theoretical architecture of the MLA, which conceptualizes discourse production as involving partially dissociable processing stages. Productivity measures (i.e., total words and speech rate) index lexical production quantity and efficiency. Lexical errors reflect distinct levels of lexical encoding. Specifically semantic errors reflect difficulties in lexical selection whereas morphological and phonological errors reflect different stages of lexical access (i.e., to morphological and phonological pieces of information stored in semantic memory). Syntactic completeness captures sentence-level grammatical structuring. Cohesion errors reflect the correct deployment of formal linguistic devices linking adjacent utterances, whereas local and global coherence errors index conceptual integration across propositions and alignment with the overarching narrative model. Finally, lexical informativeness integrates micro- and macrolinguistic processes by quantifying the proportion of contextually appropriate and pragmatically relevant lexical output. Together, these measures allow the examination of age-related changes across multiple stages of language production, from lower-level encoding to higher-order integrative mechanisms.

Productivity was indexed by the mean number of phonologically well-formed speech units (i.e., words). These words were also used to derive a measure of lexical selection efficiency, namely speech rate (words per minute ([number of words/time in minutes]).

Lexical accuracy was evaluated by computing percentages of phonologic, morphological, and semantic errors. Phonological errors comprised false starts, phonological paraphasias, and neologisms. These speech units were not counted as phonologically well-formed words. A false start was defined as an incomplete attempt to produce the target word (*fath- *fa- instead of "father"). Phonological paraphasias were words whose phonological form deviated from the target due to the inversion, omission, insertion, or substitution of phonemes (e.g., *fathir, *ather, or *farther, instead of "father"). Neologisms were non-recognizable word forms (e.g., *gerat instead of "father"). The percentage of phonological errors was calculated by dividing the number of such errors by the total number of speech units and multiplying by 100. Morphologic errors included substitutions of either morphemes in a word (for example, "questo è una coppia" "this [masc in Italian] is a couple [fem]" – in Italian "questo" should be "questa") or function words (for example, "batte da una porta" "he is knocking from a door" – in Italian "da" instead of "a"). The percentage of morphologic errors was calculated by dividing the number of such errors by the number of utterances and multiplying this value by 100. Semantic errors occurred when the speaker substituted the target word with a semantically related item (e.g., "tree" instead of "flower"). The percentage of semantic errors was obtained by dividing the number of semantic errors by the total number of words and multiplying by 100.

Grammatical accuracy was assessed through the percentage of complete sentences, defined as grammatically well-formed sentences without omissions or morphological errors. This percentage was calculated by dividing the number of complete sentences by the total number of utterances and multiplying by 100.

Discourse organization was examined by computing percentages of cohesion errors and local and global coherence errors. Cohesion referred to the appropriate use of linguistic devices to link consecutive utterances. Cohesion errors included the use of function words that failed to correctly connect two utterances or the abrupt interruption of an utterance that was then completed in the following one (e.g., "The man is … …/He is walking down the street"). The percentage of cohesion errors was calculated by dividing the number of such errors by the total number of utterances and multiplying by 100. Local coherence captured the conceptual relatedness of adjacent utterances. Local coherence errors occurred when words lacked a clear referent or when an utterance was abruptly interrupted and its content was not resumed in a subsequent utterance (topic shift; e.g., "The man is …/A flowerpot falls on his head"). The percentage of local coherence errors was obtained by dividing the number of these errors by the total number of utterances and multiplying by 100. Global coherence concerned the extent to which utterances are semantically aligned with the overall narrative. Global coherence errors included tangential or unrelated, as well as filler or repeated utterances. Tangential utterances deviated from the target information (e.g., "The man is walking down the street with a hat/Hats can be very useful when it rains"). Semantically unrelated utterances introduced content inconsistent with the story (e.g., "The man is walking down the street/an airplane crashes behind him"). Filler utterances did not add new information (e.g., "The man is waking down the street/Shall I go on?"), and repeated utterances restated already conveyed content (e.g., "The man is waking down the street/He is on the street"). The percentage of global coherence errors was calculated by dividing the number of such errors by the total number of utterances and multiplying by 100.

Finally, we computed a functional measure of communicative informativeness: the percentage of lexical informativeness. This was obtained by dividing the number of informative words by the total number of words and multiplying by 100. Informative words included phonologically well-formed words that were appropriate from both grammatical and pragmatic perspectives. Words were not considered informative if they were classified as phonological, semantic, or morphological errors, repetitions, fillers, lacked a clear referent or occurred within tangential or unrelated utterances. However, error-free words in utterances containing cohesion of local coherence errors were still counted as informative. For example, in the sequence "An old man with a hat and a cat was …/He was walking on the street, right?/I assume he is a nice person/" the underlined words would be scored as informative. In contrast, "He was" at the beginning of the second utterance would not, as it is a repetition. "Cat" would be scored as a semantic paraphasia if the target was "dog" and therefore not informative. The word "right" at the end of the second utterance and all words in the third utterance would be coded as fillers and excluded from the informative words count. The percentage of lexical informativeness thus indexes the speaker's ability to retrieve and produce contextually appropriate target words within the narrative.

3.2. Interrater reliability

Scoring was carried out by two expert raters (F.*P. and* G.G.) under the supervision of A.M. Both raters were blinded to the participants' age groups and preliminarily analyzed 685 narratives (five stories for 137 participants). Interrater reliability was assessed using both Kappa statistic (Carletta, 1996) and Intra-class correlation (ICC). Agreement between the two raters was almost perfect

across all measures (see Marini et al., 2025): Words (Kappa = .96; ICC (2,1) = 1.00); Speech rate (Kappa = .96; ICC (2,1) = 1.00); % Phonological errors (Kappa = .99; ICC (2,1) = 1.00); % Semantic errors (Kappa = .87; ICC (2,1) = .99); % Complete sentences (Kappa = .99; ICC (2,1) = 1.00); % Cohesion errors (Kappa = .99; ICC (2,1) = 1.00); % Local coherence errors (Kappa = .85; ICC (2,1) = 1.00); % Global coherence errors (Kappa = .99; ICC (2,1) = .99); % Lexical informativeness (Kappa = .78; ICC (2,1) = .99).

3.3. Statistical analyses

Analyses leading to normative values were conducted using JASP, version .95.4. All other analyses were conducted in R (R Core Team) version 4.5.2 using the packages *segmented* (Muggeo, 2008), *dplyr* (Wickham, François, Henry, Müller, & Vaughan, 2023), *readr* (Wickham, Hester, & Bryan, 2023), and *haven* (Wickham et al., 2023). Statistical procedures had two complementary aims: 1) to provide normative, decade-based reference values for each discourse measure; and 2) to identify data-driven age bands by modelling age as a continuous predictor and estimating points at which the relationship between age and performance changed.

For the computation of normative, decade-based reference values, outliers were removed prior to analyses for each discourse measure using a ±3 SD criterion based on z-scores computed from raw values in each decade band. This procedure reduced the influence of extreme observations and ensured that normative estimates reflected typical performance. After outlier removal, we computed descriptive statistics (N, mean, standard deviation) for each measure. Skewness and kurtosis were also calculated to describe the shape of their distributions. Following Kline (2011), skewness values within ±3 and kurtosis values within ±10 are considered consistent with approximate normality. Given the large sample size (N = 717), these indices were used for descriptive purposes only, as even minor deviations from a Gaussian distribution can inflate skewness and kurtosis in large datasets (N > 200; Tabachnick & Fidell, 2001). Distributional assumptions were therefore evaluated through visual inspection of histograms and Q–Q plots, as well as by examining model residuals. Parametric analyses were retained, as such methods are robust to moderate non-normality in large samples. Linear multiple regressions were conducted to assess the influence of chronological age and years of formal education on each discourse variable. The unstandardized residuals from these regressions were used to compute education-corrected normative values. For each measure, we tabulated raw scores by decade and provided education-based correction grids derived from the regression models (see supplementary tables). Next, to examine age-group differences and the shape of age-related change, when education had a significant effect in the regression models we ran ANCOVAs with age group as the independent variable, education as the covariate, and the discourse measures as dependent variables. We tested the main effect of age group, education, and the age*education interaction. For measures where education was not a significant predictor, we ran a one-way ANOVA with age group as the independent variable. For each AN(C)OVA, we reported the F-statistic, associated p-value, and effect size (η^2) to quantify the magnitude of age effects. To explore the shape of the age trajectory, polynomial contrasts were applied across age groups. Finally, to identify specific between-group differences, we conducted Tukey's HSD post-hoc tests on models without the education covariate. These pairwise comparisons allowed us to determine which age decades differed from which others for each discourse measure. As a final analysis, to address potential influences of participants' language background and geographic origin, supplementary hierarchical linear regression analyses were conducted for the target variables. In these models, participants' age and years of education were entered in the first step. Geographic area (North, Centre, South) was entered in the second step, and bilingual exposure (monolingual vs. bilingual participants) in the third step. This hierarchical approach allowed us to examine whether regional provenance or bilingualism accounted for additional variance beyond age and education, and whether the inclusion of these variables altered the magnitude or significance of the primary age effects. Changes in explained variance (ΔR^2) and associated F-change statistics were used to evaluate the incremental contribution of each block.

To derive data-driven age bands and formally test for changes in slope in the age–performance relationship, we complemented the decade-based analyses with segmented (piecewise) regression for each discourse index, using the full cohort (N = 717). For each outcome, we first fit a baseline linear model. To test for changes in the age slope, we applied the Davies test implemented in the *segmented* package in R (α = .05). This test evaluates whether there is statistical evidence for at least one change in the slope of the relationship between the target discourse measure and age (with education held constant). Segmented regression models were fit and interpreted only when the Davies test indicated evidence of at least one slope change. When the Davies test was not significant, we retained the single-slope linear model for inference and did not report breakpoints. In addition, for outcomes with a significant Davies test, we required that the segmented model improve fit over the linear model (e.g., lower AIC) before deriving and reporting data-driven age bands.

We then fit segmented regression models with age as the segmented predictor, typically requesting the estimation of three breakpoints corresponding to four age segments. Initial breakpoint guesses were supplied to facilitate model convergence, but the final breakpoint estimates were entirely data driven. To evaluate whether the segmented model provided a better description of the data than the simple linear model, we compared the two models using ANOVAs (likelihood ratio tests) on the residual sum of squares, and we also compared their Akaike Information Criterion (AIC) values for the linear (AIC_linear) and segmented (AIC_segmented) models. A significant F-test together with a lower AIC for the segmented model was taken as evidence that the age trajectory was non-linear and more accurately captured by a piecewise function with one or more breakpoints. For each measure, the set of estimated breakpoints was ordered and used to create empirical age bands.

3.4. Sample size estimation

An *a priori* power analysis was computed using G*Power 3.1 software (Faul et al., 2009) for linear multiple regression (fixed model, R^2 deviation from zero), with age and years of education entered as predictors. Previous research on narrative skills in healthy aging

(Marini et al., 2005) reported effect sizes for age-related group differences ranging from .120 to .570 (Cohen's F). Assuming a medium effect size ($f^2 = .09$), $\alpha = .05$, and desired power of .95, the analysis indicated that a minimum sample of N = 175 participants was required. For the AN(C)OVA design with seven age-groups, education as a covariate, $\alpha = .05$, and desired power of .95, assuming an expected effect size of f = .31 the analysis showed that a minimum sample of N = 220 participants was required.

4. Results

The results are presented in two sections. The first section reports the standardization outcomes for each domain of discourse performance: productivity skills, lexical difficulties, grammatical construction abilities, macrolinguistic difficulties, and lexical informativeness. The second section presents the results of the segmented (piecewise) regressions conducted for each discourse index for the extraction of data-driven age-bands across the full cohort (N = 717). Means and standard deviations for the seven age-groups on all narrative measures are shown in Table 2. Normative values and additional descriptive statistics are provided in Supplementary materials (Tables S1–S10).

5. Section 1 - primary normative analysis of the MLA

5.1. Productivity

For total word production, after removal of outliers 704 participants were retained: N = 119 aged 20-29; N = 101 aged 30-39; N = 87 aged 40-49; N = 98 aged 50-59; N = 126 aged 60-69; N = 89 aged 70-79; N = 84 aged 80-94. Inspection of the Q–Q plot indicated that the distribution of word counts was approximately normal across groups. The linear regression assessing the influence of chronological age and years of formal education on total word count was significant, $F(2, 703) = 62.116$, $p < .001$, $R^2 = .151$. Both age ($b = -1.242$, $p = .013$) and education ($b = 17.729$, $p < .001$) significantly predicted total word production, such that older age was generally associated with fewer words, whereas higher education was associated with more words. The ANCOVA with age group as the independent variable, education as a covariate, and total words as the dependent variable confirmed the presence of age-related differences after controlling for education, $F(6, 696) = 2.143$, $p = .047$, $\eta^2 = .018$, with a significant effect of education ($p < .001$) and no age*education interaction ($p = .131$). Polynomial contrasts showed a significant linear trend across age groups ($p = .004$), consistent with a gradual decline in word production with advancing age rather than a sharp drop at a single decade. Tukey's post-hoc analyses indicated no significant differences in word production among participants in their 20s, 30s, 40s and 50s, suggesting broadly comparable discourse output across early and mid-adulthood. In contrast, individuals in their 60s produced significantly fewer words than participants in their 20s ($p = .039$), 30s and 40s (both ps = .006). No differences emerged between participants in their 50s and those in their 60s or 70s, nor between those in their 60s and 70s. Finally, participants in their 80s produced fewer words than all younger groups (vs. 20s, 30, 40s and 50s: $p < .001$; vs. 60s: $p = .004$; vs. 70s: $p = .013$). Thus, the decade-based norms indicate that word production is relatively stable from the 20s through the 50s, begins to decline in the 60s, and shows a marked reduction in the 80s. Given the effect of education, education-corrected normative values were computed using the unstandardized residuals from the regression model. Tables S1 and S2 present the raw normative values for each age group and the education-based correction values, respectively. Fig. 1 shows the trends in word production for each age-group with z-score corrected values. Further hierarchical regression analyses were conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on word production. Adding geographic area (North, Centre, South) significantly increased explained variance ($\Delta R^2 = .028$, $p < .001$) and inclusion of bilingual exposure contributed a small additional increment ($\Delta R^2 = .010$, $p = .004$). Importantly, however, the magnitude and significance of the age effect remained largely unchanged across models, indicating that the age-related patterns are not attributable to regional or multilingual variation.

Regarding speech rate, after removal of outliers the resulting dataset retained 710 participants: N = 121 aged 20-29; N = 104 aged 30-39; N = 88 aged 40-49; N = 99 aged 50-59; N = 124 aged 60-69; N = 90 aged 70-79; N = 84 aged 80-94. Inspection of Q–Q plots indicated that the distribution of speech rate was approximately normal across groups. The linear regression was significant, F (2, 709)

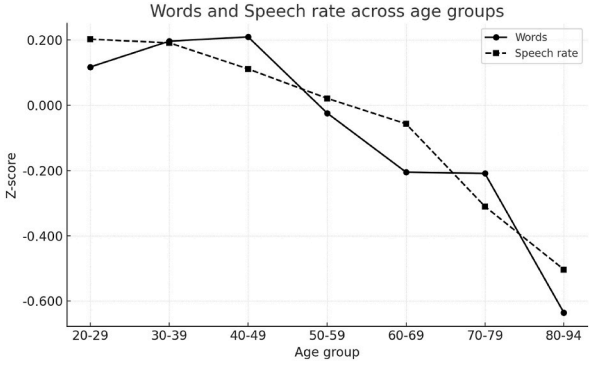

Fig. 1. – Productivity across the seven age-groups with z-score corrected values.

= 31.751, p < .001, R^2 = .082. Both age (b = −.176, p < .001) and education (b = .940, p < .001) emerged as significant predictors, indicating that older participants produced fewer words per minute, whereas more educated participants showed higher speech rates. The ANCOVA with age group as the independent variable, education as a covariate, and speech rate as the dependent variable confirmed significant age-related differences, F (6, 702) = 2.165, p = .044, η^2 = .018, together with a significant main effect of education (p < .001) with no age*education interaction (p = .419). Polynomial contrasts suggested that the age-related change in speech rate followed a predominantly linear trajectory (p < .001). Tukey's post-hoc tests revealed no significant differences in speech rate between participants in their 20s, 30s, 40s, 50s, and 60s. Participants in their 70s produced significantly fewer words per minute than those in their 20s (p = .002), 30s (p = .003), and 40s (p = .040). Finally, participants in their 80s produced fewer words per minute than all younger groups (vs. 20s, 30, and 40s: p < .001; vs. 50s: p = .003; vs. 60s: p = .012). No differences were observed between individuals in their 70s and 80s. Given the effect of education, education-corrected normative values were computed using the unstandardized residuals from the regression model. Tables S3 and S4 present the raw normative values for each age group and the education-based correction values, respectively. Fig. 1 shows the trends in speech rate for each age-group with z-score corrected values. Further hierarchical regression analyses were conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on speech rate. Adding geographic area did not significantly increase explained variance (ΔR^2 = .002, p = .593) and inclusion of bilingual exposure did not contribute additional variance (ΔR^2 = .000, p = .589). These results indicate that the age-related pattern in speech rate is not attributable to regional or multilingual variation.

5.2. Lexical difficulties

Regarding phonological accuracy, after removal of outliers the dataset retained 704 participants: N = 121 aged 20-29; N = 105 aged 30-39; N = 87 aged 40-49; N = 100 aged 50-59; N = 124 aged 60-69; N = 89 aged 70-79; N = 78 aged 80-94. Inspection of the Q–Q plots indicated that the distribution of % phonological errors was bounded at 0 and showed a floor effect, especially in the younger age groups, with many participants producing no errors. A linear regression assessing the influence of chronological age and years of formal education on % phonological errors was significant, F (2, 703) = 45.993, p < .001, R^2 = .116. Both predictors contributed uniquely: age was positively associated with phonological error rates (b = .008, p < .001), whereas education was negatively associated with them (b = −.026, p < .001). Thus, older participants produced a higher percentage of phonological errors, while more educated participants produced fewer errors. An ANCOVA with age-group as the independent variable, % phonological errors as the dependent variable, and education as a covariate confirmed significant age-related differences, F (6, 696) = 7.941, p < .001, η^2 = .064, with a significant effect of education (p = .003) and no age*education interaction (p = .774). Polynomial contrasts indicated that age-related variations in phonological error rates followed both linear (p < .001) and quadratic (p < .001) trends. Tukey's post-hoc analyses revealed no significant differences in % phonological errors among participants in their 20s, 30s, 40s, 50s, and 60s. In contrast, individuals in their 70s produced significantly more phonological errors than those in their 20s (p < .001), 30s (p < .001), 40s (p = .004), and 50s (p = .008). Finally, persons in their 80s produced more phonological errors than all younger groups (vs. 20s, 30, 40s, 50s, and 60s: p < .001; vs. 70: p = .004). Given the influence of education, education-corrected normative values were computed using the unstandardized residuals from the regression model. Tables S5 and S6 report the raw normative values by decade and the education-based correction values, respectively. Fig. 2 shows the trends in phonological errors for each age-group with z-score corrected values. Further hierarchical regression analyses were conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on % phonological errors. Adding geographic area did not significantly increase explained variance (ΔR^2 = .008, p = .061). Inclusion of bilingual exposure, however, significantly increased explained variance (ΔR^2 = .026 p < .001). Importantly, the magnitude and significance of the age effect remained largely unchanged in this model, indicating that the observed age-related patterns are not attributable to multilingual variation.

Regarding semantic accuracy, after removal of outliers the dataset retained 706 participants: N = 121 aged 20-29; N = 105 aged 30-39; N = 89 aged 40-49; N = 101 aged 50-59; N = 127 aged 60-69; N = 86 aged 70-79; N = 77 aged 80-92. Inspection of the Q–Q plots indicated that the distribution of % Semantic errors was bounded at 0 and showed a floor effect, especially in the younger age groups, with many participants producing no errors. A linear regression was conducted to assess the influence of chronological age and years of formal education on % semantic errors. The model was statistically significant, F (2, 705) = 116.527, p < .001, R^2 = .249. Both age and

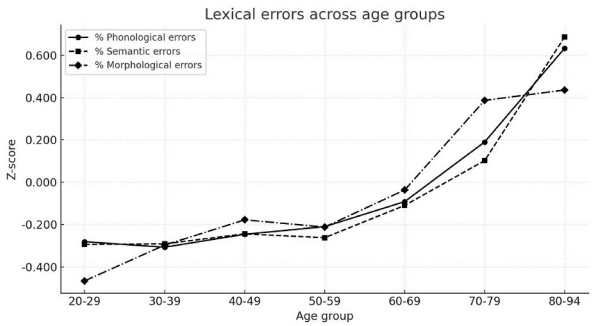

Fig. 2. – Lexical errors across the seven age-groups with z-score corrected values.

education emerged as significant unique predictors: age was positively associated with semantic error rates (b = .033, p < .001), whereas education was negatively associated with such errors (b = −.099, p < .001). Therefore, older participants tended to produce a higher percentage of semantic errors, whereas more educated participants produced fewer such errors. An ANCOVA with age group as the independent variable, education as covariate, and % semantic errors as the dependent variable confirmed robust age-related differences (F (6, 698) = 29.376, p < .001, η^2 = .202) with a significant effect of education (p < .001) and no age*education interaction (p = .771). Polynomial contrasts showed significant linear, quadratic and cubic effects (all ps < .001), indicating that the effect of age on semantic errors was non-linear. Tukey's post-hoc tests showed that participants in their 60s produced more semantic errors than those in their 20s (p = .022) and 30s (p = .036). Furthermore, adults in their 70s and 80s produced more such errors than participants in all younger groups (20s, 30s, 40s, 50s, and 60s; all ps < .001), and those in their 80s also produced more errors than participants in their 70s (p < .001). Given the influence of education, education-corrected normative values were computed using unstandardized residuals from the regression model. Tables S7 and S8 present the raw normative values for each age group and the education-based correction values, respectively. Fig. 2 shows the trends in semantic errors for each age-group with z-score corrected values. The hierarchical regression analyses conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on % semantic errors showed that adding geographic area did not significantly increase explained variance (ΔR^2 = .001, p = .652). Inclusion of bilingual exposure, however, significantly increased explained variance (ΔR^2 = .015 p < .001). Nonetheless, the magnitude and significance of the age effect remained largely unchanged in this model, indicating that, also for this variable, the age-related patterns are not attributable to multilingual variation.

Regarding morphological accuracy, after removal of outliers the dataset retained 700 participants: N = 121 aged 20-29; N = 104 aged 30-39; N = 87 aged 40-49; N = 100 aged 50-59; N = 123 aged 60-69; N = 87 aged 70-79; N = 78 aged 80-94. Inspection of the Q–Q plots indicated that the distribution of % morphological errors was bounded at 0 and showed a floor effect, especially in the younger age groups, with many participants producing no errors. A linear regression was conducted to assess the influence of chronological age and years of formal education on % morphological errors. The model was significant, F (2, 699) = 57.757, p < .001, R^2 = .142. Both predictors emerged as significant: age was positively associated (b = .011, p < .001), whereas education was negatively associated with morphological errors (b = −.030, p < .001). This suggests that older participants tended to produce a higher percentage of morphological errors, while more educated participants produced fewer such errors. An ANCOVA with age group as the independent variable, education as a covariate, and % morphological errors as the dependent variable confirmed significant age-related differences, F (6, 692) = 8.562, p < .001, η^2 = .069, with a significant effect of education (p < .001) but no age*education interaction (p = .150). Tukey's post-hoc analyses showed no systematic differences in morphological error rates among participants in their 20s, 30s, 40s, and 50s; participants in their 60s differed significantly only from those in their 20s (p < .001). In contrast, individuals in their 70s produced significantly more morphological errors than participants in their 20s, 30s, 40s, 50s (all p < .001), and 60s (p = .002). Finally, participants in their 80s produced more errors than all younger groups (vs. 20s, 30, 40s, 50s, and 60s: all ps < .001) but did not differ significantly from those in their 70s. Polynomial contrasts showed a significant linear (p = .002) and a quartic (p = .015) effect, suggesting a largely monotonic increase in morphological errors with age, modulated by some curvature in the oldest age ranges. Given the effect of education, education-corrected normative values were computed using the unstandardized residuals from the regression model. Tables S9 and S10 present the raw normative values for each age group and the education-based correction values, respectively. Fig. 2 shows the trends in morphological errors for each age-group with z-score corrected values. Hierarchical regression analyses were conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on % morphological errors. Neither the inclusion of geographic area (ΔR^2 = .006, p = .455) nor bilingual exposure (ΔR^2 = .002, p = .264) significantly increased explained variance. These findings indicate that, also in this case, the age-related pattern is not attributable to regional or multilingual variation.

5.3. *Grammatical construction*

For syntactic completeness, after removing outliers the dataset retained 715 participants: N = 121 aged 20-29; N = 105 aged 30-39; N = 88 aged 40-49; N = 101 aged 50-59; N = 127 aged 60-69; N = 90 aged 70-79; N = 83 aged 80-94. Inspection of the Q–Q plots indicated that the distribution of % complete sentences was approximately normal across groups. A linear regression was conducted to assess the influence of chronological age and years of formal education on this measure. The model was significant, F (2, 714) = 49.444, p < .001, R^2 = .122. Both predictors were significant: age was negatively associated with % complete sentences (b = - .166, p < .001), whereas education was positively associated with this measure (b = .462, p < .001). Thus, older participants tended to produce fewer complete sentences, while more educated participants produced more syntactically complete sentences. An ANCOVA with Age group as the independent variable, education as a covariate, and % complete sentences as the dependent variable confirmed significant age-related differences (F (6, 707) = 7.828, p < .001, η^2 = .062) with a significant effect of education (p = .004) but no age*education interaction (p = .138). Polynomial contrasts revealed significant linear (p < .001) and quadratic effects (p < .001), indicating that the decline in syntactic completeness is not strictly linear but becomes more pronounced in later life. Tukey's post-hoc tests showed no differences in % complete sentences among participants in their 20s, 30s, 40s, and 50s, and 60s. In contrast, individuals in their 70s produced significantly fewer complete sentences than those in their 20s through 50s (all ps < .001), while performing at a similar level to adults in their 60s. Participants in their 80s produced fewer complete sentences than all younger groups (vs. 20s, 30, 40s, 50s, and 60s: all ps < .001), but did not differ significantly from those in their 70s. Given the influence of education, education-corrected normative values were computed using the unstandardized residuals from the regression model. Tables S11 and S12 present the raw normative values for each age group and education-based correction values, respectively. Fig. 3 shows the trends in syntactic completeness for each age group using z-score corrected values. Hierarchical regression analyses were conducted to

examine whether geographic origin and bilingual exposure influenced the observed age effects on % complete sentences. Adding geographic area did not significantly increase explained variance ($\Delta R^2 = .006$, $p = .102$). Inclusion of bilingual exposure, however, significantly increased explained variance ($\Delta R^2 = .010$ $p < .008$), but the magnitude and significance of the age effect remained largely unchanged in this model. These findings indicate that the observed age-related patterns are not attributable to multilingual variation.

5.4. Macrolinguistic difficulties

For cohesion errors, after outlier removal the dataset retained 715 participants: N = 119 aged 20-29; N = 105 aged 30-39; N = 89 aged 40-49; N = 101 aged 50-59; N = 127 aged 60-69; N = 90 aged 70-79; N = 84 aged 80-94. Inspection of Q–Q plots indicated that the distribution of % cohesion errors was approximately normal across age groups. A linear regression assessing the influence of chronological age and years of formal education on cohesion error rates was significant, $F(2, 714) = 4.713$, $p = .009$, $R^2 = .013$, although it explained only a small proportion of variance (1.3%). Age emerged as a significant positive predictor ($b = .052$, $p = .009$), indicating a slight increase in cohesion errors with advancing age, whereas education did not predict such errors ($b = -.005$, $p = .957$). Consistent with these findings, an ANOVA with age group as the independent variable and % cohesion errors as the dependent variable yielded a small but significant main effect of age, $F(6, 708) = 2.285$, $p = .034$, $\eta^2 = .019$. However, Tukey's post hoc tests revealed no systematic pattern of pairwise differences: the only significant contrast was that adults in their 80s produced more cohesion errors than those in their 40s ($p = .032$). All other between-decade comparisons were not significant. Polynomial contrasts showed significant linear ($p = .002$) and quadratic ($p = .029$) effects, suggesting a very modest age-related increase in cohesion errors with slight curvature, but without a clear, stepwise deterioration across successive decades. Tables S13 and S14 present the raw normative values for each age group. No education-based correction values were required for this measure. Fig. 4 shows the trends in macrolinguistic errors for each age-group using z-score corrected values. The hierarchical regression analyses conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on % cohesion errors showed that neither geographic area ($\Delta R^2 = .002$, $p = .680$) nor bilingual exposure ($\Delta R^2 = .002$, $p = .235$) significantly increased explained variance. Therefore, also in this case the age-related pattern is not attributable to regional or multilingual variation.

For local coherence errors, after removal of outliers the dataset retained 703 participants: N = 121 aged 20-29; N = 105 aged 30-39; N = 89 aged 40-49; N = 101 aged 50-59; N = 127 aged 60-69; N = 88 aged 70-79; N = 72 aged 80-94. Q–Q plots indicated that the distribution of such errors was approximately normal across age groups. A linear regression assessing the influence of age and education was significant, $F(2, 702) = 137.666$, $p < .001$, $R^2 = .282$. Both predictors contributed uniquely: age was positively associated with local coherence errors ($b = .280$, $p < .001$), whereas education was negatively associated ($b = -.845$, $p < .001$). Thus, advancing age was linked to more local coherence errors, while higher education was associated with fewer errors. An ANCOVA with age group as the independent variable, education as a covariate, and % local coherence errors as the dependent variable confirmed robust age-related differences, $F(6, 695) = 26.937$, $p < .001$, $\eta^2 = .189$, along with a strong effect of education ($p < .001$). Importantly, there was also a significant age*education interaction ($p = .009$), indicating that the protective effect of education on local coherence is not uniform across the adult lifespan, but varies as a function of age. Inspection of the interaction revealed that the positive association between age and local coherence errors was substantially steeper in individuals with lower levels of education, whereas participants with higher education showed a markedly attenuated age-related increase. This pattern suggests that education moderates the effect of ageing on discourse coherence, with higher schooling conferring resilience against age-related decline. Tukey's post-hoc tests showed no significant differences among participants in their 20s, 30s, 40s, and 50s, supporting the view that local coherence is relatively preserved across early and mid-adulthood. In contrast, individuals in their 60s produced more such errors than those in their 20s ($p = .036$), 30s ($p = .003$), and 40s ($p = .022$), while performing at a similar level to adults in their 50s. Adults in their 70s showed further deterioration, producing significantly more local coherence errors than all younger age-groups (20s-60s; all ps $< .001$). Finally, participants in their 80s produced more local coherence errors than every other decade group (20s, 30, 40s, 50s, 60s, and 70s; all ps $< .001$). Polynomial contrasts showed significant linear ($p = .002$) and quadratic ($p < .001$) effects, consistent with a pattern of monotonic age-related worsening that accelerates in later life rather than a simple straight-line trend. Given the effect of education, education-corrected normative values were derived using the unstandardized residuals from the regression model. Tables S15 and S16

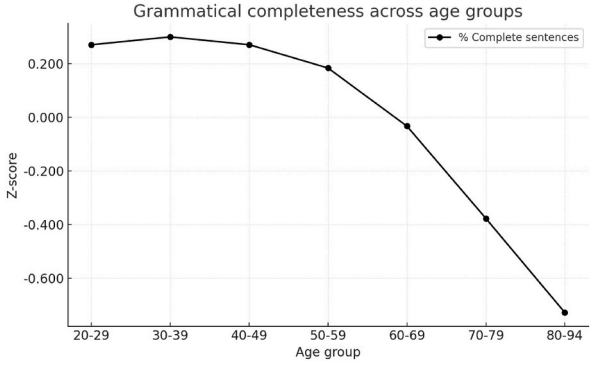

Fig. 3. – Grammatical completeness across the seven age-groups with z-score corrected values.

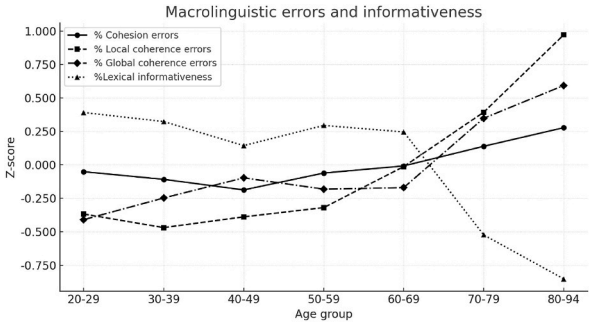

Fig. 4. – Macrolinguistic errors and lexical informativeness across the seven age-groups with z-score corrected values.

report the raw normative values for each age group and the education-based correction grids, respectively. Fig. 4 shows the trends in macrolinguistic errors for each age-group with z-score corrected values. Hierarchical regression analyses were conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on % local coherence errors. Adding geographic area did not significantly increase explained variance ($\Delta R^2 = .003$, $p = .263$). Inclusion of bilingual exposure, however, significantly increased explained variance ($\Delta R^2 = .020$, $p < .001$), but the magnitude and significance of the age effect remained largely unchanged in this model. This indicates that the age-related patterns are not attributable to multilingual variation.

Considering global coherence errors, after outlier removal the dataset retained 703 participants: N = 121 aged 20-29; N = 105 aged 30-39; N = 89 aged 40-49; N = 101 aged 50-59; N = 127 aged 60-69; N = 84 aged 70-79; N = 76 aged 80-94. Inspection of Q–Q plots indicated that the distribution of these errors was approximately normal across groups. A linear regression assessing the influence of age and education on % global coherence errors was significant, $F(2, 702) = 40.020$, $p < .001$, $R^2 = .103$. Age emerged as a positive predictor ($b = .089$, $p < .001$), indicating that older adults produced more global coherence errors, whereas education did not significantly predict performance ($b = -.090$, $p = .119$). Consistent with this, an ANOVA with age group as the independent variable and % global coherence errors as the dependent variable revealed robust age-related differences ($F(6, 696) = 17.168$, $p < .001$, $\eta^2 = .129$). Tukey's post-hoc analyses showed that participants in their 70s and 80s produced more global coherence errors than adults in their 20s, 30s, 40s, 50s, and 60s (all ps < .001, except 70s vs. 40s: $p = .006$; 80s vs. 40s: and $p < .001$). Participants in their 70s and 80s did not differ from each other. Polynomial contrasts showed significant linear ($p < .001$), quadratic ($p < .001$), cubic ($p = .018$), and quintic ($p = .019$) effects, underscoring that the effect of age on global coherence errors is clearly non-linear, with relatively stable performance up to late midlife and pronounced deterioration in later decades. Given the non-significant effect of education, Tables S17 and S18 present the raw normative values for each age group with no education-based correction. Fig. 4 shows the trends in macrolinguistic errors for each age-group with z-score corrected values. The hierarchical regression analyses conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on % global coherence errors showed that the neither geographic area ($\Delta R^2 = .007$, $p = .082$) nor bilingual exposure ($\Delta R^2 = .001$, $p = .370$) significantly increased explained variance. This suggests that, also in this case, the age-related pattern is not attributable to regional or multilingual variation.

5.5. *Informativeness*

For the percentage of lexical informativeness, after outlier removal the dataset retained 705 participants: N = 121 aged 20-29; N = 105 aged 30-39; N = 89 aged 40-49; N = 101 aged 50-59; N = 127 aged 60-69; N = 87 aged 70-79; N = 75 aged 80-94. Inspection of Q–Q plots indicated that the distribution of % lexical informativeness was approximately normal across groups. A linear regression was conducted to assess the influence of age and education on this variable. The model was significant, $F(2, 704) = 66.105$, $p < .001$, $R^2 = .158$. Age was negatively associated with lexical informativeness ($b = -.143$, $p < .001$), indicating that older adults produced fewer informative words, whereas education showed a small positive effect ($b = .155$, $p < .033$), suggesting that individuals with more schooling produced slightly more informative narratives. An ANCOVA with age group as the independent variable, education as a covariate, and % lexical informativeness as the dependent variable confirmed robust age-related differences ($F(6, 697) = 24.277$, $p < .001$, $\eta^2 = .173$). In this model, however, the effect of education was not significant ($p = .436$), and there was no age*education interaction ($p = .111$), indicating that the age effect on lexical informativeness was similar across education levels. Tukey's post-hoc tests showed no differences among adults in their 20s, 30s, 40s, 50s, and 60s, whereas participants in their 70s and 80s produced significantly fewer informative words than all younger groups (all ps < .001) and did not differ from each other. Polynomial contrasts showed significant linear ($p < .001$), quadratic ($p < .001$), cubic ($p = .010$), quartic ($p = .023$), and quintic ($p < .001$) effects, suggesting a non-linear age trajectory characterized by preserved performance across early and mid-adulthood and a sharp decline beginning in the 70s. Given the small but detectable effect of Education in the regression model, education-corrected normative values were computed using the unstandardized residuals from that model. Tables S19 and S20 present the raw normative values for each age group and education-based correction values, respectively. Fig. 4 shows the trends in lexical informativeness for each age-group with z-score corrected values. The hierarchical regression analyses conducted to examine whether geographic origin and bilingual exposure influenced the observed age effects on % lexical informativeness revealed that adding geographic area did not increase explained variance ($\Delta R^2 = .000$, $p = .825$). Inclusion of bilingual exposure significantly increased explained variance ($\Delta R^2 = .006$, $p = .036$) but

the magnitude and significance of the age effect remained largely unchanged also for this model, indicating that the age-related patterns in lexical informativeness are not attributable to multilingual variation.

6. Section 2 - exploratory age-trajectory analysis: extraction of data-driven age-bands for variables showing non-linear trends

6.1. Productivity

The baseline linear regression model including age and education as predictors showed that, after adjusting for education, age was negatively associated with total words produced ($b = -1.240$, $p = .032$). In contrast, higher education was associated with a greater number of words ($b = 20.300$, $p < .001$). The model explained 14% of the variance in total words, $F(2, 714) = 58.12$, $p < .001$; $R^2 = .14$. The Davies test did not indicate a statistically significant change in slope ($p = .267$), therefore we retained the linear model. Model diagnostics and decade-based analyses suggested a gradual decline in word production with advancing age, with education positively associated with output (see Table 3).

Considering speech rate, the baseline linear regression model with age and education as predictors indicated that, after adjusting for education, age was negatively associated with this variable ($b = -.55$, $p = .003$). In contrast, higher educational attainment was associated with slightly faster speech ($b = 1.080$, $p < .001$). Overall, the model accounted for a modest proportion of variance in speech rate, $F(2, 713) = 30.59$, $p < .001$; $R^2 = .079$. The Davies test for a change in the age slope did not support a significant breakpoint ($p = .231$). We therefore retained the linear model confirming that the association between age and speech rate is linear rather than characterized by abrupt transitions (see Table 3).

6.2. Lexical difficulties

The baseline linear regression model including age and education as predictors indicated that both variables were significantly associated with the phonological error rate. After adjusting for education, age was positively related to phonological errors ($b = .009$, $p < 001$), such that older participants produced more phonological errors. Conversely, higher education was associated with fewer phonological errors ($b = -.039$, $p < 001$). The model explained about 14% of the variance in the outcome, $F(2, 714) = 56.01$, $p < .001$; $R^2 = .136$. The Davies test provided strong evidence for a non-linear pattern, indicating a significant change in slope with a best candidate breakpoint at approximately 68 years ($p < .001$). This provided justification for fitting a segmented regression model. We

Table 3
– Results of the piecewise segmented regression models showing data-driven age-bands for the different measures of narrative production. The "Conclusion" column contains a general conclusion derived by analyses of the seven-decade age-bands and piecewise segmented regression models. The asterisk (*) shows when Davies test for a change in the age slope was significant.

Target variable	Data-driven age groups	Conclusion
Words	No significant changes in slope	Gradual linear decline in total word production with advancing age.
Speech rate	No significant changes in slope	Gradual linear decline in speech rate with advancing age.
% Phonological errors*	1) 20-42 years (M = .65; SD = .53) 2) 42-60 years (M = .82; SD = .82) 3) 60-83 years (M = 1.07; SD = .85) 4) 83+ years (M = 1.88; SD = 1.45)	Phonological errors remain low from the 20s through the late 50s, accelerate after age 60, and increase sharply in the 80s.
% Semantic errors*	1) 20-59 years (M = .79; SD = .94) 2) 59-72 years (M = 1.50; SD = 1.77) 3) 72-83 years (M = 3.83; SD = 5.00) 4) 83+ years (M = 7.47; SD = 9.26)	Semantic errors remain infrequent through late midlife, begin to increase in the 60s, and rise sharply in the 80s.
% Morphological errors	No significant changes in slope	Gradual linear decline in the production of morphological errors with advancing age.
% Complete sentences*	1) 20-24 years (M = 72.8; SD = 13.20) 2) 24-26 years (M = 73.3; SD = 13.10) 3) 26-60 years (M = 73.0; SD = 12.30) 4) 60+ years (M = 64.7; SD = 14.40)	The production of complete sentences remains stable from early through late mid-adulthood, with a decline emerging after age 60 and becoming more pronounced in individuals in their 80s.
% Cohesion errors*	The segmented model did not improve fit relative to the linear model.	Cohesion errors in narrative discourse remain stable across adulthood, showing at most a very small increase in late old age.
% Local coherence errors*	1) 20-36 years (M = 18.2; SD = 10) 2) 36-58 years (M = 19.9; SD = 9.62) 3) 58-76 years (M = 27.5; SD = 13.9) 4) 76+ years (M = 49.4; SD = 28.7)	Local coherence errors remain stable in young adults (in their 20s and mid-30s), rise slightly in individuals from their mid-30s through their late-50s and then show a steeper increase later on, particularly in those in their 70s and 80s.
% Global coherence errors*	1) 20-42 years (M = 6.87; SD = 5.17) 2) 42-61 years (M = 6.99; SD = 5.32) 3) 61-83 years (M = 10.5; SD = 8.0) 4) 83+ years (M = 17.9; SD = 10.4)	Global coherence errors remain relatively stable in young and middle-aged adults (in their 20s through their 60s), rise slightly in individuals in their 60s and 70s, and then show a steeper increase in those in their 80s.
% Lexical informativeness*	1) 20-42 years (M = 85.6; SD = 6.03) 2) 42-60 years (M = 84.7; SD = 6.4) 3) 60-83 years (M = 79.7; SD = 10.01) 4) 83+ years (M = 67.9; SD = 12.9)	Lexical informativeness remains stable throughout early and mid-adulthood (from the 20s through the 60s). A clearer decline emerges afterward, with a pronounced drop in individuals in their 80s.

then attempted to fit a segmented (piecewise) linear regression model with up to three age breakpoints, controlling for education. The algorithm converged on breakpoints at approximately 42.0, 60.0, and 83.0 years. However, the segmented model did not improve fit relative to the linear model and yielded a substantially higher AIC, indicating that the additional breakpoint parameters were not supported (AIC_segmented = 4974.34; AIC_linear = 1700.06). This likely reflected the floor-distributed nature of this variable. Considering this, we did not interpret the segmented regression coefficients further and retained the simpler linear model for inferential purposes, treating any age-band results as descriptive only. Four age bands were identified: 20–42 years ($n = 247$), 42–60 years ($n = 187$), 60–83 years ($n = 243$), and 83+ years ($n = 40$). Mean phonological error rates increased monotonically across these bands: .65 ($SD = .53$) in the 20–42 group, .82 ($SD = .82$) in the 42–60 group, 1.07 ($SD = .85$) in the 60–83 group, and 1.88 ($SD = 1.45$) in the 83+ group. Overall, these analyses are consistent with the decade-based results, suggesting that the production of phonological errors remains low from the 20s through the late 50s, accelerates after approximately age 60, and increases sharply in the 80s (see Table 3).

The baseline linear regression on semantic errors again showed a strong positive effect of age (b = .048, p < .001) and a negative effect of education (b = −.18, p < .001), explaining a substantial proportion of variance, F (2, 714) = 85.52, p < .001, $R^2 = .19$. In line with previous analyses, the Davies test for a change in the age slope provided strong evidence for a non-linear pattern, indicating a highly significant change in slope with a best candidate breakpoint at approximately 76 years ($p < 001$). This suggested that the relationship between age and semantic errors becomes steeper in advanced old age. We therefore fitted a segmented (piecewise) regression model with up to three age breakpoints, again controlling for education. The segmented model identified breakpoints at 59.17, 72.19, and 83.00 years. Relative to the linear model, the segmented model substantially improved fit (likelihood-ratio test: F (6, 708) = 23.53, p < .001; AIC_segmented = 3582.81 vs. AIC_linear = 3701.21) and accounted for a larger proportion of variance ($R^2 = .33$). In this model, the overall age term was no longer significant (reflecting the fact that the age effect is concentrated in specific age segments), whereas education remained a significant negative predictor (b = − .10, p < .001). For descriptive purposes, we used the estimated breakpoints to derive four empirically defined age bands: 20–59 years ($n = 416$), 59–72 years ($n = 155$), 72–83 years ($n = 106$), and 83+ years ($n = 40$). Mean semantic error rates increased monotonically and non-linearly across these bands: .79 ($SD = .94$) in the 20–59 group, 1.50 ($SD = 1.77$) in the 59–72 group, 3.83 ($SD = 5.00$) in the 72–83 group, and 7.47 ($SD = 9.26$) in the 83+ group. Overall, the segmented regression and age-band descriptives indicate that semantic errors remain relatively infrequent through late midlife, begin to increase in the 60s, and rise sharply in the 80s (see Table 3).

Considering the percentage of morphological errors, the baseline linear regression model including Age and Education as predictors showed that both variables were significantly associated with such difficulties. After adjusting for education, higher age was related to more morphological errors (b = .042, p < .001), whereas higher education was associated with fewer errors (b = − .13, p < .001). The linear model accounted for 15% of the variance in morphological error rate, F (2, 714) = 63.56, p < .001; $R^2 = .15$. Contrary to polynomial analyses, the Davies test for a breakpoint in the age effect did not support a reliable breakpoint (p = .343). We therefore retained the linear specification for inferential purposes (see Table 3).

6.3. Grammatical construction

Considering complete sentences, the baseline linear regression model with age and education as predictors showed that both variables were significantly associated with syntactic completeness. After adjusting for education, higher age was related to a lower percentage of complete sentences ($b = − .16$, $p < 001$), while higher education was associated with a higher percentage of complete sentences ($b = .49$, $p < 001$). The linear model accounted for 12% of the variance in syntactic completeness, F (2, 714) = 48.61, $p < .001$; $R^2 = .12$. The Davies test indicated evidence for a change in the age slope, with the most likely breakpoint around 60 years (p < .001), suggesting that the association between age and syntactic completeness is not strictly linear across the adult lifespan and accelerates in later life. We therefore fitted a segmented (piecewise) regression model with three candidate breakpoints, which were estimated at 24, 26, and 60 years. The segmented model modestly increased explained variance ($R^2 = .14$) and significantly improved fit compared to the linear model, F (6, 708) = 3.26, p < .004), with a slightly lower AIC than the linear model (AIC_segmented = 5716.53; AIC_linear = 5724.09), supporting the presence of nonlinearity in the age effect. In this model, education remained positively associated with syntactic completeness (b = .34, p = .009), whereas the overall age term was no longer significant, and the age-related pattern was captured through changes in slope around the identified breakpoints. For descriptive purposes, we derived four empirical age bands based on the estimated breakpoints: 20–24 years (n = 55), 24–26 years (n = 44), 26–60 years (n = 335), and 60+ years (n = 283). Mean syntactic completeness was very similar in the first three age-bands (20–24: $M = 72.8\%$, $SD = 13.2$; 24–26: $M = 73.3\%$, $SD = 13.1$; 26–60: $M = 73.0\%$, $SD = 12.3$), with a noticeable reduction in the in the 60+ group ($M = 64.7\%$, $SD = 14.4$). Therefore, in line with previous analyses using decade age-bands, the first three segments were collapsed into a single age group for interpretability. Taken together, these results suggest that syntactic completeness is relatively stable from early through late mid-adulthood, with a decline emerging after age 60 and becoming more pronounced in individuals in their 80s (see Table 3).

6.4. Macrolinguistic difficulties

Considering the percentage of cohesion errors, the baseline linear regression model with age and education as predictors showed a small but statistically significant association between age and cohesion errors. After adjusting for education, higher age was associated with a slightly higher percentage of cohesion errors (b = .04, p = .030), whereas education did not significantly predict cohesion errors (b = −.02, p = .83). Overall, the linear model explained only 1% of the variance in cohesion errors, F (2, 714) = 3.573, p = .029; $R^2 = .01$, indicating that, although the age effect is reliable, its magnitude is very small. Although the Davies test suggested potential slope instability (p = .026), the segmented model did not improve fit relative to the linear model. Indeed, the segmented model showed only

a modest increase in explained variance ($R^2 = .02$), and comparison with the linear model was not statistically compelling, F (6, 708) = 1.784, p = .10, with virtually identical AIC values (AIC_linear = 5238.43; AIC_segmented = 5239.67). Thus, although there is some indication of nonlinearity, the segmented solution does not clearly outperform the simpler linear specification. We therefore retained the linear model and did not derive age bands for % Cohesion errors. Both the decade-based normative analysis and the continuous/segmented models converge on the conclusion that cohesion errors in narrative discourse are remarkably stable across adulthood, showing at most a very small increase in late old age (see Table 3).

Regarding local coherence errors, the baseline linear model again revealed strong, opposing effects of age (more errors: b = .33, p < .001) and education (fewer errors: b = −1.24, p < .001), together explaining 32% of the variance, F (2, 714) = 167.40, p < .001; R^2 = .32. The Davies test provided strong evidence for a non-linear age effect, with the most likely breakpoint around 68 years (p < .001), prompting a segmented regression. The segmented model identified change points at 36, 58, and 76 years. In this piecewise model, the overall age coefficient was not significant (b = − .11, p = .593), reflecting that age-related differences were captured by the slope changes surrounding the breakpoints, whereas education retained a strong protective effect (b = − .83, p < .001). The segmented model explained more variance ($R^2 = .40$), and a formal comparison confirmed that it fit the data significantly better than the simple linear model, F (6, 708) = 15.831, p < .001, with a notably lower AIC (AIC_segmented = 5896.14 vs. AIC_linear = 5974.40). Four empirical age bands were derived: 20–36 years ($n = 206$), 36–58 years ($n = 191$), 58–76 years ($n = 206$), and 76+ years ($n = 114$). Mean local coherence error rates increased only slightly between the first two age-bands (20–36: $M = 18.2\%$, $SD = 10$; 36–58: $M = 19.9\%$, $SD = 9.62$), with a more pronounced increase in the next two age-bands (58–76: $M = 27.5\%$, $SD = 13.9$; and 76+: $M = 49.4\%$, $SD = 28.7$). These findings refine observations from the decade-based analyses, showing that local coherence errors remain relatively stable in early adulthood (20s – mid 30s), begin to rise through midlife (mid 30s – late 50s), and then increase sharply in later life, particularly among adults in their 70s and 80s (see Table 3).

As for the percentage of global coherence errors, the linear regression showed a significant positive effect of age on global coherence errors (b = .12, p < .001) and no reliable effect of education (b = − .09, p = .18), accounting for 14% of the variance, F (2, 714) = 57.22, p < .001; $R^2 = .14$. A Davies test provided strong evidence for a non-linear age effect, identifying a likely change in slope around 60 years (p < .001). A segmented regression with three candidate breakpoints converged on change points at 42, 61, and 83 years. In this piecewise model, the overall age slope remained significant (b = .16, p = .005), while education again showed no meaningful association with global coherence errors (b = .03, p = .69). The segmented model explained a larger proportion of variance ($R^2 = .22$) and significantly improved model fit over the linear specification, F (6, 708) = 11.801, p < .001, also yielding a lower AIC (AIC_segmented = 4707.63 vs. AIC_linear = 4763.97). Four empirical age bands were derived: 20–42 years ($n = 260$), 42–61 years ($n = 174$), 61–83 years ($n = 243$), and 83+ years ($n = 40$). Mean global coherence error rates were relatively low and stable up to early old age: 20–42: $M = 6.87\%$, $SD = 5.17$; 42–61: $M = 6.99\%$, $SD = 5.32$. A clear increase emerged in the 61–83 group ($M = 10.5\%$, $SD = 8.00$), followed by a further steep rise in the oldest adults (83+: $M = 17.9\%$, $SD = 10.4$). These findings refine the decade-based age observations, showing that global coherence errors remain relatively stable from the 20s through the 60s, increase modestly in the 60s and 70s, and then show a marked escalation among adults in their 80s (see Table 3).

6.5. *Informativeness*

Considering lexical informativeness, the baseline linear regression with age and education as predictors showed that higher age was clearly associated with lower lexical informativeness ($b = -.18$, $p < .001$), while education exerted a small positive effect ($b = .20$, $p = .013$). This model explained 19% of the variance in lexical informativeness, F (2, 714) = 84.98, p < .001; $R^2 = .19$, indicating a robust linear decline in informativeness with advancing age, partially offset by higher schooling. However, a Davies test strongly indicated that the age effect is non-linear, identifying a likely change in slope around 68 years (p < .001). A segmented regression with three candidate breakpoints identified change points at 42, 60, and 83 years. In this piecewise model, the overall age slope remained negative and significant ($b = -.18$, $p < .023$), but the pattern of the slope-change parameters (U1, U2, U3) showed that the decline is shallow across early and mid-adulthood and steepens markedly in later life. Notably, once this non-linear structure is considered, the effect of education was no longer detectable ($b = .00$, $p = .98$). The segmented model accounted for substantially more variance ($R^2 = .32$) and significantly improved fit relative to the linear model, F (6, 708) = 21.586, p < .001; $R^2 = .317$, with a lower AIC (AIC_segmented = 4974.34; AIC_linear = 5082.79). In this model, age remained a strong predictor of lexical informativeness, whereas the contribution of education was no longer reliable, suggesting that once the non-linear shape of the age trajectory is considered, chronological age emerges as the primary determinant of performance. Four empirical age bands were derived: 20–42 years ($n = 247$), 42–60 years ($n = 187$), 60–83 years ($n = 243$), and 83+ years ($n = 40$). Lexical informativeness was very high and relatively stable through early and mid-adulthood: 20–42: $M = 85.6\%$, $SD = 6.0$; 42–60: $M = 84.7\%$, $SD = 6.4$. A clearer decline emerged in the 60–83 group ($M = 79.7\%$, $SD = 10.1$), with a pronounced drop in the oldest adults (83+: $M = 67.9\%$, $SD = 12.9$; Mean Age = 86.2 years) (see Table 3).

7. Discussion

The present study aimed to provide a comprehensive characterization of narrative discourse production in healthy Italian speakers across the adult lifespan. Specifically, it addressed key methodological and empirical gaps that have long limited the clinical and research utility of discourse analysis (e.g., Bryant et al., 2016). By analyzing the performance of more than 700 healthy adults spanning seven decades of life, and by applying both traditional normative approaches and data-driven modelling of age effects, our findings offer the most extensive adult lifespan reference framework currently available for narrative discourse production. Indeed, previous

studies establishing normative data have usually included relatively small samples: Kong et al. (2025), who presented norms for Cantonese language, enrolled 149 healthy adults; Boucher et al. (2022), who reported French Canadian norms, recruited 62 healthy adults; Berube et al. (2019), who updated the English *Cookie Theft* normative data, collected 50 healthy speakers; and Richardson and Hudspeth (2016), who also worked with English speakers, included 92 healthy adults. Furthermore, consistent with previous work suggesting that aging exerts heterogeneous effects on micro- and macrolinguistic processes, our results indicate that discourse production does not follow a uniform, linear decline across adulthood. Rather, age-related differences in productivity, lexical difficulties, grammatical accuracy, macrolinguistic impairments, and lexical informativeness emerged selectively across discourse components and displayed distinct trajectories highlighting the multidimensional nature of discourse (Hilviu et al., 2025; Marini et al., 2025).

A major contribution of the present study lies in the provision of age- and education-adjusted normative values for the MLA derived from rigorous regression-based procedures following the removal of outliers. These norms address an important need in clinical practice: they allow clinicians to evaluate discourse performance not only against broad age expectations but also in relation to an individual's educational background, a factor known to influence linguistic and cognitive performance (Malcorra et al., 2022; Marini et al., 2025). This is a particularly delicate issue as the effect potentially exerted by education level has often not been considered in previous studies and may lead to inaccurate or imprecise clinical-diagnostic assessments. In our investigation, the normative tables and corrected scores (see Supplementary materials) provide a precise foundation for identifying atypical discourse patterns and for disentangling healthy age-related variability from signs of underlying pathology. Beyond these traditional normative outcomes, the application of segmented regression analyses represents a second key advance. By treating age as a continuous variable and allowing the data to determine where changes in slope occur, this approach revealed critical inflection points in the ageing trajectory of some discourse measures. These points were not predictable *a priori* and do not align strictly with chronological decades. The empirically derived age bands offer a more nuanced understanding of when discourse abilities begin to show subtle shifts and when more pronounced declines emerge. Importantly, they complement the decade-based normative results rather than replacing them, offering clinicians and researchers two complementary interpretive frameworks: one aligned with conventional normative groupings, and one reflecting the underlying structure of adult lifespan changes. Taken together, these approaches converge on a coherent picture: narrative discourse is largely preserved through early and mid-adulthood but shows domain-specific and often non-linear changes from later midlife onwards, with the most pronounced alterations emerging in the 70s and 80s.

For productivity measures, the data support a gradual (approximately linear) decline with age rather than discrete transitions. For total word output, both the normative and trajectory analyses point to a gradual age-related decline. The baseline regression confirmed that age is negatively associated with word production, whereas education has a strong positive effect. The Davies test did not support a statistically reliable change in slope, and the segmented model did not improve fit over the simple linear model. This result is partially coherent with the findings by Marini et al. (2025) and further refines the effects of age and education on word production. A similar picture emerged for speech rate. Age was again negatively associated with fluency, and education exerted a modest positive effect. The Davies test did not provide strong evidence for a breakpoint, and the segmented model did not outperform the linear one. Again, this is in line with previous findings by Marini et al. (2025) but at odds with those by Hilviu et al. (2025) and Leeper and Culatta (1995) who observed a non-linear pattern in speech rate. However, the relatively small number of participants in those studies (N = 60 in Hilviu et al., 2025; N = 78 in Leeper & Culatta, 1995) and the inclusion of just few age-bands (young adults in their 20s and 30s, mature adults aged 65 through 75, and older participants aged 76 to 86 in Hilviu et al., 2025 and 5 age-groups without a group of middle-aged adults in Leeper & Culatta, 1995) may account for such discrepancy confirming the need for adequate sample sizes and the inclusion of comprehensive age-ranges in such studies. Therefore, for productivity measures, the data-driven modelling confirms that age-related changes are best conceptualized as broadly linear trends with late-life intensification, rather than sharp transitions at specific ages. In clinical terms, substantial reductions in both total word output and speech rate are not typical before late midlife, and more pronounced changes are especially characteristic of adults in their late 70s and 80s.

Considering lexical difficulties, phonological, semantic, and morphological errors increase with age and are mitigated by higher education. However, they differ in how strongly and how non-linearly they change. For phonological errors, the linear model already captured a sizeable portion of variance, with age positively and education negatively associated with error rates. The Davies test suggested non-linearity, but the segmented model performed worse than the linear one. The derived age bands are therefore best viewed as descriptive: phonological errors remain low from the 20s through the late 50s, increase modestly in the 60s and 70s, and rise more sharply in the 80s. This pattern is consistent with the decade-based findings, but there is no strong statistical justification for replacing the linear specification. Overall, this is coherent with the results from Marini et al. (2025) who observed a linear increase in the production of such errors confirming that a large age-sample enhances the sensitivity to subtle, progressive changes not captured in previous investigations with smaller age groups (e.g., Hilviu et al., 2025). Consistently with previous investigations (e.g., Au et al., 1995; Marini et al., 2005), semantic errors followed a nonlinear pattern: they are rare and relatively stable through late midlife, increase in the 60s, and then rise steeply from the 70s onward, especially in the oldest adults. As these errors likely reflect a failure to efficiently inhibit semantic competitors during the process of lexical selection while keeping active the concepts that the speaker needs to convey, this pattern indicates that the efficiency of the process of lexical selection remains stable between young, middle-aged, and mature adults, and that the 60–70s represent a transition from relative stability to accelerated decline (see also Connor et al., 2004). Nonetheless, in our study even in the group of oldest individuals the production of such errors remained limited, and the high interindividual variability suggests that these difficulties may occur only in some individuals. Likely, other factors not explicitly addressed here (e.g., cognitive resilience; Oosterhuis et al., 2023) may have played a role. For morphological errors, age and education again showed clear, opposing linear effects. The Davies test did not support a reliable breakpoint, and segmented regression did not improve fit. Overall, these results show that while all lexical error types increase with age, semantic errors exhibit a distinctly non-linear, late-life escalation, whereas phonological and morphological errors follow more gradual trajectories. This pattern

reinforces the importance of examining lexical components separately and suggests that semantic integrity in narrative discourse may offer a particularly sensitive marker of ageing-related change.

Regarding grammatical accuracy, the analyses of syntactic completeness revealed a trajectory that is largely consistent across methodological approaches. The baseline regression showed that older age is associated with fewer complete sentences, and that higher education is associated with greater syntactic completeness (see also D'Ortenzio et al., 2025 for similar results). The derived age bands showed very similar levels of syntactic completeness across the first three segments and a clear drop in the 60+ group. For interpretability, these were collapsed into a "<60" and a "60+" band. This solution is highly consistent with the decade-based ANCOVA results, which also indicated preserved syntactic completeness from the 20s through the 50s and a subsequent decline from the 60s onward, becoming more marked in the 80s. These converging findings suggest that sentence-level grammatical encoding in narrative discourse is remarkably stable through early and mid-adulthood, with age-related decline emerging primarily after 60. The segmented analysis refines this picture by pinpointing 60 years as a meaningful transition point, rather than an arbitrary chronological threshold.

The assessment of macrolinguistic organization revealed a clear dissociation between cohesion and coherence, and the segmented analyses help clarify the timing and shape of these changes. For cohesion errors, age was associated with a small increase, and education was not a significant predictor. The linear model explained only about 1% of the variance, and although the Davies test suggested non-linearity, the segmented solution did not meaningfully improve fit. The data-driven bands showed virtually identical cohesion error rates across all segments. Both the decade-based and segmented analyses therefore converge on the same conclusion: cohesion is strikingly stable across adulthood, with at most a very small increase in late old age. Formal devices for linking propositions (e.g., pronouns, connectives) appear to be largely resilient to ageing. By contrast, local coherence errors showed a robust, clearly non-linear trajectory. The linear model already captured a large proportion of variance, with age increasing and education decreasing error rates, but the Davies test and segmented regression demonstrated that a piecewise specification provides a much better fit. The empirical age bands illustrate that local coherence errors are minimal and stable from the 20s to mid 30s, increase modestly from the mid 30s through the late 50s, and show substantial, accelerating increases from the late 50s into the 70s and 80s. This pattern refines the decade-based results by indicating that the deterioration in local coherence is not confined to old age but begins subtly in midlife, then becomes much more pronounced in later decades. Coherently with previous studies this study highlights an increasing difficulty in linking adjacent utterances via conceptual means (Duong & Ska, 2001). The persistent protective effect of education across segments further supports the idea that cognitive reserve plays a role in maintaining coherence. The significant age*education interaction observed for local coherence further refines this interpretation. Specifically, the age-related increase in local coherence errors was substantially steeper in individuals with lower levels of education, whereas participants with higher education showed a markedly attenuated age-related slope. In other words, although local coherence declines with advancing age, this deterioration is considerably moderated by educational attainment. This pattern is consistent with cognitive reserve accounts, according to which formal education enhances the efficiency or flexibility of higher-order control mechanisms that support discourse production (e.g., Marini et al., 2026). Notably, this moderating effect emerged most clearly for local coherence (a measure that requires maintaining conceptual links between adjacent propositions and suppressing irrelevant semantic competitors). This aligns with the findings of Hoffman et al. (2018) and shows that educational attainment may buffer these processes across the adult lifespan, attenuating the impact of aging on integrative discourse processes. Importantly, this protective effect does not eliminate age-related change, but rather reduces its magnitude, suggesting that education enhances resilience rather than preventing decline entirely.

A similar but temporally shifted pattern emerged for global coherence errors rates. The linear model indicated a positive effect of age and no reliable education effect, and the segmented model significantly improved fit and explained variance. The data-driven bands showed stable, low global coherence error rates up to the early 60s, followed by a clear increase from approximately 60 to the early 80s and a further steep rise in the oldest adults. Thus, while local coherence begins to deteriorate earlier and more gradually, global coherence remains relatively preserved until early old age, then declines sharply, particularly after 80. Taken together, these findings show that coherence (especially in later life) is substantially more vulnerable than cohesion. The ability to maintain a thematically consistent, contextually appropriate narrative is sensitive to age, with distinct phases of decline, whereas the formal mechanisms that bind sentences together remain largely intact. Clinically, this underscores the need to complement surface-level measures of discourse structure with more conceptually oriented indices of coherence.

Finally, the measure of lexical informativeness integrates aspects of productivity and relevance by quantifying the proportion of words that contribute meaningfully to the narrative. The baseline regression showed a strong negative effect of age and a small positive effect of education. However, the Davies test and segmented regression clearly favored a non-linear specification, suggesting that the decline in informativeness is shallow in early and mid-adulthood and steepens in later life. The segmented model substantially increased explained variance, and once the non-linear age structure was considered, the effect of education was no longer significant. The derived age bands showed high and stable informativeness through the first two segments, a noticeable decline in the 60–83 group, and a pronounced drop in adults aged 83 and above. These results refine the decade-based findings by identifying 60 and 83 years as meaningful inflection points: lexical informativeness is preserved through early and mid-adulthood, begins to decline from around 60, and deteriorates sharply in the oldest old. They also suggest that once the shape of the age trajectory is modelled appropriately, chronological age becomes the dominant driver of lexical informativeness, and the apparent protective effect of education largely reflects differences in where individuals fall along this non-linear curve.

From a methodological point of view, the combination of decade-based norms and segmented regression offers complementary perspectives. For variables such as total word output, speech rate, phonological and morphological errors, and cohesion errors, simple linear models with education as a covariate are sufficient to capture the main age-related trends; the segmented models do not meaningfully improve fit and are best used descriptively. In these cases, decade-based normative values and education-corrected

residuals provide a clear and interpretable reference frame for clinical use. For other variables (semantic errors, syntactic completeness, local and global coherence, and lexical informativeness) the segmented models clearly outperform linear ones, revealing distinct age phases that align with functional changes rather than arbitrary chronological cut-offs. These data-driven bands highlight: a transition around 60 years for syntactic completeness, global coherence, and lexical informativeness; earlier, more gradual changes in local coherence beginning from the mid 30s; and marked accelerations in error rates and loss of informativeness in the 70s and 80s, especially beyond 80–83 years. These findings are coherent with recent neuroimaging evidence showing the existence of four major topological turning points across the lifespan around 9, 32, 66, and 83 years old (Mousley et al., 2025). These insights are particularly relevant for studies of healthy and pathological aging, as they suggest that the sensitivity of discourse measures can be enhanced by adopting empirically derived age bands that reflect genuine shifts in performance. An additional set of hierarchical regression analyses examined whether geographic and bilingual exposure influenced age-related patterns. Across discourse measures, regional provenance did not account for meaningful additional variance beyond age and education. Bilingual exposure explained small increments of variance for some indices (e.g., percentages of phonological errors, semantic errors, lexical informativeness), yet, critically, the magnitude and statistical significance of the age effects remained stable across models. These findings indicate that the trajectories reported here are robust and cannot be attributed to regional variation or multilingual background. Rather, age-related differences in discourse production appear consistent across diverse geographic and linguistic experiences within the Italian-speaking population.

These findings have relevant clinical and theoretical implications. From a clinical perspective, education emerges as a consistent protective factor in the linear models, supporting the notion that schooling contributes to a form of cognitive reserve for discourse production (Sharp & Gatz, 2011; Stern et al., 1994). In several segmented models, however, the role of education diminishes once the non-linear age structure is considered, indicating that age phase (i.e., where individuals sit along the trajectory) can be more informative than age alone. The present norms and trajectories show that reductions in productivity and phonological, morphological and cohesion abilities show a gradual, linear, trend across adulthood. Syntactic production abilities remain quite stable until late midlife, with a decline emerging after age 60 and becoming more pronounced in individuals in their 80s. Our results also show that increases in semantic errors, coherence errors, and reductions in informativeness in the 70s and 80s may be compatible with healthy aging within certain limits, but marked deviations from the normative bands, or early emergence of such patterns, should prompt further investigation. Furthermore, coherence and informativeness (especially when examined using data-driven age bands) may provide sensitive markers for distinguishing typical aging from early neurodegenerative changes, even when more traditional microlinguistic measures appear relatively preserved. At a theoretical level, the differentiated trajectories across components support models that view discourse production as the outcome of interacting micro- and macrolinguistic processes with distinct susceptibilities to aging. Lower-level formal mechanisms (e.g., basic grammatical structure) are relatively robust, whereas higher-level integrative mechanisms (coherence, informativeness) show more pronounced late-life decline.

From a theoretical perspective, the differentiated trajectories observed across discourse components provide insight into the organization of language production in aging. Measures reflecting lower-level structural encoding, such as total word output, speech rate, and morphological accuracy, followed predominantly linear trajectories with relatively modest age effects. In contrast, measures requiring higher-level integrative processing (i.e., semantic error production, local and global coherence errors, and lexical informativeness) displayed clear non-linear patterns with late-life acceleration. This dissociation supports models proposing that language production depends on interacting but partially separable mechanisms, with higher-order semantic integration and executive regulation being more vulnerable to aging than lexical and morphosyntactic encoding processes. These findings resonate with a previous investigation by Kavé and Goral (2017), who showed that age-related word retrieval difficulties may be attenuated in connected speech where semantic and contextual support can scaffold production. In our data, productivity measures remained relatively stable through midlife, suggesting that contextualized discourse may compensate for isolated lexical retrieval challenges. However, when tasks require maintaining conceptual coherence or suppressing irrelevant semantic competitors, age effects become more pronounced. The non-linear increases in the production of local and global coherence errors align closely with Hoffman et al. (2018), who demonstrated that poor coherence in older adults is better explained by impaired semantic control and executive processes than by degradation of semantic knowledge per se. Importantly, this interpretation aligns with broader evidence indicating that discourse planning and conceptual preparation rely on distributed cognitive networks (Arbuckle et al., 2000; Pistono et al., 2017) and is consistent with recent multilevel investigations showing that discourse production is not purely linguistic in nature but is deeply intertwined with executive and socio-cognitive processes (Gallo et al., 2025; Hilviu et al., 2025; Marini et al., 2025). For example, in a prior study adopting the same multilevel framework (Marini et al., 2025), an "executive-working memory efficiency" component showed the strongest negative association with semantic-coherence breakdown, suggesting that the ability to maintain the global conceptual organization of discourse depends on sustained attention, inhibitory control, flexible updating of a mental model, and efficient suppression of irrelevant semantic competitors during the process of lexical selection. This interpretation is also coherent with the Inhibition deficit hypothesis (Hasher & Zacks, 1988), which suggests that aging compromises the ability to suppress irrelevant information and maintain focus on goal-relevant representations. Applied to discourse production, this reduced inhibitory efficiency may lead to increased inclusion of irrelevant concepts and events, thereby undermining both local and global coherence and reducing lexical informativeness.

This study has some limitations. First, its cross-sectional design prevents us from distinguishing true ageing effects from potential cohort influences. Second, narrative performance was assessed using a single genre picture-based storytelling which does not capture the full range of discourse types used in everyday communication. Future studies should replicate such a comprehensive assessment using also other types of discourse (e.g., procedural or expository). A third limitation concerns the fact that the segmented regression analyses were exploratory and, for some variables, did not improve upon simpler linear models, meaning that some data-driven age bands should be interpreted cautiously. Finally, although age and education were modelled, other factors known to influence discourse

such as executive functions and socioeconomic status were not directly assessed in this study. In addition, bilingualism was only indirectly captured through self-reported language background, without detailed information regarding proficiency, frequency of use, age of acquisition, or language switching habits. Future research should therefore adopt more fine-grained measures of multilingual experience to clarify its potential role in shaping discourse trajectories across the lifespan (Bialystok, 2021).

In conclusion, by combining traditional decade-based normative values with segmented, data-driven age bands in a large, well-stratified sample, this study provides a nuanced and clinically useable view of how narrative discourse changes from early adulthood to very old age. The decade-based norms offer a unique practical framework for assessment and standardization of narrative production abilities in Italian speakers, while the data-driven age bands reveal critical phases in the trajectory of discourse in adulthood that can guide future research on healthy ageing, cognitive reserve, and the early detection of pathological language change. The norms of the MLA and the derived age bands are intended to support future work in at least three domains: (i) clinical assessment, by helping clinicians differentiate typical age-related changes from pathological discourse profiles; (ii) research on healthy ageing, by providing well-characterized age ranges for group comparisons and longitudinal studies; and (iii) early detection of neurodegenerative conditions, where sensitive discourse markers, interpreted against robust normative and age-banded benchmarks, may reveal subtle changes before standard tests do. Finally, beyond clinical and aging research, the availability of standardized, multilevel discourse measures also opens promising avenues for future studies aimed at monitoring cognitive efficiency in extreme and isolated environments, such as long-duration deep-space missions, where subtle changes in discourse production may provide sensitive markers of cognitive functioning over time.

CRediT authorship contribution statement

A. Marini: Conceptualization, Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis. **F. Petriglia:** Writing – review & editing, Software, Data curation. **G. Gasparotto:** Writing – review & editing, Methodology, Formal analysis, Data curation. **S. D'Ortenzio:** Writing – review & editing, Methodology, Formal analysis, Data curation. **S. Andreetta:** Writing – review & editing, Formal analysis. **M. Gobbo:** Writing – review & editing, Formal analysis.

Funding

This research was supported by PRIN 2022 PN RR, Prot. n. P2022M9JCM, project title: "Standardization of the Multilevel prOcedure for discOurse analysis and Training program for narrative production in Healthy adults - SMOOTH". Avviso pubblico n. 1409 del 14/09/2022 – PRIN 2022 PNRR M4C2 Inv. 1.1. Ministero dell'Università e della Ricerca (Financed by EU, NextGenerationEU) – CUP G53D23007250001.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This manuscript was proofread and edited for language clarity using ChatGPT 5.0. All scientific content, interpretations, and conclusions remain the responsibility of the authors. The authors wish to thank Cinzia Baldanzi and Chiara Vitali (IRCCS Don Gnocchi), Giulia Fusari (University of Pavia), Francesco Ferretti (University of Roma Tre), and Paola Marangolo (University of Naples). The authors wish to thank also Gianmarco Veronesi for drawing the new version of the *Flowerpot* story used for this standardization and Laura Bellini, Irene Castellani and Cristina Reverberi for their help.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jneuroling.2026.101343.

References

Aita, S. L., Beach, J. D., Taylor, S. E., Borgogna, N. C., Harrell, M. N., & Hill, B. D. (2019). Executive, language, or both? An examination of the construct validity of verbal fluency measures. *Applied Neuropsychology Adult., 26*(5), 441–451. https://doi.org/10.1080/23279095.2018.1439830

Andreetta, S., Cantagallo, A., & Marini, A. (2012). Narrative discourse in anomic aphasia. *Neuropsychologia, 50*(8), 1787–1793. https://doi.org/10.1016/j.neuropsychologia.2012.04.003

Andreetta, S., Marini, A., Menichelli, A., Furlanis, G., Vincis, E., Caruso, P., Naccarato, M., & Manganotti, P. (2025). Language assessment in persons with aphasia early after thrombolysis: The utility of multilevel procedures of discourse analysis. *Aphasiology, 39*(2), 182–213. https://doi.org/10.1080/02687038.2024.2340797

Arbuckle, T. Y., Nohara-LeClair, M., & Pushkar, D. (2000). Effect of off-target verbosity on communication efficiency in a referential communication task. *Psychology and Aging, 15*(1), 65–77. https://doi.org/10.1037//0882-7974.15.1.65

Au, R., Joung, P., Nicholas, M., Obler, L. K., Kass, R., & Albert, M. L. (1995). Naming ability across the adult life span. *Aging, Neuropsychology, and Cognition, 2*(4), 300–311. https://doi.org/10.1080/13825589508256605

Berube, S., Nonnemacher, J., Demsky, C., Glenn, S., Saxena, A., Wright, A., Tippett, D. C., & Hillis, A. E. (2019). Stealing cookies in the twenty-first century: Measures of spoken narrative in healthy versus speakers with Aphasia. *American Journal of Speech-Language Pathology, 28*(1S), 321–329. https://doi.org/10.1044/2018_AJSLP-17-0131

Bialystok, E. (2021). Bilingualism: Pathway to cognitive reserve. *Trends in Cognitive Sciences, 25*(5), 355–364. https://doi.org/10.1016/j.tics.2021.02.003

Boucher, J., Brisebois, A., Slegers, A., Courson, M., Désilets-Barnabé, M., Chouinard, A.-M., Gbeglo, V., Marcotte, K., & Brambati, S. M. (2022). Picture description of the Western Aphasia battery picnic scene: Reference data for the French Canadian population. *American Journal of Speech-Language Pathology, 31*(1), 257–270. https://doi.org/10.1044/2021_AJSLP-20-00388

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample-size and test-retest stability of connected speech measures for adults with aphasia. *Journal of Speech & Hearing Research, 37*(2), 399–407. https://doi.org/10.1044/jshr.3702.399

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

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*(2), 49–57. https://doi.org/10.1016/j.tics.2006.11.004

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*, 249–254. https://doi.org/10.48550/arXiv.cmp-lg/9602004

Connor, L. T., Spiro, A., Obler, L. K., & Albert, M. L. (2004). Change in object naming ability during adulthood. *The Journals of Gerontology, Series B: Psychological Science, 59*(5), 203–209. https://doi.org/10.1093/geronb/59.5.p203

Conti, S., Bonazzi, S., Laiacona, M., Masina, M., & Coralli, M. V. (2015). Montreal cognitive assessment (MoCA) - Italian version: Regression based norms and equivalent scores. *Neurological Sciences, 36*, 209–214. https://doi.org/10.1007/s10072-014-1921-3

Cupit, J., Graham, N. L., Leonard, C., Tang-Wai, D., Black, S. E., & Rochon, E. (2016). Wh-questions and passive sentences in non-fluent variant PPA and semantic variant PPA: Longitudinal findings of an anagram production task. *Cognitive Neuropsychology, 33*, 329–342. https://doi.org/10.1080/02643294.2016.1179179

De Renzi, E., & Vignolo, L. A. (1962). The token test: A sensitive test to detect receptive disturbances in aphasics. *Brain: Journal of Neurology, 85*, 665–678. https://doi.org/10.1093/brain/85.4.665

Dipper, L., Marshall, J., Boyle, M., Hersh, D., Botting, N., & Cruice, M. (2021). Creating a theoretical framework to underpin discourse assessment and intervention in aphasia. *Brain Sciences, 11*, 183. https://doi.org/10.3390/brainsci11020183

D'Ortenzio, S., Petriglia, F., Gasparotto, G., Andreetta, S., Gobbo, M., & Marini, A. (2025). Aging, cognitive efficiency, and lifelong learning: Impacts on simple and complex sentence production during storytelling. *Brain Sciences, 15*, 1120. https://doi.org/10.3390/brainsci15101120

Duong, A., & Ska, B. (2001). Production of narratives: Picture sequence facilitates organizational but not conceptual processing in less educated subjects. *Brain and Cognition, 46*(1–2), 121–124. https://doi.org/10.1016/s0278-2626(01)80047-6

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gallo, E., Bosco, F. M., & Marini, A. (2025). An assessment of the relation between narrative productivity, subjectivity expression, emotion description, and ToM abilities in healthy aging. *Scientific Reports, 15*, Article 39837. https://doi.org/10.1038/s41598-025-23442-9

Glosser, G., & Deser, T. (1990). Patterns of discourse production among neurological patients with fluent language disorders. *Brain and Language, 40*(1), 67–88. https://doi.org/10.1016/0093-934x(91)90117-j

Goodglass, H., & Kaplan, E. (1972). *The Boston diagnostic Aphasia examination*. Philadelphia: Lea & Febiger.

Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Boston diagnostic Aphasia examination* (3rd ed.). Lippencott: Williams & Wilkins.

Gordon, J. K., Young, M., & Garcia, C. (2018). Why do older adults have difficulty with semantic fluency? *Aging, Neuropsychology, and Cognition, 25*(6), 803–828. https://doi.org/10.1080/13825585.2017.1374328

Harry, A., & Crowe, S. F. (2014). Is the Boston Naming Test still fit for purpose? *The Clinical Neuropsychologist, 28*(3), 486–504. https://doi.org/10.1080/13854046.2014.892155

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of Learning and Motivation, 22*, 193–225. https://doi.org/10.1016/S0079-7421(08)60041-9

Hilviu, D., Parola, A., Bosco, F. M., Marini, A., & Gabbatore, I. (2025). Grandpa, tell me a story! narrative ability in healthy aging and its relationship with cognitive functions and Theory of Mind. *Language, Cognition and Neuroscience, 40*(1), 103–121. https://doi.org/10.1080/23273798.2024.2401027

Hoffman, P., Loginova, E., & Russell, A. (2018). Poor coherence in older people's speech is explained by impaired semantic and executive processes. *eLife, 7*, Article e38907. https://doi.org/10.7554/eLife.38907

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural Language processing in python. https://doi.org/10.5281/zenodo.1212303.

Huber, W., & Gleber, J. (1982). Linguistic and non-linguistic processing of narratives in aphasia. *Brain and Language, 16*, 1–18. https://doi.org/10.1016/0093-934x(82)90069-4

Indefrey, P. (2012). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology, 2*, 255. https://doi.org/10.3389/fpsyg.2011.00255

Indefrey, P., & Levelt, W. J. M. (2000). The neural correlates of language production. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 845–865). MIT Press.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge University Press.

Juncos-Rabadán, O., Pereiro, A. X., & Rodríguez, M. S. (2005). Narrative speech in aging: Quantity, information content, and cohesion. *Brain and Language, 95*(3), 423–434. https://doi.org/10.1016/j.bandl.2005.04.001

Kavé, G., & Goral, M. (2017). Do age-related word retrieval difficulties appear (or disappear) in connected speech? *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition, 24*(5), 508–527. https://doi.org/10.1080/13825585.2016.1226249

Kavé, G., & Yafé, R. (2014). Performance of younger and older adults on tests of word knowledge and word retrieval: Independence or interdependence of skills? *American Journal of Speech-Language Pathology, 23*, 36–45. https://doi.org/10.1044/1058-0360(2013/12-0136

Kertesz, A. (1982). *Western Aphasia battery*. New York, NY: Grune & Stratton.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Kong, A. P.-H., Cheung, C. Y.-N., & Wong, C. W.-Y. (2025). Establishing norm of connected speech measures for descriptive discourses in Cantonese-speaking adults. *International Journal of Language & Communication Disorders, 60*, Article e70055. https://doi.org/10.1111/1460-6984.70055

Leeper, L. H., & Culatta, R. (1995). Speech fluency: Effect of age, gender and context. *Folia Phoniatrica et Logopaedica, 47*(1), 1–14. https://doi.org/10.1159/000266337

Luzzatti, C., Willmes, K., & De Bleser, R. (1996). *Aachener Aphasie Test: Versione Italiana (Seconda Edizione); Organizzazioni speciali, Firenze*.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Malcorra, B. L. C., Wilson, M. A., Schilling, L. P., & Hübner, L. C. (2022). Lower education and reading and writing habits are associated with poorer ora discourse production in typical adults and older adults. *Frontiers in Psychology, 18*(13), Article 740337. https://doi.org/10.3389/fpsyg.2022.740337

Marangolo, P., Fiori, V., Calpagnano, M. A., Campana, S., Razzano, C., Caltagirone, C., & Marini, A. (2013). tDCS over the left inferior frontal cortex improves speech production in aphasia. *Frontiers in Human Neuroscience, 7*(539), 1–10. https://doi.org/10.3389/fnhum.2013.00539

Marcotte, K., Roy, A., Brisebois, A., Jutras, C., Leonard, C., Rochon, E., & Brambati, S. (2024). Reliability of the picture description task of the Western Aphasia Battery – Revised in Laurentian French persons without brain injury. *The Clinical Neuropsychologist, 38*(8), 1980–2008. https://doi.org/10.1080/13854046.2024.2340777

Marini, A. (2012). Characteristics of narrative discourse processing after damage to the right hemisphere. *Seminars in Speech and Language, 33*(1), 68–78. https://doi.org/10.1055/s-0031-1301164
Marini, A. (2022). Cognitive and linguistic characteristics of narrative discourse production in healthy aging. In C. Coelho, L. R. Cherney, & B. Shadden (Eds.), *Discourse analysis in adults with and without communication disorders: A resource for clinicians and researchers* (pp. 15–31). Plural Publishing, Inc.
Marini, A., & Andreetta, S. (2016). Age-related effects on language production: A combined psycholinguistic and neurocognitive perspective. In H. Harris Wright (Ed.), *Cognition, Language and aging* (pp. 55–79). John Benjamins Publishing Company.
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
Marini, A., Carlomagno, S., Caltagirone, C., & Nocentini, U. (2005). The role played by the right hemisphere in the organization of complex textual structures. *Brain and Language, 93*(1), 46–54. https://doi.org/10.1016/j.bandl.2004.08.002
Marini, A., Galetto, V., Zampieri, E., Vorano, L., Zettin, M., & Carlomagno, S. (2011). Narrative language in traumatic brain injury. *Neuropsychologia, 49*, 2904–2910. https://doi.org/10.1016/j.neuropsychologia.2011.06.017
Marini, A., Petriglia, F., D'Ortenzio, S., Bosco, F. M., & Gasparotto, G. (2025). Unveiling the dynamics of discourse production in healthy aging and its connection to cognitive skills. *Discourse Processes, 62*(6–7), 479–501. https://doi.org/10.1080/0163853X.2025.2507548
Marini, A., Petriglia, F., D'Ortenzio, S., Gabbatore, I., & Gasparotto, G. (2026). Cognitive reserve effects on discourse production processing in healthy aging. *Journal of Neurolinguistics*. , Article 101315. https://doi.org/10.1016/j.jneuroling.2026.101315
Marini, A., Spoletini, I., Rubino, I. A., Ciuffa, M., Bria, P., Marinotti, G., Banfi, G., Boccascino, R., Strom, P., Siracusano, A., Caltagirone, C., & Spalletta, G. (2008). The Language of schizophrenia: An analysis of Micro and macrolinguistic abilities and their neuropsychological correlates. *Schizophrenia Research, 105*, 144–155. https://doi.org/10.1016/j.schres.2008.07.011
Marini, A., Zettin, M., Bencich, E., Bosco, F. M., & Galetto, V. (2017). Severity effects on discourse production after TBI. *Journal of Neurolinguistics, 44*, 91–106. https://doi.org/10.1016/j.jneuroling.2017.03.005
Mazzon, G., Ajčević, Cattaruzza, T., Menichelli, A., Guerriero, M., Capitanio, S., Pesavento, V., Dore, F., Sorbi, S., Manganotti, P., & Marini, A. (2019). Connected speech deficit as an early hallmark of CSF-defined Alzheimer's disease and correlation with cerebral hypoperfusion pattern. *Current Alzheimer Research, 16*, 1–12. https://doi.org/10.2174/1567205016666190506141733
Mousley, A., Bethlehem, R. A. I., Yh, F.–C., & Astle, D. (2025). Topological turning points across the human lifespan. *Nature Communications, 16*, Article 10055. https://doi.org/10.1038/s41467-025-65974-8
Muggeo, V. M. R. (2008). Segmented: An R package to fit regression models with broken-line relationships. *R News, 8/1*, 20–25. https://cran.r-project.org/doc/Rnews/.
Nicholas, L., & Brookshire, R. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech & Hearing Research, 36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338
Oosterhuis, E. J., Slade, K., Smith, E., May, P. J. C., & Nuttall, H. E. (2023). Getting the brain into gear: An online study investigating cognitive reserve and word-finding abilities in healthy ageing. *PLoS One, 18*(4), Article e0280566. https://doi.org/10.1371/journal.pone.0280566
Paradis, M., & Libben, G. (1987). *The assessment of bilingual aphasia*. Lawrence Erlbaum Associates.
Pekkala, S. (2012). Verbal fluency tasks and the neuropsychology of language. In M. Faust (Ed.), *The handbook of the neuropsychology of language* (pp. 619–634). Blackwell Publishing Ltd.
Petriglia, F., Gasparotto, G., D'Ortenzio, S., Gabbatore, I., & Marini, A. (2025). Assessing lexical diversity and informativeness across the adult lifespan: A comprehensive investigation. *Research Methods in Applied Linguistics, 4*, Article 100276. https://doi.org/10.1016/j.rmal.2025.100276
Pistono, A., Pariente, J., Bézy, C., Pastor, J., Tran, T. M., Renard, A., Fossard, M., Nespoulous, J. L., & Jucla, M. (2017). Inter-individual variability in discourse informativeness in elderly populations. *Clinical Linguistics and Phonetics, 31*(5), 391–408. https://doi.org/10.1080/02699206.2016.1277390
Prins, R., & Bastiaanse, R. (2004). Analysing the spontaneous speech of aphasic speakers. *Aphasiology, 18*(12), 1075–1091. https://doi.org/10.1080/02687030444000534
Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *202. Proceedings of the 40th international conference on machine learning, Honolulu, HI, USA, 23–29 July 2023* (pp. 28492–28518). New York, NY, USA: PMLR.
Richardson, J. D., & Hudspeth, S. G. (2016). Main concepts for three different discourse tasks in a large non-clinical sample. *Aphasiology, 30*(1), 45–73. https://doi.org/10.1080/02687038.2015.1057891
Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (1989). The quantitative analysis of agrammatic production. *Procedure and data. Brain and Language, 37*, 440–479. https://doi.org/10.1016/0093-934x(89)90030-8
Sharp, E. S., & Gatz, M. (2011). Relationship between education and dementia: An updated systematic review. *Alzheimer Disease and Associated Disorders, 25*(4), 289–304. https://doi.org/10.1097/WAD.0b013e318211c83c
Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., Ramage, A. E., & Roberts, A. C. (2021). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *Journal of Speech, Language, and Hearing Research, 64*(11), 4366–4389. https://doi.org/10.1044/2021_JSLHR-20-00708
Stern, Y., Gurland, B., Tatemichi, T. K., Wilder, D., & Mayeaux, R. (1994). Influence of education and occupation on the incidence of Alzheimer's disease. *JAMA, 271*(13), 1004–1010. PMID: 8139057.
Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Allyn and Bacon.
Thornton, R., & Light, L. L. (2006). Language comprehension and production in normal aging. In J. E. Birren, & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 261–287). Academic Press.
Ulatowska, H. K., Freedman-Stern, R., Doyel, A. W., Macaluso-Haynes, S., & North, A. J. (1983). Production of narrative discourse in aphasia. *Brain and Language, 19*(2), 317–334. https://doi.org/10.1016/0093-934x(83)90074-3
Weiss, E. M., Ragland, J. D., Brensinger, C. M., Bilker, W. B., Deisenhammer, E. A., & Delazer, M. (2006). Sex differences in clustering and switching in verbal fluency. *Journal of the International Neuropsychological Society, 12*, 502–509. https://doi.org/10.1017/s1355617706060656
Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). _dplyr: A grammar of data manipulation_. *R package version 1.1.4*. https://CRAN.R-project.org/package=dplyr.
Wickham, H., Hester, J., & Bryan, J. (2023). _readr: Read rectangular text data_. *R package version 2, 1*(4). https://CRAN.R-project.org/package=readr.
Wickham, H., Miller, E., & Smith, D. (2023). _haven: Import and Export 'SPSS', 'Stata' and 'SAS' files_. R package version 2.5.4. https://CRAN.R-project.org/package=haven.