Automated Measures of Syntactic Complexity in Natural Speech Production: Older and Younger Adults as a Case Study

Galit Agmon* [1], Sameer Pradhan [2], Sharon Ash [1], Naomi Nevler [1], Mark Liberman [2], Murray Grossman† [1], Sunghye Cho [2]

1. *Frontotemporal Degeneration Center, Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA*
2. *Linguistic Data Consortium, University of Pennsylvania, Philadelphia PA*

ORCID IDs: Galit Agmon: 0000-0001-9588-5953
Sameer Pradhan: 0000-0002-5537-2181
Sharon Ash: 0000-0002-6656-1842
Naomi Nevler: 0000-0001-8745-5945
Mark Liberman: 0000-0002-8605-9024
Murray Grossman: 0000-0002-7477-6218
Sunghye Cho: 0000-0003-1569-7608

All authors have nothing to disclose.

This work was supported by: Department of Defense [W81XWH-20-1-0531 (MG, NN, GA)]; National Institute of Aging [AG073510-01 (NN), AG066597 (MG)]; Alzheimer's Association [AARF-21-851126 (SC)].

Keywords: syntactic complexity, aging, natural language processing (NLP), automatic speech recognition (ASR)

† Deceased.

* Corresponding author at:

Anatomy-Chemistry building, room 219

3620 Hamilton Walk, Philadelphia PA 19104

galit.agmon@pennmedicine.upenn.edu

Abstract

Purpose: Multiple methods have been suggested for quantifying syntactic complexity in speech. We compared the performance of eight automated syntactic complexity metrics to determine which best captured differences in syntactic complexity between two age groups.

Method: We used natural speech samples produced in a picture description task by younger (n=76) and older (n=36) healthy participants, manually transcribed and segmented into sentences. We manually verified that older participants produced fewer complex structures. We developed a metric of syntactic complexity using automatically extracted syntactic structures as features in a multi-dimensional metric. Then, we compared our methods to seven other different methods: Yngve score, Frazier score, Frazier-Roark score, d-level, syntactic frequency, mean dependency distance and sentence length. We examined the success of each method in distinguishing the age group of speakers using logistic regression models. We repeated the same analysis with automatic transcription and segmentation using an ASR system.

Results: Our multi-dimensional metric was successful in predicting age group (AUC=0.87), and it performed better than all the other metrics. High AUCs were also achieved by Yngve score (0.84) and sentence length (0.84). However, in a fully automated pipeline with ASR, their performance dropped, while the performance of the multi-dimensional metric remained high.

Conclusions: Syntactic complexity in spontaneous speech can be quantified by directly assessing syntactic structures. It can be derived automatically, saving considerable time, cost and effort compared to manually analyzing large-scale corpora, while maintaining high face validity and parsimony.

1. Introduction

Words in a sentence do not come in a random order. They are systematically organized by a language's syntax, rules by which words can be combined to create larger units of meaning. Native speakers' implicit knowledge of syntax is assumed to be a basic cognitive capacity (Chomsky, 1980; Fodor et al., 1974). Therefore, studying syntax has been focal in psycholinguistics and neurolinguistics, where researchers have been trying to link syntactic structures with online language processing, focusing mostly on comprehension (Grodzinsky et al., 2021; Grodzinsky & Friederici, 2006; Lewis & Phillips, 2015). In particular, syntactic processing has been associated with cognitive measures such as reaction times, accuracy rates, and brain activation, providing an index of complexity (Cooke et al., 2002; Friederici et al., 2002; Ben-Shachar et al., 2003; Wingfield et al., 2003; Grodzinsky and Santi, 2008 among many others).

Cognitive methods for assessing individual linguistic capacity are challenging to implement when studying speech production. Yet, assessment of linguistic capacity is an important goal when it concerns clinical populations (Ash & Grossman, 2015), when linguistic capacity has deteriorated or is impaired (Friedmann, 2002; Grodzinsky, 1986; Grodzinsky et al., 1999; Zurif et al., 1993). Analyzing language production, particularly in spontaneous speech, offers new ways for assessing linguistic capacity at the individual level. Previous literature has shown that syntactic complexity in language production can be quantified and is useful for assessing neural pathologies that affect language in general and syntax in particular (Calzà et al., 2021; Eyigoz et al., 2020; Fraser et al., 2015; Roark et al., 2007, 2011; Silva et al., 2022; Tavabi et al., 2022).

To make methods for syntactic complexity applicable to a large-scale dataset, we focus on automated methods. Automated scoring systems have been previously developed to assess proficiency or coherence in language learning or language development (Channell, 2003; L. Chen et al., 2018; Graesser et al., 2014; Hassanali et al., 2014; Kyle, 2016; X. Lu, 2009, 2010; McNamara et al., 2014; Polio & Yoon, 2018; Sheehan et al., 2014; Yoon et al., 2020; Zechner et al., 2017). Although these automated methods often contain a grammatical component, they are less geared towards detecting fine syntactic distinctions, which is the focus of our current study. In particular, subtle changes in syntax can be a result of cognitive decline due to healthy aging or pathological degeneration. To this end, we compared seven of the most frequently employed methods of quantifying syntactic complexity in spontaneous speech and one novel metric that we developed. We used known and verified syntactic differences between two age groups as a test case, based on the well-attested decline in the processing of syntax in older persons (Burke & Shafto, 2008; Kemper et al., 2003; Kynette & Kemper, 1986; Obler et al., 1991; Peelle, 2019; Poulisse et al., 2019; Zhu et al., 2018; Zurif et al., 1995). A well-performing metric is expected to be sensitive to the decrease of complex syntactic structures in the older participants' speech and to allow accurate predictions of the age of the speaker.

1.1. Quantifying syntactic complexity

According to phrase structure grammar, sentence structure is hierarchical: words are combined into phrases, which are combined to form larger phrases, through a recursive set of rules (Bar-Hillel, 1953; Chomsky, n.d.; Hauser et al., 2002). The syntactic integration of words into phrases and sentences is cognitively costly (Brennan et al., 2016; Nelson et al., 2017), and therefore it is assumed that the degree of the cognitive cost for these syntactic integrational processes can be quantified from the sentence structure itself (e.g., T-unit length, Yngve score, Frazier score, mean dependency distance; see below). Other metrics assign a complexity score to characteristics of identified rules or structure, such as their frequency of use (Kyle & Crossley, 2017; Rezaii et al., 2022) or their expected age of acquisition (Botel & Granowsky, 1972; Lee, 1974; Rosenberg & Abbeduto, 1987;

Scarborough, 1990). For comparison to all these unidimensional scores, we developed a method that assessed individual complex syntactic structures and used them in a multi-dimensional model (see 1.1.8). We explain below and in Fig. 1 the metrics that we employed.

1.1.1. **Utterance length**: Syntactic complexity is correlated with the length of the utterance, as complex syntactic structures inevitably require more words (Ferrer-i-Cancho & Liu, 2014; Mandel Glazer, 1974; J. W. Miller & Hintzman, 1975; Szmrecsanyi, 2004). Utterance length on its own does not necessarily reflect syntactic complexity, because length can theoretically be increased without increasing complexity (e.g., by conjoining words). However, it has been used as a simple proxy for syntactic complexity (Nutter, 1981; O'Donnell, 1974; Pallier et al., 2011; Szmrecsanyi, 2004). Reduced utterance length both in writing and in speech has been shown to be associated with Alzheimer's disease (Kemper et al., 1993; Pakhomov et al., 2011) and with healthy aging (Cheung & Kemper, 1992).

1.1.2. **Yngve score**: This model was developed by Victor Yngve, a pioneer in computational linguistics, to reflect syntactic complexity based on the hierarchical phrase structure of the sentence (Yngve, 1960). Yngve's system assigns a score to each node in the hierarchy, to reflect the word-by-word short-term memory cost during the representation build-up in a top-down left-to-right traversal (Fig. 1a and Supp. Material). The total score per utterance is usually taken as the average of the word-level scores. The Yngve score has been shown to be reduced in older people (Cheung & Kemper, 1992; Kemper et al., 2001; Kemper & Rash, 1988) and in states of dementia (Fraser et al., 2015; Pakhomov et al., 2011; Roark et al., 2011).

1.1.3. **Frazier score**: Like the Yngve score, the method suggested by Frazier (1985) also relies on the hierarchical phrase structure representation of the sentence. The scoring of the tree nodes in Frazier's method is through a bottom-up traversal that examines the incremental built-up of the phrase structure representation (Fig. 1b). Each additional word in the sentence is scored by the number of nodes that it introduces in the partial representation. Sentence complexity increases when a large number of nodes are introduced within a short interval (~3 words). Although Frazier's scoring system was intended to quantify syntactic complexity in comprehension, it has also been shown to decrease in speech production during healthy aging (Cheung & Kemper, 1992).

1.1.4. **Frazier-Roark score**: A variation on Frazier's score takes the average of all word-level scores rather than just considering short intervals within the sentence (Fig. 1b). To highlight the fact that this score is a variation on Frazier’s original proposal (see Discussion), and since we were able to track its usage only to Roark et al. (2007), Roark et al. (2011) and Pakhomov et al. (2011), we termed it the Frazier-Roark score.

1.1.5. **Mean dependency distance (MDD)**: MDD reflects the average distance between related words in a sentence, and it is derived from Dependency Grammar (DG), which is an alternative way of representing the structure of a sentence (Hudson, 1984; Mel’čuk, 1988; Tesnière, 2015). Unlike in phrase structure grammar, words in DG are not grouped into constituents, but rather, they are related to other individual words in an asymmetrical relationship, called a head-dependent relationship (Fig. 1c). A dependency distance is defined as the linear distance between a dependent word and its head. The arithmetic average of all dependency distances in one sentence is the sentence's mean dependency distance (MDD) (H. Liu, 2008). MDD is based on the idea that it is easier to integrate syntactically related words when they are closer to each other (Gibson, 1998, 2000; Gibson & Pearlmutter, 1998). Previous studies have shown that MDD is increased for certain complex syntactic structures (M. X. Collins, 2014; Hudson, 1995; Jaeger & Tily, 2011) and have suggested that a larger MDD is associated with increased cognitive demands (Gildea & Temperley, 2010; Hudson, 1995; Lin, 1996; H. Liu, 2008; H. Liu et al., 2017). Reduced MDD in dementia has been attested (Aronsson et al., 2021; Pakhomov et al., 2011), although some reports have produced conflicting findings (Fors et al., 2018; Orimaye et al., 2017).

1.1.6. **Syntactic Frequency**: A different approach from computing a complexity metric out of the tree structure itself is to assign a score to the structure based on external features. One of these features is the frequency of use, which was implemented by Rezaii et al. (2022) to demonstrate reduced syntactic complexity in speech production of patients with primary progressive aphasia. In this method, syntactic rules are extracted from the DG representation of the sentence (Fig. 1d) and assigned frequency scores that were previously derived from an analysis of a large corpus (see Supp. Material for additional information).

1.1.7. **D-level**: In this scoring system for developmental level syntactic complexity (d-level), the sentence is given a score based on the expected developmental stage of its syntactic structures in language acquisition (Fig. 1e). The scale was developed by Rosenberg & Abbeduto (1987), revised by Covington (2006), and fully automated by Lu (2009). D-level was shown to decline in healthy aging and in dementia (Cheung & Kemper, 1992; Kemper et al., 2001; Kemper & Sumner, 2001).

1.1.8. **Syntactic Structures**: We developed a novel metric, which instead of extracting one single number to represent syntactic complexity, examines multiple complex syntactic structures multi-dimensionally. These syntactic structures include subordination, center embedding, relative clauses and modification in noun phrases and adjectival phrases (Fig. 1f). **Subordination** is the embedding of a clause within another clause. It is cognitively effortful, as corroborated by cognitive studies on language comprehension and by clinical studies on production in older adults and in agrammatic aphasia (Cheung & Kemper, 1992; Friedmann, 2001, 2006; Friedmann & Grodzinsky, 1997; Holmes et al., 1987; Kemper, 1986, 1987a; Kemper et al., 2003; Shetreet et al., 2009). Previous studies have even used the total number of clauses per sentence as an index of syntactic complexity (Beaman, 1984; C. Lu et al., 2019; Szmrecsanyi, 2004).

A relative clause is a particular case of subordination: a relativized noun appears at the head of a relative clause, yet it is semantically interpreted within the relative clause (Fig. 1f, blue). Notice, for example, that in the sentence "The mother, who is washing dishes, is not aware …", the word "mother" is interpreted twice: as the subject of "washing dishes" and as the subject of “not aware”. Such constructions are cognitively costly (Ben-Shachar et al., 2003, 2004; Kaan et al., 2000; Kluender & Kutas, 1993; Lau & Tanaka, 2021). Older adults perform more poorly than younger adults in processing such constructions (Baum, 1993). Difficulties of agrammatic aphasia patients in processing relative constructions are also reported (Caramazza & Zurif, 1976; Grodzinsky, 1986, 1995; Zurif et al., 1993).

Finally, although an embedded clause usually comes after the main clause (final embedding), this is not always the case, as it can be embedded within the main clause, in a construction called **center embedding**[1] (Fig. 1f, green). When processing a subordinate clause while the main clause has not been concluded yet, working memory load increases (Caplan et al., 1998; G. A. Miller & Isard, 1964; Pattamadilok et al., 2016). In particular, older adults perform worse than younger adults in recall tasks of such constructions (Kemper, 1987b; Norman et al., 1991). Note that relative clauses and centrally embedded clauses are special types of subordinate clauses (others include complement clauses and adverbial clauses).

Finally, cognitive cost can emerge through **word integration** below the clause level, such as when adjectives modify nouns (Poortman & Pylkkänen, 2016; Pylkkänen, 2019; Ziegler & Pylkkänen, 2016). There is evidence such integrational processes are affected by aging (Huang et al., 2012).

---

[1] More accurately, "left-branching" is the more general term for both center embedding and initial embedding. Cases of left-branching can emerge either by subordination or by generation of other heavy phrases, such as noun phrases or prepositional phrases (e.g., Stallings & MacDonald, 2011). To keep nomenclature as simple as possible, we will use the term "center embedding" to refer to all cases of left-branching.

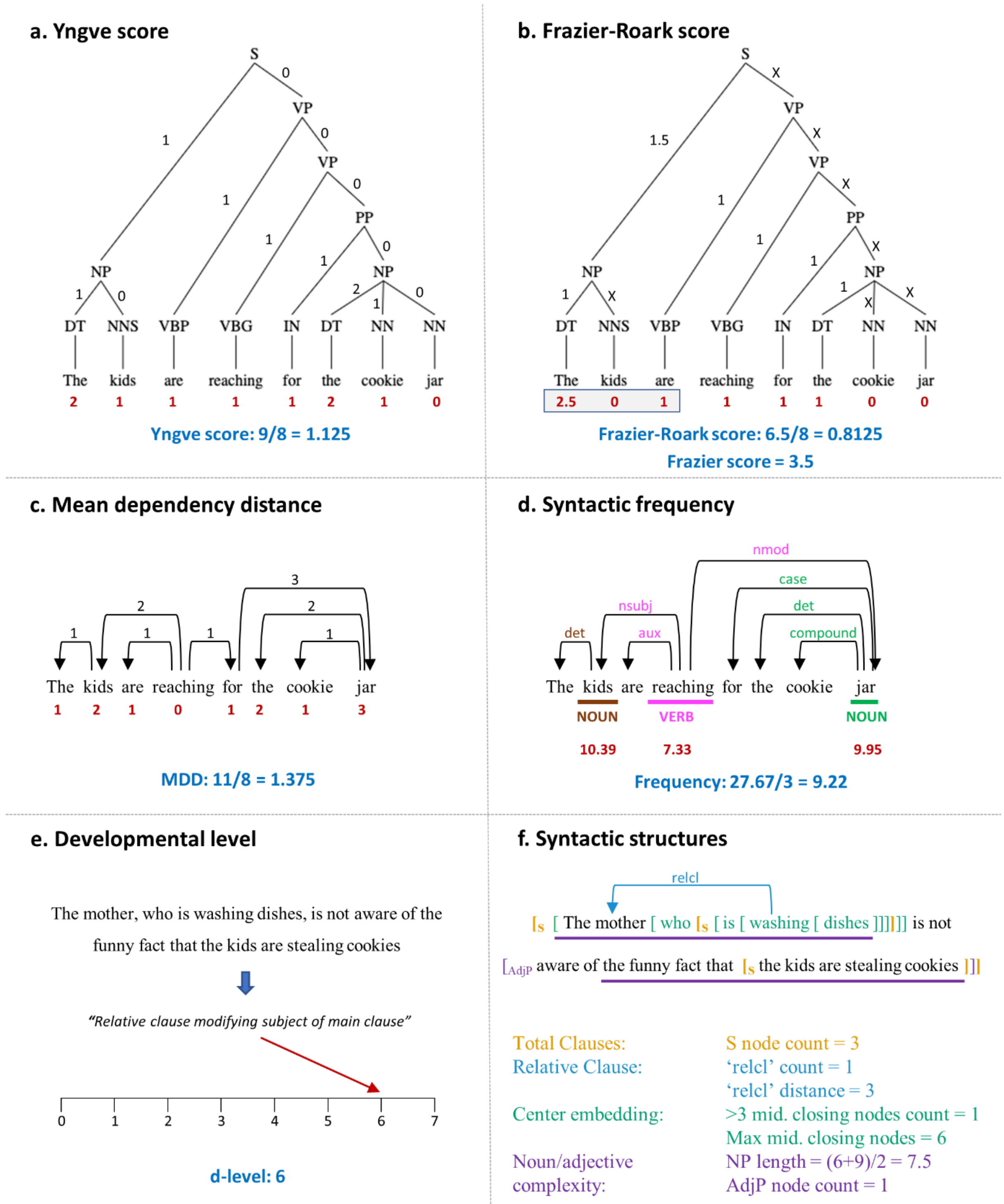

Figure 1: The calculation of complexity metrics. The first row (a, b) depicts a phrase structure tree representation of the syntactic structure of the sentence "The kids are reaching for the cookie jar". Node annotations are abbreviations of the Penn Treebank Part-of-Speech Tags (Bies et al., 1995) (e.g., S=sentence, NP=Noun Phrase, VP=Verb Phrase). The second row (c, d) depicts the same sentence in a dependency grammar representation. Each arrow represents a dependency between a head (plain end) and its dependent (pointed end). **(a)** The Yngve score assigns a score to each node of the number of its right siblings (siblings=nodes that share a parent). The path of each word is defined as the nodes that connect that word to the root of the tree (the S node). The score of each word is the sum of scores along its path (red), and the score of the sentence is the average of the word-level scores (blue). **(b)** The Frazier and Frazier-Roark scores assign a score of 1 to any node with no left siblings. When this node is headed by an S, the score is 1.5. The x symbol represents a node without a score.

The path of the word is defined as all the nodes that connect that word with either the root of the sentence (S) or the first x symbol. As with the Yngve score, the word-level score is the sum of scores along its path (red). The Frazier-Roark sentence-level score is the average of the word-level scores. The Frazier sentence-level score is the maximum of the sums of the word-level scores of every three adjacent words (here, the first three words, in the frame). **(c)** We used SpaCy's output for dependency grammar representation. The dependency distance of each word is defined as the number of words it is separated from its head. The word-level dependency distances (red) are averaged to obtain the sentence-level MDD score (blue). **(d)** We used enhanced dependency to match the Stanford Enhanced Universal Dependencies representation (Schuster & Manning, 2016) used by Rezaii et al. (2022). A rule is defined as a head (underlined word) with all its dependency relations. Each rule (color-coded) was given a frequency score. The frequency scores for each rule were averaged to calculate sentence-level scores. **(e)** According to the revised scale of expected developmental stage (X. Lu, 2009), syntactic features of a sentence are located on a scale from 0 to 6 (a higher score being a later acquisition stage). If two features of different developmental stages co-occur in the same sentence, that sentence is given a score of 7. To obtain the individual level scores, all sentences' d-level scores were averaged. **(f)** A simplified phrase structure representation. Phrases are represented here with brackets rather than tree nodes. Subscripts on opening brackets represent the node label. From this representation, we extracted the number of S nodes (orange, 3). We counted the number of relative clause ('relc') dependencies (1) and their average dependency distance (3). Heavy phrases associated with center embedding (green) were detected by counting the number of closing nodes. Modifications on the noun and adjective level (purple) were quantified by averaging the length of noun phrases which are not embedded under another noun phrase (purple underline), and by counting the number of AdjP (adjectival phrase) and AdvP (adverbial phrase) nodes per sentence.

In sum, all metrics have previously been shown to be sensitive to aging or dementia. The multiplicity of metrics for quantifying syntactic complexity calls for investigating the relationship among them. To test the different metrics, we focused on age-related differences. In this study, we used cohorts of old and young healthy speakers, where we manually identified and labeled syntactic differences, and we tested which of the above metrics was the most successful in capturing these differences.

2. Methods

2.1. Participants

We examined speech samples produced by two groups, a group of young adults (n=76) and a group of older healthy participants (n=36). Demographic characteristics of the participants are summarized in Table 1. The younger participants were mostly undergraduates at the University of Pennsylvania. The older participants were mostly caregivers of patients at the Frontotemporal Degeneration Center of the Hospital of the University of Pennsylvania. None of the older participants reported any hearing or speaking difficulties, nor did they report any medical conditions that could have interfered with their speech such as stroke, closed head injury, brain surgery or hypothyroidism. All reported being native speakers of English (two participants from the older group did not provide primary language). The young participants have not completed yet their bachelor's degree and therefore had fewer years of formal education compared to the older group (13.5 vs. 15.8). We previously used the same dataset to test the possibility of applying automated acoustic and lexical pipelines in studying natural, spontaneous speech (Cho et al., 2021).

Table 1: Demographic characteristics of the participants

Characteristic	Older (n=36)	Younger (n=76)	*p*
Age (y)			
Mean±SE	67.9±1.3	20.0±0.1	<.001
Range	53-89	18-22	
Sex (M)			
Count	11	40	.03
Percentage	31%	53%	
Education (y)			
Mean±SE	15.8±0.4	13.5±0.1	<.001
Range	12-20	11.5-15.5	

2.2. Task

All participants were asked to describe the Cookie Theft picture, a picture of a mother washing dishes while two children are stealing cookies from the cookie jar behind her. This picture is part of the clinical protocol of the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983). Participants described the picture for 70 seconds on average. The younger participants were recorded while sitting in a quiet booth. The older participants were recorded by an interviewer sitting with them in the same room. The Institutional Review Board of the University of Pennsylvania approved the study of human participants, and all participants provided written consent to participate in the study.

2.3. Transcription and preprocessing

The audio files were transcribed in two ways, manually and automatically. For the manual pre-processing, all audio files were transcribed by trained annotators and a linguist (SA). Fillers ("um", "uh"), repetitions, partial words and false starts were manually flagged during transcription and later removed from the analysis. All transcripts were then manually segmented into utterances, defined as a predicate in an independent clause with all its arguments and adjuncts (also known as a T-unit (Hunt, 1965)). This was used as the basic unit for our syntactic complexity analysis. The segmentation into utterances (T-units) was done by a trained linguist (GA) and reviewed by a second trained linguist (SA). The categorization of clauses was discussed and agreed upon by the two linguists. All transcripts included punctuation marks (commas, hyphens, and a period at the end of each utterance).

To compare with manual pre-processing, we also implemented a fully automated pipeline, using a state-of-the-art automatic speech recognition (ASR) system, OpenAI's Whisper (Radford et al., 2022). Whisper is a speech-to-text algorithm that automatically transcribes audio files as text. The transcribed output is clean of disfluencies, and it also includes punctuation marks (periods and commas), which allowed us to automatically segment the transcript into utterances (sentences) based on the position of the period. For the automated transcription and segmentation, we used Whisper's medium model, which includes 769M parameters and transcribes with a word error rate (WER) of 2.7%-43.0% (average of 12.5% across multiple types of speech), implemented through the python package whisper (https://pypi.org/project/openai-whisper/).

The cleaned and segmented transcript provided the utterances (T-units in the manual pre-processing, sentences in the automated pre-processing) that served as input to the automatic parsing. For meaningful parsing, we considered only utterances that were at least 2 words long. 1-word utterances were exclamations with no syntax, such as “yes”, “okay” or "great" and were produced only by the older group, probably as a pragmatic signal to the interviewer who was present in the room. We also performed the same analyses after excluding all utterances that were shorter than three words. Results from this second analysis did not differ qualitatively from the first one, so we report in this paper only the results of the first analysis with all utterances of 2 or more words. See Supplemental Table S1 for a summary of results of the analysis with utterances of 4 or more words.

2.4. Automated parsing

The syntactic structure of utterances was automatically analyzed using two different parsers: a dependency parser and a phrase structure parser. To obtain the dependency structure, we processed the speech data samples using spaCy 3.2.2 (Honnibal & Johnson, 2015; https://spacy.io), an NLP library in Python, using one of its largest language models for English ("en_core_web_lg"). To obtain the phrase structure, we used the Charniak-Johnson Parser, which performed N-best parse fusion (Charniak & Johnson, 2005; Choe et al., 2015), implemented through the python package bllipparser (Johnson & Charniak, 2006). From these parses, we extracted our automated syntactic measures, described in the following section.

2.5. Syntactic complexity scores derived by unidimensional metrics

We followed the algorithms that were used in previous studies to measure these metrics. Please find a general description in Section 1.1 and in Fig. 1. For a detailed description, see Supplemental Material.

2.6. Syntactic complexity scores derived by measuring syntactic structures

We compared the seven previously described unidimensional metrics with a novel multi-dimensional metric, for which we automatically approximated the prevalence of the four complex syntactic structures in the transcripts, using seven features that were automatically extracted from the phrase structure and dependency representations (GA, SP and SC).

a) Total clauses: We automatically counted the number of S nodes per utterance from the output of the phrase structure parser and averaged this number across utterances to obtain a score per subject. Included in this count are all nodes labeled as S, SQ and SINV. Notice that this number included the main clause in addition to the subordinate clauses, as the main clause was also marked with an S tag.

b) Relative clauses: We automatically counted the number of relative clauses (marked with a 'relcl' label) from the output of the dependency parser, then averaged this number across utterances to obtain a score per subject. Since ‘relc’ is not assigned in cases of headless WH-clauses (e.g., “I know [what this is supposed to be]”), we complemented this measure by counting the number of WHNP nodes from the phrase structure parser. In addition to counting the number of relative clauses, we extracted the distance associated with the 'relcl' label, assuming that a longer distance should be associated with increased complexity (Cooke et al., 2002; Fiebach et al., 2002; Grodzinsky & Santi, 2008; Lau & Tanaka, 2021; Müller et al., 1997) and particularly with lower scores for older adults (Davis & Ball, 1989; X. Liu & Wang, 2019). We averaged these distances within utterances (in case there was more than one relative clause in an utterance), and then averaged across all utterances where the parser identified a relative clause (i.e., that had a ‘relcl’ label), to compute the relative clause distance per subject.

c) Center embeddings: We assessed initial and center embedding in an utterance by examining the number of closed nodes per word, as obtained from the phrase structure parser (Fig. 1f, green). For each utterance, we calculated the number of nodes that were closed by each word (excluding the last word), assuming that closing a syntactic node is a source of cognitive effort (Brennan et al., 2016; Nelson et al., 2017). A large number of closed nodes in a non-final position in a sentence should indicate a heavy phrase in the beginning or middle of the sentence. To count the number of centrally embedded constructions, we employed a threshold of 3 on the number of mid-utterance closing nodes. We experimented with other threshold values and chose 3 because a smaller threshold captured many simple noun phrases that were not considered center embeddings. "A big kid", for example, is a phrase where the word "kid" closes 3 nodes. A higher threshold missed many cases of short center embeddings, thus increasing the chance of having a floor effect on this measure. For example, in "the woman [who [is [the [mother]]]] is washing a dish", the word “mother” closes 4 nodes, marked by having 4 right brackets. In addition to counting center embeddings as defined above, we calculated the maximal number of mid-utterance closing nodes as an approximation of the depth of a centrally embedded phrase in an utterance, assuming that deeper center embeddings result in increased complexity. We averaged the depths of center embeddings across the relevant utterances to compute scores per subject.

d) Complex NP and adjectival modifications: We extracted three features that reflect the level of nominal, adjectival and adverbial modification in a sentence. For noun phrases, we extracted all of the NPs that were not embedded under another NP and counted the number of words. We then averaged this number within utterances and across utterances to obtain an individual-level score. For adjectival and adverbial phrases, we counted the number of AdjP and AdvP nodes in each utterance and then averaged this number across utterances to obtain a score per individual.

2.7. Validation of syntactic differences and automated measures

We first verified true differences in syntactic structures between the two groups. Subordinate clauses, and in particular relative clauses and centrally embedded (or initially embedded[2]) constructions were manually identified by the two linguists (GA and SA). We averaged these

[2] Initially embedded constructions included an initial subordinate clause followed by a main clause, topicalized noun phrases and fronted prepositional phrases.

counts across utterances to get the manual scores of total clauses, relative clauses and center embeddings. We then compared the scores of manual measurements of total clauses, relative clauses and center embeddings by group. The distributions could not be considered normal due to the lower bound at zero. Hence, significance was assessed using one-tailed Mann-Whitney tests. When the directionality of the effect was not expected (i.e., higher complexity for older adults), we ran a two-tailed Mann-Whitney as a post-hoc test. Due to the slight sex imbalance between the groups, we also adjusted for sex-related differences by including sex as a covariate in a regression analysis. Since sex did not turn out to be significant and did not change the significance of the syntactic scores compared to the Mann-Whitney tests, we report only the latter in the Results section.

To test the validity of the multi-dimensional Syntactic Structures method, we correlated the syntactic structures that were derived automatically with their manual counterparts (if available), using Spearman correlations to avoid susceptibility to extreme scores. To test the validity of the unidimensional automated metrics, we tested for group differences, using one-tailed Mann-Whitney tests (assuming higher scores for younger speakers for all metrics but frequency).

2.8. Statistical Analysis

We examined the correlations among the different metrics. Note that for syntactic frequency, we expected to find a negative correlation with the other metrics, since it is assumed that more complex syntax is associated with lower frequency in use (Rezaii et al., 2022). For the multi-dimensional Syntactic Structures metric, the score for the correlational analysis was taken from the predicted values (logit scores) of a logistic regression predicting Group from all the features described in Section 2.6.

Next, we tested which metric best explained age-related group differences. For this analysis, missing values of Syntactic Structures features were replaced with zeros (i.e., the average relative clause distance of a participant that produced no relative clauses was set to 0). We fitted a logistic model that predicted Group using each of the eight automated metrics: utterance length, Yngve score, Frazier score, Frazier-Roark score, MDD, syntactic frequency, D-level and the multi-dimensional Syntactic Structures. Since the multi-dimensional model was more specified than the unidimensional models, to avoid over-fitting, we employed a 5-fold cross-validation: We divided the data into 5 balanced folds and trained the data on a pool of 4 of the 5 folds. We used the parameters from the training to predict the logit scores of the fifth fold. We repeated this procedure five times, once for each of the five folds, to obtain the predicted values (logit scores) for the full data set. Model performance was assessed by the area under the curve (AUC) of the receiver-operating characteristic (ROC), provided by R's pROC package (Robin et al., 2011). We calculated the AUC of the logit scores for each fold, from which we calculated the mean and standard deviation of the AUC for the metric. We performed this analysis twice: one time with transcripts that were manually pre-processed and a second time with transcripts that were automatically transcribed and segmented into sentences using ASR.

3. Results

3.1. Validation of group differences in manual and automated measures

We found a significant group difference in the manual counts of syntactic structures (Fig. 2). Compared to the younger group, the older group exhibited fewer subordinate clauses per utterance (W = 781.5, $p < .001$), fewer relative clauses per utterance (W = 589.5, $p < .0001$) and fewer center embeddings per utterance (W= 828, $p < .001$).

The manual counts were significantly correlated with their automated counterparts. The automated counts **of total clauses** were strongly correlated with their corresponding manual counts ($\rho = .90$, $p < .0001$). The automated counts of **headed ('relcl') and headless (WHNP) relative clauses** were strongly correlated with the corresponding manual counts (ρ

= .93, $p < .0001$). The automated counts **of center embeddings**, which were inferred and not counted directly from the parser output, were also correlated with our manual counts of center embeddings ($\rho = .37$, $p < .0001$). The correlation between the automated and manual scores of the center embedding measures was lower than those of the other two measures, likely due in part to the floor effect in the manual count: Some participants in both age groups did not produce center embeddings according to our manual counts, while the automated counts assigned a score higher than zero in the majority of cases. After removing participants with a manual count of zero (23 [64%] old and 31 [41%] young), we obtained a stronger correlation with 58 participants ($\rho = .53$, $p < .0001$).

The group differences in counts of syntactic structures were replicated using our automated measures for all features except relative clause distance ($p < .001$ for all the others). Among those who were automatically detected as producing relative clauses, the older participants' automated score for distance was larger (3.3) than that of the younger participants (2.9). Since this was not in the predicted direction, the planned one-tailed test was not significant, but when employing post-hoc a two-tailed test, the difference turned out to be significant (W = 1073, $p = .04$).

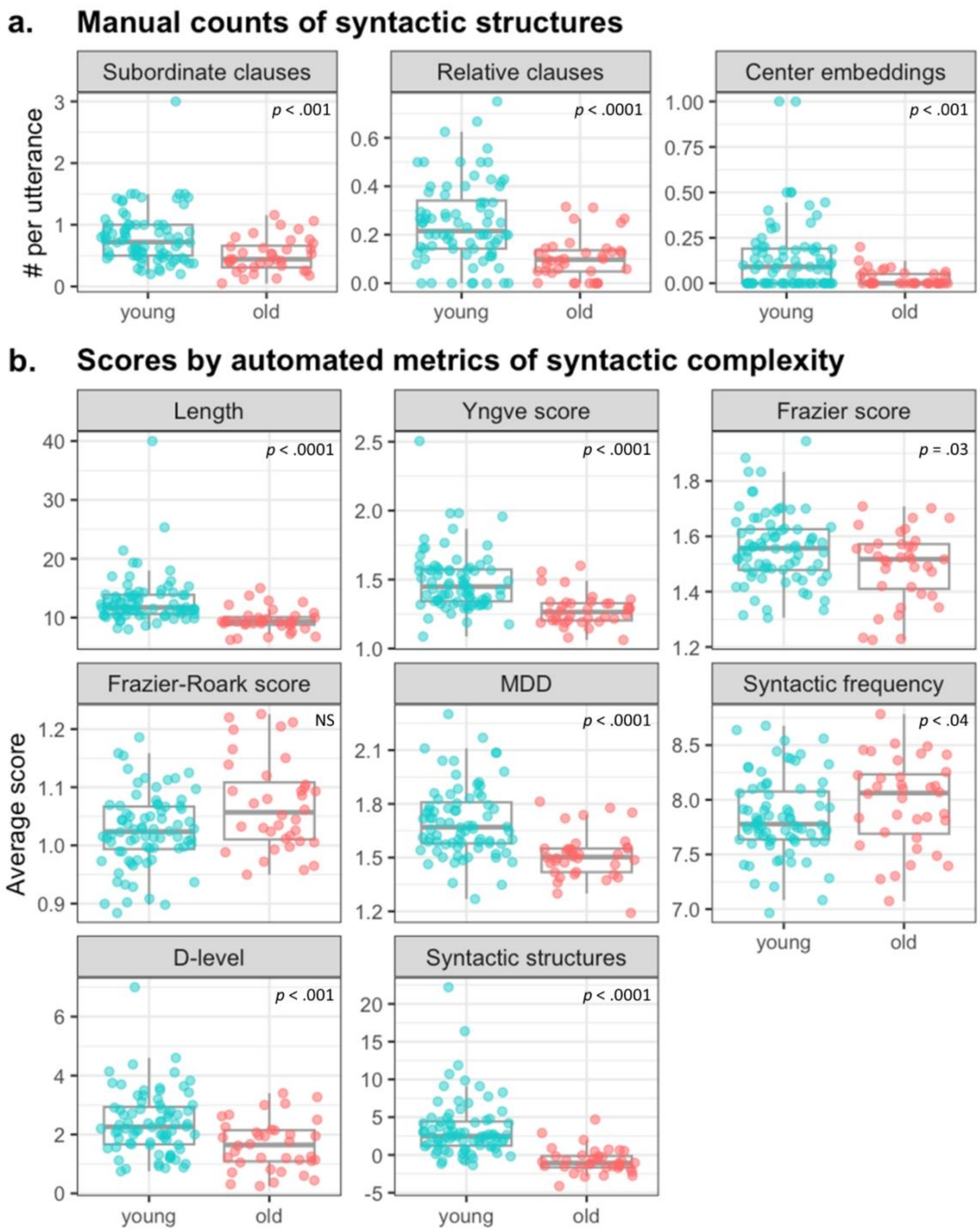

Figure 2: (a) Group differences in frequency of syntactic structures produced. Each point represents an individual. *P*-values from a one-tailed Mann-Whitney test are given. **(b) Group differences in syntactic complexity scores**. Scores are derived after manual pre-processing. Length represents number of words in a T-unit. All *p*-values are from a one-tailed Mann-Whitney test. Notice that for Syntactic Frequency, lower scores correspond to more complex syntax.

Group differences based on scores from the automated metrics were almost all in the expected direction: younger participants scored higher on utterance length (W = 424.5, $p < .0001$), Yngve score (W = 438, $p < .0001$), MDD (W = 515, $p < .0001$), d-level (W = 793, $p < .001$), Frazier score (W = 1064.5, $p = .03$), and lower on frequency (W = 1653, $p = .04$). Logit scores of Syntactic Structures also showed the expected group difference of young >

old (W = 253, $p < .0001$). Only the Frazier-Roark metric, which averages word-level scores rather than taking the maximum, showed the opposite trend, with higher scores for the older participants. Since this direction was unexpected, we tested its significance post-hoc using a two-tailed test (W = 1771, $p = .01$).

Examining the correlations between the metrics, we found that besides the Frazier-Roark score, all metrics were highly correlated with each other. The strongest correlations were between Syntactic Structures, utterance length, Yngve score and MDD. Frequency, as expected, had an inverse correlation with all the metrics, as lower complexity was expected to be associated with higher frequency.

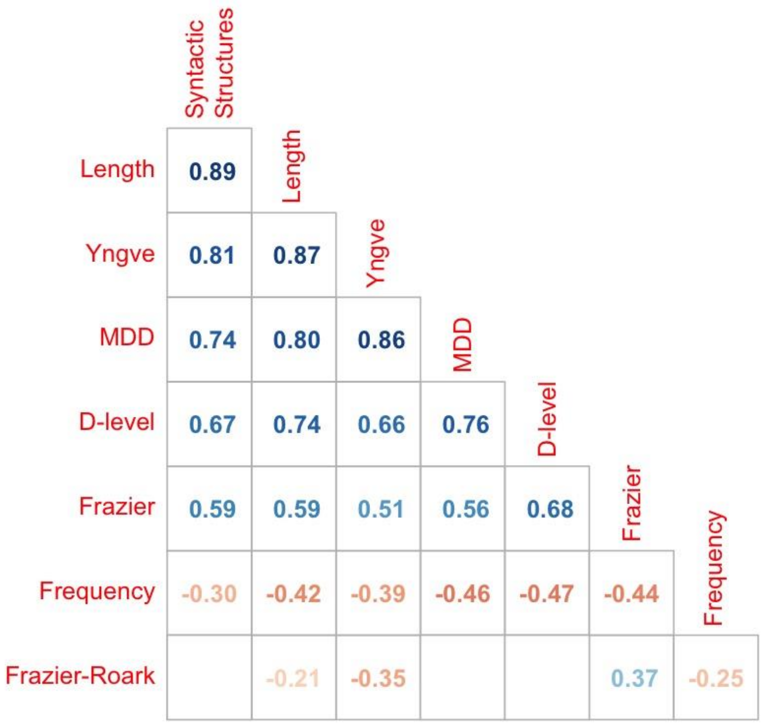

Figure 3: Correlation matrix of syntactic complexity scores. Correlations among the eight metrics, across all participants, ordered by Syntactic Structures score. For the multi-dimensional Syntactic Structures metric, scores were the weighted sum of syntactic features in logit space, weights extracted from a logistic regression that predicts Group from syntactic features. Only significant correlations ($p < .05$) are shown.

3.2. Comparing metrics of syntactic complexity

In an examination of the automated metrics, the Syntactic Structures model performed better than any of the other metrics in predicting Group, with AUC = 87.0% (Table 2). The Yngve score and T-unit length were not far behind, both with AUCs of 84.0%. In a fully automated pipeline with ASR, we also observed that the highest performance was that of the Syntactic Structures model (AUC = 78.8%). Importantly, while utterances manually defined based on T-units were significantly different between groups, sentences defined by ASR (Whisper) showed no group difference (p = 0.9). This made the performance of sentence length drop to an AUC of 46.4%. The performance of the Yngve score, the second highest performing metric, dropped to 72.5%. All the other metrics performed at less than 69%, suggesting the sensitivity of syntactic complexity metrics to the way a sentence is defined.

Table 2: Performance of the automated metrics in distinguishing between age groups: Sample mean and standard deviation of AUC, measured over the five folds of test set.

	Manual transcription and sentence segmentation		**Automatic transcription and sentence segmentation**	
	AUC	**SD**	**AUC**	**SD**
Syntactic structures	87.0%	12.9%	78.8%	19.3%
Yngve score	84.0%	8.9%	72.5%	13.2%
Sentence length	84.0%	7.8%	46.4%	25.0%
Mean dependency distance	80.8%	7.4%	68.0%	5.2%
Developmental level	71.2%	7.1%	66.3%	10.1%
Frazier-Roark score	63.8%	8.7%	67.7%	8.8%
Frazier score	60.1%	10.6%	49.8%	11.9%
Syntactic frequency	37.8%	8.6%	33.8%	8.6%

4. Discussion

Many metrics have been proposed for quantifying syntactic complexity (e.g., Covington et al., 2006; DiStefano & Howie, 1979; Frazier, 1985; Gibson, 1998; H. Liu, 2008; Rezaii et al., 2022; Scarborough, 1990; Uddén et al., 2022; Yngve, 1960). In this study we compared seven automated metrics that quantify syntactic complexity and have been shown to be associated with aging or dementia. In addition, we proposed a new multi-dimensional metric that assessed the prevalence of syntactic structures that were previously shown to be cognitively costly and found that this metric was the most sensitive of all in detecting group differences in syntactic complexity. Our metric is easy to interpret, grounded in the psycho-linguistic literature, and offers a fast and easy-to-implement protocol for the analysis of syntactic complexity in speech. Previous studies of spontaneous speech have been able to distinguish healthy participants from patients such as those with mild cognitive impairment (Calzà et al., 2021; Roark et al., 2011), Alzheimer's Disease (Eyigoz et al., 2020; Tavabi et al., 2022) and schizophrenia (Silva et al., 2022). In future work, we plan to use our automated syntactic measures to assess syntactic complexity in speech production in clinical populations with neural degeneration.

This study is consistent with past results and suggests that aging affects the syntactic complexity of language production. In line with previous literature, our cross-sectional comparison shows that the speech of older speakers contains less complex syntax, with fewer clauses, relative clauses and center embeddings per utterance. Surprisingly, the distance of relative clauses was longer for older adults, contrary to previous findings (Davis & Ball, 1989; X. Liu & Wang, 2019; Peelle et al., 2010; Wingfield et al., 2003; Zurif et al., 1995). This result, although not very strong, was still significant with an alpha of .05. Yet we were not able to replicate this finding when we tried to approximate relative clause distance manually. This issue could profitably be investigated further in future research.

A possible reason for not finding a longer distance in relative clauses of the younger group could be due to the automated method that was used. It is possible that using a dependency parser is not the best way to assess long-distance relationships, particularly for relative clauses. When a dependency parser analyzes a relative clause, it relates the relativized noun to the main verb of the relative clause. However, according to linguistic theory, the distance should be between the noun and the verb that assigns that noun its thematic role, which is not necessarily the main verb. For example, the sentence "The dishes [which <I guess the mother is cleaning>] are on the counter" contains a relative clause (square brackets), which itself contains another embedded clause that starts with "I guess" (triangular brackets). Our method of approximating the relative clause distance was to calculate the dependency distance of the 'relcl' arc, which connects "dishes" with the verb "guess". That is, it is the distance between the relativized noun ("dishes") and the *main* verb in the relative clause ("guess"), which turns out to be 3. However, according to linguistic theory, the distance that is associated with cognitive cost should be to the verb that gives the noun its semantic interpretation ("cleaning"), which is actually 7. Using dependency distance therefore truncates long distances in cases of multiply embedded sentences. To assess aging effects on the distance of relative clauses more reliably, it is important to correctly identify the constituents that are dislocated from the position where they are semantically interpreted.

4.1. Dependency Grammar: Mean Dependency Distance and syntactic frequency

Dependency distance should increase for more complex structures. Although there is not much research on the psychological reality of dependency grammar (DG) (though see Lopopolo et al. (2021)), in theory a higher MDD is associated with structures of increased syntactic complexity (M. X. Collins, 2014; Hudson, 1995). Subordination, relative clauses and center embedding all increase dependency distances, which explains the relatively well performance of MDD. However, it seems that a single-dimensional score like MDD flattens the richness of syntactic structures and washes out some of the group differences. For example, it could be that a center embedded clause is more cognitively costly than a relative clause, yet in the dependency framework, dependencies of both structures weigh similarly in their contribution to MDD. Moreover, it could even be that some variables weigh in different directions, as we report in the current study, where older participants scored lower on all measures but the distance of the relative clause. A metric like MDD, which takes into account linear distances regardless of the structure that they stem from or its depth, is liable to be weaker than a metric that considers each structure individually.

Various versions and modifications to dependency distances exist. Some suggest that the distance should not be measured linearly, but structurally, as nodes in the syntactic tree or as hierarchical distance (Baumann, 2014; R. Chen et al., 2021; Jing & Liu, 2015) or a more intricate distance measure that takes utterance length into account (Lei & Jockers, 2020). We expect these metrics to suffer from similar weaknesses for reasons discussed above, but future research might determine the usefulness of other dependency metrics in modelling syntactic complexity.

Syntactic frequency was a second metric we considered that was based on DG. Although group differences were significant and in the predicted direction, the effect was not very strong, and this metric was not very successful compared to the other metrics in predicting age group. This can be explained if we consider the psychological reality of DG, and particularly of DG rules. There are over 70,000 DG rules in Rezaii et al. (2022). From a cognitive perspective, it is unlikely that the language system is sensitive to rules or encodes rules at this level of detail. For example, relative clauses are considered a difficult structure with high cognitive cost, and therefore we should expect a high complexity score assigned to them. This score should be similar across realizations of the relative clause which are trivially different, such as whether the head has a definite article or not. However, there are multiple rules that match a relative clause in the list of rules constructed by Rezaii et al. (2022), such as *det + NOUN + acl:relcl* and *NOUN + acl:relcl*, which differ only in the presence of a determiner. Yet, each rule has its own frequency score. If frequency is indeed associated with cognitive cost, it should be evaluated with respect to rules that have a cognitive representation. As mentioned, as far as we know, the cognitive reality of DG rules has never been investigated. Future cognitive research should address this question.

4.2. Frazier score and Frazier-Roark score

The metric that performed differently from all the other measures in this study was the Frazier-Roark score. Group differences in this measure were actually in the unpredicted direction, with the older adults scoring higher than the younger adults. Moreover, this scoring system did not positively correlate with any of the other systems. A negative correlation between Frazier's score and Yngve's score was reported also in Roark (2011), who compared the two scoring systems in classifying mild cognitive impairment. The explanation for this seemingly unexpected low performance is actually quite simple: Given that by the end of the sentence all nodes are eventually introduced, then averaging all word-level scores approximates no more than the ratio between total number of nodes and total number of words. A sophisticated algorithm is not needed for simply counting the nodes and dividing them by the number of words. A node count across the entire sentence is not sensitive to the distribution of nodes within the sentence and hence is not sensitive to syntactic structures. It has even been criticized by Frazier herself (1985, p. 157): "The major problem with the nonterminal-to-terminal node ratio stems from the fact that it is not sensitive to the precise distribution of non-terminals over the lexical string."

For this reason, in this study we diverged from Roark's (2011, 2007) algorithm and computed a second version of the Frazier score which was more in the spirit of her original proposal. Yet, the Frazier score in our study, although showing the expected group differences and being correlated with the other metrics, did not perform as well as the other metrics in capturing group differences. The reason for this could be due to the fact that even our version was still not exactly what Frazier had in mind. As mentioned in the Introduction, Frazier's original proposal was to examine sentence tree representation incrementally, as it unfolds word-by-word, to explain complexity in speech *comprehension*, rather than production. Each word is scored by the number of nodes that are introduced into the partial representation at that point. Yet, current NLP parsers do not provide partial representations, and therefore our algorithm is also not the full implementation of this bottom-up incremental build-up of syntactic representations[3]. Based on our results, it seems that the Frazier score, when computed based on the final tree representation, is not a good representation of syntactic complexity in speech production.

4.3. Sensitivity to sentence definition and automatic transcription

We implemented ASR to transcribe and segment spontaneous speech automatically, and we calculated the same automated measures of syntactic complexity in order to test the possibility of fully automating the process. We confirmed that the results were similar to those produced by a semi-automated pipeline, with the Syntactic Structures metric still performing the best of all the metrics. However, we also noticed that the performance of the models that were trained with automated transcripts dropped substantially from their manually transcribed counterparts, replicating previous findings on reduced parser performance when employed on ASR output (L. Chen & Yoon, 2012; M. Chen & Zechner, 2011). While all metrics dropped in performance by 4%-38%, the performance of utterance

---

[3] For example, consider the sentence "A friend from Milwaukee came". According to the incremental proposal of Frazier, the word "a" introduces two non-terminal nodes to the partial representation ([$_S$ [$_{NP}$ a]]), since upon receiving only "a" as input, listeners can only minimally assume a noun phrase (NP) and a sentence (S). At this point, it is not yet known that "a" is actually embedded under a second noun phrase ([$_S$ [$_{NP}$ [$_{NP}$ a friend] [$_P$ from Milwaukee]]). This fact will be revealed and incorporated into the structure only later on, upon hitting the word "from". However, an algorithm based on the final tree representation scores ends up ascribing the word "a" the score of 3.5 rather than 2.5, due to that extra noun phrase.

length decreased the most (about 38%). Considering that the performance of utterance length in manually segmented transcripts showed a much higher AUC (over 80%) compared to the one trained with ASR transcripts (AUC = 46%), this result seems to suggest that utterance length in automated transcripts is not reliable enough to capture minor group differences. When utterance boundaries were not accurate, it was inevitable that the other measures of syntactic complexity were also affected. Future research on fully automating the process of measuring syntactic complexity should develop a model (ASR or NLP) that segments speech into utterances in a way that represents T-units more closely.

4.4. Limitations

There are several limitations of this study that future research needs to address. First, when comparing metrics, we used a heterogeneous set of parsers. These included the blipparser for the Frazier and Yngve scores and SpaCy for MDD. For d-level analysis, we used the algorithm of Lu (2009), which makes use of the Collins parser (M. Collins, 1996). For syntactic frequency we used SpaCy and modified its output to match the enhanced DG representation provided by the Stanford Lexicalized Parser (Klein & Manning, 2003). All these parsers may perform at different levels of accuracy and therefore might affect a fair comparison between the metrics. Although we believe that the use of different parsers should not have such a large effect as we report in this paper, future research should examine different NLP parsers to find the most accurate one for measuring syntactic complexity.

A limitation to the approach of counting syntactic structures is the risk of floor effects in cases where complex syntactic structures are not present in the input. Such floor performance could result in low sensitivity of this metric, making it less useful for monitoring pathological cases with severe syntactic deficits. Future research should consider syntactic features that can be detected even in such cases.

Finally, despite statistically robust findings, our study is limited in the conclusions that can be drawn about healthy aging. Without longitudinal data, any cross-sectional difference might be the result of generational differences. For example, it could be that the younger adults were speaking more casually, which resulted in an increase of subject relative clauses. In addition, some factors were not controlled for in our study, such as the presence of a human interviewer or years of education. Regarding education, considering that most of the younger participants would finish their BA degrees within a couple of years and all participants' education level was at ceiling given their age, we assumed that the small gap in years of education did not reflect a meaningful group difference. Future research should use larger, longitudinal samples and identical data collection methods to test how healthy aging affects syntactic complexity.

Conclusion

To evaluate heterogeneous methods of quantifying individual-level scores of syntactic complexity, we compared eight automated ways of measuring syntactic complexity. We advocate a method that considers individual structures that are known to be cognitively costly. Our implementation of syntactic complexity measures has proven useful in examining spontaneous speech samples produced by two age groups of speakers.

Data Availability Statement

Anonymized transcripts of the recordings analyzed in this study, as well as the code used to analyze them, are available from the authors on reasonable request.

References

Aronsson, F. S., Kuhlmann, M., Jelic, V., & Östberg, P. (2021). Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis. *Aphasiology*, *35*(7), 900–913. https://doi.org/10.1080/02687038.2020.1742282

Ash, S., & Grossman, M. (2015). Why study connected speech production? In R. M. Willems (Ed.), *Cognitive Neuroscience of Natural Language Use* (pp. 29–58). Cambridge University Press.

Bar-Hillel, Y. (1953). A quasi-arithmetical notation for syntactic description. *Language*, *29*(1), 47–58. https://www.jstor.org/stable/410452

Baum, S. R. (1993). Processing of center-embedded and right-branching relative clause sentences by normal elderly individuals. *Applied Psycholinguistics*, *14*(1), 75–88. https://doi.org/10.1017/S0142716400010158

Baumann, P. (2014). Dependencies and Hierarchical Structure in Sentence Processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*, 36.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen (Ed.), *Coherence in spoken and written discourse* (pp. 45–80). Praeger.

Ben-Shachar, M., Hendler, T., Kahn, I., Ben-Bashat, D., & Grodzinsky, Y. (2003). The Neural Reality of Syntactic Transformations. *Psychological Science*, *14*(5), 433–440. https://doi.org/10.1111/1467-9280.01459

Ben-Shachar, M., Palti, D., & Grodzinsky, Y. (2004). Neural correlates of syntactic movement: converging evidence from two fMRI experiments. *NeuroImage*, *21*(4), 1320–1336. https://doi.org/10.1016/j.neuroimage.2003.11.027

Bies, A., Ferguson, M., Katz, K., MacIntyre, R., Tredinnick, V., Kim, G., Marcinkiewicz, M. A., & Schasberger, B. (1995). *Bracketing guidelines for Treebank II style Penn Treebank project*.

Botel, M., & Granowsky, A. (1972). A formula for measuring syntactic complexity: A directional effort. *Elementary English*, *49*(4), 513–516.

Brennan, J. R., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., & Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, *157*, 81–94. https://doi.org/10.1016/j.bandl.2016.04.008

Burke, D. M., & Shafto, M. A. (2008). Language and aging. In F. I. M. Craik & T. A. Salthouse (Eds.), *The Handbook of Aging and Cognition* (Third edit, pp. 373–443). Psychology Press. https://doi.org/10.4324/9780203837665

Calzà, L., Gagliardi, G., Rossini Favretti, R., & Tamburini, F. (2021). Linguistic features and automatic classifiers for identifying mild cognitive impairment and dementia. *Computer Speech & Language*, *65*, 101113. https://doi.org/10.1016/J.CSL.2020.101113

Caplan, D., Alpert, N., & Waters, G. (1998). Effects of Syntactic Structure and Propositional Number on Patterns of Regional Cerebral Blood Flow. *Journal of Cognitive Neuroscience*, *10*(4), 541–552. https://doi.org/10.1162/089892998562843

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, *3*(4), 572–582. https://doi.org/10.1016/0093-934X(76)90048-1

Channell, R. W. (2003). Automated developmental sentence scoring using computerized profiling software. *American Journal of Speech-Language Pathology*, *12*(3), 369–375. https://doi.org/10.1044/1058-0360(2003/082)

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 173–180.

Chen, L., & Yoon, S.-Y. (2012). Application of structural events detected on ASR outputs for automated speaking assessment. *INTERSPEECH 2012: ISCA's 13th Annual Conference*, 767–770.

Chen, L., Zechner, K., Yoon, S.-Y., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). Automated Scoring of Nonnative Speech Using the SpeechRaterSM v. 5.0 Engine. *ETS Research Report Series*, *2018*(1), 1–31. https://doi.org/10.1002/ets2.12198

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 722–731.

Chen, R., Deng, S., & Liu, H. (2021). Syntactic complexity of different text types: From the perspective of dependency distance both linearly and hierarchically. *Journal of Quantitative Linguistics*, *29*(4), 510–540. https://doi.org/10.1080/09296174.2021.2005960

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, *13*(1), 53–76. https://doi.org/10.1017/S0142716400005427

Cho, S., Nevler, N., Shellikeri, S., Parjane, N., Irwin, D. J., Ryant, N., Ash, S., Cieri, C., Liberman, M., & Grossman, M. (2021). Lexical and Acoustic Characteristics of Young and Older Healthy Adults. *Journal of Speech, Language, and Hearing Research*, *64*(2), 302–314. https://doi.org/10.1044/2020_JSLHR-19-00384

Choe, D. K., McClosky, D., & Charniak, E. (2015). Syntactic parse fusion. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Chomsky, N. (n.d.). *Syntactic Structures*.

Chomsky, N. (1980). Rules and Representations. *The Behavioral and Brain Sciences*, *3*, 1–61.

Collins, M. (1996). A New Statistical Parser Based on Bigram Lexical Dependencies. *ArXiv Preprint*, 1–8. https://doi.org/10.48550/arXiv.cmp-lg/9605012

Collins, M. X. (2014). Information Density and Dependency Length as Complementary Cognitive Models. *Journal of Psycholinguistic Research*, *43*, 651–681. https://doi.org/10.1007/s10936-013-9273-3

Cooke, A., Zurif, E. B., DeVita, C., Alsop, D., Koenig, P., Detre, J., Gee, J., Pinãngo, M., Balogh, J., & Grossman, M. (2002). Neural basis for sentence comprehension: Grammatical and short-term memory components. *Human Brain Mapping*, *15*(2), 80–94. https://doi.org/10.1002/HBM.10006

Covington, M. A., He, C., Brown, C., Naçi, L., & Brown, J. (2006). *How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale*. http://lorinanaci.org/wp-content/uploads/2012/06/2006-01-Covington.pdf

Davis, G. A., & Ball, H. E. (1989). Effects of age on comprehension of complex sentences in adulthood. *Journal of Speech, Language, and Hearing Research*, *32*(1), 143–150. https://doi.org/10.1044/jshr.3201.143

DiStefano, P., & Howie, S. (1979). Sentence weights: An alternative to the T-Unit. *English Education*, *11*(2), 98–101. https://www.jstor.org/stable/40172289

Eyigoz, E., Mathur, S., Santamaria, M., Cecchi, G., & Naylor, M. (2020). Linguistic markers predict onset of Alzheimer's disease. *EClinicalMedicine*, *28*, 100583. https://doi.org/10.1016/J.ECLINM.2020.100583

Ferrer-i-Cancho, R., & Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, *5*(2), 143–155. https://doi.org/10.1515/GLOT-2014-0014

Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2002). Separating syntactic memory costs and syntactic integration costs during parsing: the processing of German WH-questions. *Journal of Memory and Language*, *47*(2), 250–272. https://doi.org/10.1016/S0749-596X(02)00004-9

Fodor, J. A., Bever, T. G., & Garrett, M. F. (1974). The psychological reality of grammatical structure. In *The psychology of language: an introduction to psycholinguistics and generative grammar* (pp. 221–274). McGraw-Hill.

Fors, K. L., Fraser, K., & Kokkinakis, D. (2018). Automated Syntactic Analysis of Language Abilities in Persons with Mild and Subjective Cognitive Impairment. In A. Ugon, D. Karlsson, G. O. Klein, & A. Moen (Eds.), *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth (Proceedings of MIE 2018)* (pp. 705–709). IOS Press BV. https://doi.org/10.3233/978-1-61499-852-5-705

Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2015). Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, *49*(2), 407–422. https://doi.org/10.3233/JAD-150520

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural Language Parsing* (pp. 129–189). Cambridge University Press.

Friederici, A. D., Hahne, A., & Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. *Journal of Psycholinguistic Research*, *31*(1), 45–63. https://doi.org/10.1023/A:1014376204525

Friedmann, N. (2001). Agrammatism and the Psychological Reality of the Syntactic Tree. *Journal Pf Psycholinguistic Research*, *31*, 71–90. https://doi.org/10.1023/A:1005256224207

Friedmann, N. (2002). Question Production in Agrammatism: The Tree Pruning Hypothesis. *Brain and Language*, *80*(2), 160–187. https://doi.org/10.1006/BRLN.2001.2587

Friedmann, N. (2006). Speech production in Broca's agrammatic aphasia: Syntactic tree pruning. In Y. Grodzinsky & K. Amunts (Eds.), *Broca's region* (pp. 63–82). Oxford University Press.

Friedmann, N., & Grodzinsky, Y. (1997). Tense and agreement in agrammatic production: Pruning the syntactic tree. *Brain and Language*, *56*(3), 397–425. https://doi.org/10.1006/brln.1997.1795

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based approach of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, Language, Brain: Papers from the first mind articulation project symposium* (pp. 95–126). MIT Press.

Gibson, E., & Pearlmutter, N. J. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences*, *2*(7), 262–268. https://doi.org/10.1016/S1364-6613(98)01187-5

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286–310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders* (Second edi). Lea & Febiger.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix Measures Text Characteristics at Multiple Levels of Language and Discourse. *The Elementary School Journal*, *115*(2), 210–229. https://doi.org/10.1086/678293

Grodzinsky, Y. (1986). Language deficits and the theory of syntax. *Brain and Language*, *27*(1), 135–159. https://doi.org/10.1016/0093-934X(86)90009-X

Grodzinsky, Y. (1995). Trace deletion, Θ-roles, and cognitive strategies. *Brain and Language*, *51*(3), 469–497. https://doi.org/10.1006/brln.1995.1072

Grodzinsky, Y., & Friederici, A. D. (2006). Neuroimaging of syntax and syntactic processing. *Current Opinion in Neurobiology*, *16*(2), 240–246. https://doi.org/10.1016/J.CONB.2006.03.007

Grodzinsky, Y., Pieperhoff, P., & Thompson, C. (2021). Stable brain loci for the processing of complex syntax: A review of the current neuroimaging evidence. *Cortex*, *142*, 252–271. https://doi.org/10.1016/J.CORTEX.2021.06.003

Grodzinsky, Y., Piñango, M. M., Zurif, E., & Drai, D. (1999). The Critical Role of Group Studies in Neuropsychology: Comprehension Regularities in Broca's Aphasia. *Brain and Language*, *67*(2), 134–147. https://doi.org/10.1006/BRLN.1999.2050

Grodzinsky, Y., & Santi, A. (2008). The battle for Broca's region. *Trends in Cognitive Sciences*, *12*(12), 474–480. https://doi.org/10.1016/j.tics.2008.09.001

Hassanali, K., Liu, Y., Iglesias, A., Solorio, T., & Dollaghan, C. (2014). Automatic generation of the index of productive syntax for child language transcripts. *Behavior Research Methods*, *46*, pages254–262. https://doi.org/10.3758/s13428-013-0354-x

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Holmes, V. M., Kennedy, A., & Murray, W. S. (1987). Syntactic structure and the garden path. *Quarterly Journal of Experimental Psychology*, *39A*(2), 277–293. https://doi.org/10.1080/14640748708401787

Honnibal, M., & Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1373–1378.

Huang, H.-W., Meyer, A. M., & Federmeier, K. D. (2012). A "concrete view" of aging: Event related potentials reveal age-related changes in basic integrative processes in language. *Neuropsychologia*, *50*(1), 26–35. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2011.10.018

Hudson, R. A. (1984). *Word Grammar*. Blackwell.

Hudson, R. A. (1995). Measuring syntatic difficulty. In *Manuscript*.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processingcomplexity and communicativeefficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 323–335. https://doi.org/10.1002/wcs.126

Jing, Y., & Liu, H. (2015). Mean hierarchical distance augmenting mean dependency distance. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 161–170.

Johnson, M., & Charniak, E. (2006). *BLLIP reranking parser*. https://github.com/BLLIP/bllip-parser

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes*, *15*(2), 159–201. https://doi.org/10.1080/016909600386084

Kemper, S. (1986). Imitation of complex syntactic constructions by elderly adults. *Applied Psycholinguistics*, *7*(3), 277–287. https://doi.org/10.1017/S0142716400007578

Kemper, S. (1987a). Life-span Changes in Syntactic Complexity. *Journal of Gerontology*, *42*(3), 323–328. https://doi.org/10.1093/geronj/42.3.323

Kemper, S. (1987b). Syntactic complexity and elderly adults' prose recall. *Experimental Aging Research*, *13*(1), 47–52. https://doi.org/10.1080/03610738708259299

Kemper, S., Herman, R. E., & Lian, C. H. T. (2003). The costs of doing two things at once for young and older adults: Talking while walking, finger tapping, and ignoring speech of noise. *Psychology and Aging*, *18*(2), 181–192. https://doi.org/10.1037/0882-7974.18.2.181

Kemper, S., LaBarge, E., Ferraro, F. R., Cheung, H., Cheung, H., & Storandt, M. (1993). On the preservation of syntax in Alzheimer's Disease: Evidence from written sentences. *Archives of Neurology*, *50*(1), 81–86. https://doi.org/10.1001/archneur.1993.00540010075021

Kemper, S., & Rash, R. (1988). Speech and writing across the life-span. In P. E. Morris & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (pp. 107–112). Wiley.

Kemper, S., & Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging*, *16*(2), 312–322. https://doi.org/10.1037/0882-7974.16.2.312

Kemper, S., Thompson, M., & Marquis, J. (2001). Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content. *Psychology and Aging*, *16*(4), 600–614. https://doi.org/10.1037/0882-7974.16.4.600

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 423–430.

Kluender, R., & Kutas, M. (1993). Bridging the Gap: Evidence from ERPs on the Processing of Unbounded Dependencies. *Journal of Cognitive Neuroscience*, *5*(2), 196–214. https://doi.org/10.1162/JOCN.1993.5.2.196

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Georgia State University]. https://doi.org/10.57709/8501051

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Kynette, D., & Kemper, S. (1986). Aging and the loss of grammatical forms: A cross-sectional study of language performance. *Language & Communication*, *6*(1/2), 65–72. https://doi.org/10.1016/0271-5309(86)90006-6

Lau, E., & Tanaka, N. (2021). The subject advantage in relative clauses: A review. In *Glossa* (Vol. 6, Issue 1). Ubiquity Press. https://doi.org/10.5334/GJGL.1343

Lee, L. L. (1974). *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians*. Northwestern University Press.

Lei, L., & Jockers, M. L. (2020). Normalized dependency distance: Proposing a new measure. *Journal of Quantitative Linguistics*, *27*(1), 62–79. https://doi.org/10.1080/09296174.2018.1504615

Lewis, S., & Phillips, C. (2015). Aligning Grammatical Theories and Language Processing Models. *Journal of Psycholinguistic Research*, *44*(1), 27–46. https://doi.org/10.1007/s10936-014-9329-z

Lidz, J., & Musolino, J. (2002). Children's command of quantification. *Cognition*, *84*(2), 113–154. https://doi.org/10.1016/S0010-0277(02)00013-6

Lin, D. (1996). On the Structural Complexity of Natural Language Sentences. *Proceedings of COLING-96*, 729–733.

Liu, H. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. *Journal of Cognitive Science*, *9*, 159–191.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. In *Physics of Life Reviews* (Vol. 21, pp. 171–193). Elsevier B.V.

https://doi.org/10.1016/j.plrev.2017.03.002

Liu, X., & Wang, W. (2019). The effect of distance on sentence processing by older adults. *Frontiers in Psychology*, *10*, 2455. https://doi.org/10.3389/FPSYG.2019.02455/BIBTEX

Lopopolo, A., van den Bosch, A., Petersson, K.-M., & Willems, R. M. (2021). Distinguishing Syntactic Operations in the Brain: Dependency and Phrase-Structure Parsing. *Neurobiology of Language*, *2*(1), 152–175. https://doi.org/10.1162/nol_a_00029

Lu, C., Bu, Y., Ding, Y., Torvik, V., Schnaars, M., & Zhang, C. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology*, *70*, 462–475. https://doi.org/10.1002/asi.24126

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, *14*(1), 3–28. https://doi.org/10.1075/ijcl.14.1.02lu

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Mandel Glazer, S. (1974). Is sentence length a valid measure of difficulty in readability formulas? *The Reading Teacher*, *27*(5), 464–468. https://www.jstor.org/stable/20193535

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Mel'čuk, I. A. (1988). *Dependency Syntax: Theory and Practice*. State University of New York Press.

Miller, G. A., & Isard, S. (1964). Free recall of self-embedded english sentences. *Information and Control*, *7*(3), 292–303. https://doi.org/10.1016/S0019-9958(64)90310-9

Miller, J. W., & Hintzman, C. A. (1975). Syntactic complexity of Newberry award winning books. *The Reading Teacher*, *28*(4), 750–757. https://www.jstor.org/stable/20193907

Müller, H. M., King, J. W., & Kutas, M. (1997). Event-related potentials elicited by spoken relative clauses. *Cognitive Brain Research*, *5*(3), 193–203. https://doi.org/10.1016/S0926-6410(96)00070-5

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., & Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), E3669–E3678. https://doi.org/10.1073/pnas.1701590114

Norman, S., Kemper, S., Kynette, D., Cheung, H., & Anagnopoulos, C. (1991). Syntactic complexity and adults' running memory span. *Journal of Gerontology*, *46*(6), P346–P351. https://doi.org/10.1093/geronj/46.6.P346

Nutter, N. (1981). Relative merit of mean length of T-Unit and sentence weight as indices of syntactic complexity in oral language. *English Education*, *13*(1), 17–19.

O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech*, *49*(1/2), 102–110. https://doi.org/10.2307/3087922

Obler, L. K., Fein, D., Nicholas, M., & Albert, M. L. (1991). Auditory comprehension and aging: Decline in syntactic processing. *Applied Psycholinguistics*, *12*(4), 433–452. https://doi.org/10.1017/S0142716400005865

Orimaye, S. O., Wong, J. S.-M., Golden, K. J., Wong, C. P., & Soyiri, I. N. (2017). Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC Bioinformatics*, *18*(34), 1–13. https://doi.org/10.1186/s12859-016-1456-0

Pakhomov, S., Chacon, D., Wicklund, M., & Gundel, J. (2011). Computerized assessment of syntactic complexity in Alzheimerś disease: A case study of Iris Murdochś writting. *Behavior Research Methods*, *43*, 136–144. https://doi.org/10.3758/s13428-010-0037-9

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(6), 2522–2527. https://doi.org/10.1073/PNAS.1018711108/-/DCSUPPLEMENTAL

Pattamadilok, C., Dehaene, S., & Pallier, C. (2016). A role for left inferior frontal and posterior superior temporal cortex in extracting a syntactic tree from a sentence. *Cortex*, *75*, 44–55. https://doi.org/10.1016/j.cortex.2015.11.012

Peelle, J. E. (2019). Language and aging. In G. I. De Zubicaray & N. O. Schiller (Eds.), *The Oxford Handbook of Neurolinguistics* (pp. 295–316). Oxford University Press.

Peelle, J. E., Troiani, V., Wingfield, A., & Grossman, M. (2010). Neural processing during older adults' comprehension of spoken sentences: Age differences in resource allocation and connectivity. *Cerebral Cortex*, *20*(4), 773–782. https://doi.org/10.1093/cercor/bhp142

Polio, C., & Yoon, H.-J. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*, *28*(1), 165–188. https://doi.org/10.1111/ijal.12200

Poortman, E. B., & Pylkkänen, L. (2016). Adjective conjunction as a window into the LATL's contribution to conceptual combination. *Brain and Language*, *160*, 50–60. https://doi.org/10.1016/j.bandl.2016.07.006

Poulisse, C., Wheeldon, L., & Segaert, K. (2019). Evidence Against Preserved Syntactic Comprehension in Healthy Aging. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/XLM0000707

Pylkkänen, L. (2019). The neural basis of combinatory syntax and semantics. In *Science* (Vol. 366, Issue 6461, pp. 62–66). American Association for the Advancement of Science. https://doi.org/10.1126/science.aax0050

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust speech recognition via large-scale weak supervision*. https://cdn.openai.com/papers/whisper.pdf

Rezaii, N., Mahowald, K., Ryskin, R., & Gibson, E. (2022). A syntax–lexicon trade-off in language production. *PNAS*, *119*(25), e2120203119. https://doi.org/10.1073/pnas.212020311

Roark, B., Mitchell, M., & Hollingshead, K. (2007). Syntactic complexity measures for detecting Mild Cognitive Impairment. *BioNLP 2007: Biological, Translational, and Clinical Language Processing*, 1–8.

Roark, B., Mitchell, M., Hosom, J.-P., Hollingshead, K., & Kaye, J. (2011). Spoken language derived measures for detecting Mild Cognitive Impairment. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(7), 2081–2090. https://doi.org/10.1586/14737175.2013.856265

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Rosenberg, S., & Abbeduto, L. (1987). Indicators of linguistic competence in the peer group conversational behavior of mildly retarded adults. *Applied Psycholinguistics*, *8*(1), 19–32. https://doi.org/10.1017/S0142716400000047

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, *11*(1), 1–22. https://doi.org/10.1017/S0142716400008262

Schuster, S., & Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. *Proceedings of the Tenth International Conference on Language*

Resources and Evaluation (LREC'16), 2371–2378. https://aclanthology.org/L16-1376

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator Tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal*, *115*(2), 184–209. https://doi.org/10.1086/678294

Shetreet, E., Friedmann, N., & Hadar, U. (2009). An fMRI study of syntactic layers: Sentential and lexical aspects of embedding. *NeuroImage*, *48*(4), 707–716. https://doi.org/10.1016/J.NEUROIMAGE.2009.07.001

Silva, A. M., Limongi, R., MacKinley, M., Ford, S. D., Alonso-Sánchez, M. F., & Palaniyappan, L. (2022). Syntactic complexity of spoken language in the diagnosis of schizophrenia: A probabilistic Bayes network model. *Schizophrenia Research*, *March*. https://doi.org/10.1016/j.schres.2022.06.011

Stallings, L. M., & MacDonald, M. C. (2011). It's not Just the "Heavy NP": Relative phrase length modulates the production of heavy-NP shift. *Journal of Psycholinguistic Research*, *40*, 177–187. https://doi.org/10.1007/s10936-010-9163-x

Szmrecsanyi, B. (2004). On operationalizing syntactic complextity. *JADT 2004 7es Journées Internationales d'Analyse Statistique Des Données Textuelles*, 1031–1038.

Tavabi, N., Stück, D., Signorini, A., Karjadi, C., Hanai, T. Al, Sandoval, M., Lemke, C., Glass, J., Hardy, S., Lavallee, M., Wasserman, B., Ang, T. F. A., Nowak, C. M., Kainkaryam, R., Foschini, L., & Au, R. (2022). Cognitive digital biomarkers from automated transcription of spoken language. *The Journal of Prevention of Alzheimer's Disease*, *9*, 791–800. https://doi.org/10.14283/jpad.2022.66

Tesnière, L. (2015). *Elements of Structural Syntax*. John Benjamins Publishing Company. https://doi.org/10.1075/z.185

Uddén, J., Hultén, A., Schoffelen, J. M., Lam, N., Harbusch, K., van den Bosch, A., Kempen, G., Petersson, K. M., & Hagoort, P. (2022). Supramodal sentence processing in the human brain: fMRI evidence for the influence of syntactic complexity in more than 200 participants. *Neurobiology of Language*, *3*(4), 575–598. https://doi.org/10.1162/nol_a_00076

Wingfield, A., Peelle, J. E., & Grossman, M. (2003). Speech rate and syntactic complexity as multiplicative factors in speech comprehension by young and older adults. *Aging, Neuropsychology and Cognition*, *10*(4), 310–322. https://doi.org/10.1076/ANEC.10.4.310.28974

Yngve, V. H. (1960). A Model and an Hypothesis for Language Structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466. https://www.jstor.org/stable/985230

Yoon, S.-Y., Lu, X., & Zechner, K. (2020). Features measuring vocabulary and grammar. In K. Zechner & K. Evanini (Eds.), *Automated Speaking Assessment: Using langauge technologies to score spontaneous speech* (pp. 123–137). Routledge.

Zechner, K., Yoon, S.-Y., Bhat, S., & Leong, C. W. (2017). Comparative evaluation of automated scoring of syntactic competence of non-native speakers. *Computers in Human Behavior*, *76*, 672–682. https://doi.org/10.1016/j.chb.2017.01.060

Zhu, Z., Hou, X., & Yang, Y. (2018). Reduced syntactic processing efficiency in older adults during sentence comprehension. *Frontiers in Psychology*, *9*(MAR), 243. https://doi.org/10.3389/FPSYG.2018.00243/BIBTEX

Ziegler, J., & Pylkkänen, L. (2016). Scalar adjectives and the temporal unfolding of semantic composition: An MEG investigation. *Neuropsychologia*, *89*, 161–171. https://doi.org/10.1016/j.neuropsychologia.2016.06.010

Zurif, E., Swinney, D., Prather, P., Solomon, J., & Bushell, C. (1993). An On-Line Analysis of Syntactic Processing in Broca′s and Wernicke′s Aphasia. *Brain and Language*, *45*(3), 448–464. https://doi.org/10.1006/BRLN.1993.1054

Zurif, E., Swinney, D., Prather, P., Wingfield, A., & Brownell, H. (1995). The allocation of memory resources during sentence comprehension: Evidence from the elderly. *Journal of Psycholinguistic Research 1995 24:3*, *24*(3), 165–182. https://doi.org/10.1007/BF02145354