Vocabulary, Grammar, Sex, and Aging

Fermín Moscoso del Prado Martín

University of California, Santa Barbara

Received 24 April 2015; received in revised form 18 November 2015; accepted 6 January 2016

Abstract

Understanding the changes in our language abilities along the lifespan is a crucial step for understanding the aging process both in normal and in abnormal circumstances. Besides controlled experimental tasks, it is equally crucial to investigate language in unconstrained conversation. I present an information-theoretical analysis of a corpus of dyadic conversations investigating how the richness of the vocabulary, the word-internal structure (inflectional morphology), and the syntax of the utterances evolves as a function of the speaker’s age and sex. Although vocabulary diversity increases throughout the lifetime, grammatical diversities follow a different pattern, which also differs between women and men. Women use increasingly diverse syntactic structures at least up to their late fifties, and they do not deteriorate in terms of fluency through their lifespan. However, from age 45 onward, men exhibit a decrease in the diversity of the syntactic structures they use, coupled with an increased number of speech disfluencies.

Keywords: Aging; Corpus study; Dialog; English; Information theory; Lexicon; Morphology; Sex differences; Syntax

1. Introduction

Language is perhaps the most unique cognitive ability of humans. Naturally, it is an ability that changes along people’s lifetimes. Beyond the evident changes in language during the early years of life (i.e., language acquisition), linguistic performance has been widely documented to be influenced by aging processes. It is long-known that the fundamental frequency of speech (F0) becomes lower with growing age (e.g., Endres, Bam-bach, & Flösser, 1971; Harrington, Palethorpe, & Watson, 2007), which is related to both cognitive and physiological aspects of aging (Ramig & Ringel, 1983). It has also been found that older people are slower and less accurate than younger people in recognizing and producing words (e.g., Lima, Hale, & Myerson, 1991; Mortensen, Meyer, & Hum-
phreys, 2006), which results in significantly reduced speech tempos for older people (e.g., Quéné, 2013). This slowing down is, however, counterbalanced by older people using richer vocabularies than younger people do (Hartshorne & Germine, 2015; Kavé, Knafo, & Gilboa, 2010; Kavé, Samuel-Enoch, & Adiv, 2009). This has led some researchers to argue that cognitive decline of linguistic abilities is a “myth” (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014), as the slowing down and decreased accuracies can be attributed to having to choose from a larger lexicon or to trying to access more detailed representations (Kavé & Nussbaum, 2012), rather than to decline in actual cognitive ability. Beyond single words, the syntactic complexity of the utterances produced by people is reported to decline in the later stages of life (e.g., Kemper, Thompson, & Maquis, 2001), and so does the ease with which people understand syntactically complex sentences (e.g., Waters & Caplan, 2001).

Most research on aging effects on language skills is based on well-controlled experimental studies in lab (and, more recently, also online) contexts or on highly edited samples such as novels, speeches, or broadcasts (e.g., Harrington et al., 2007; Le, Lancashire, Hirst, & Jokel, 2011; Quéné, 2013). However, the “ecological niche” of human language is neither picture naming or lexical decision experiments nor carefully prepared written language, but rather natural unedited conversations. Despite the evident value of experimental studies, it is also necessary to investigate the age evolution of language abilities in natural dialog situations. This is important because performance in natural dialog requires that the speakers successfully negotiate multiple social, pragmatic, and perceptual cues, imposing an additional set of constraints and conditionings on the cognitive system (e.g., Adams, Smith, Pasupathi, & Vitolo, 2002; Stine-Morrow, Soederberg Miller, & Hertzog, 2006).

Some researchers have investigated the effects of aging on linguistic performance in dialog situations. Bortfeld, Leon, Bloom, Schrober, and Brennan (2001) analyzed a corpus of conversations elicited in collaborative task situations. They report that older speakers produce more sentence-internal disfluencies than both middle-aged and young speakers do, with no difference found between the two latter groups. Closer to natural spontaneous dialog, Horton, Spieler, and Shriberg (2010) and Meylan and Gahl (2014) analyzed the Switchboard I Corpus (Godfrey, Holliman, & McDaniel, 1992), a large collection of transcribed telephone conversations between speakers of different ages and backgrounds, in which the caller chose the conversation topic from a predefined list. Horton and his colleagues found positive correlations between the age of the speakers and the lexical richness, number of filled pauses, and the length of the sentences they used (a proxy for syntactic complexity), as well as a negative correlation between the age and the rate at which people spoke. However, it was noticed that the degree of intercorrelation between all variables, and the likely influence of other properties of the speaker (i.e., sex, level of education, dialect, etc.), make plain correlations inadequate for assessing whether or not a factor is influenced by age. Working on the same corpus, Meylan and Gahl addressed the problem of possible confounds due to speaker properties by using linear mixed-effect model regression analyses instead of the plain correlations. Their results confirmed Horton et al.’s (2010) finding that the diversity of the lexicon increases with the age of the
speaker, and further added that the older speakers are less sensitive to lexical priming
than young speakers (i.e., older speakers are less likely to reuse words produced earlier
by their interlocutor). Finally, Gahl, Cibelli, Hall, and Sprouse (2014) analyzed a longitudi-
dinal corpus of spontaneous speech following 10 speakers from ages 7 to 49 years, con-
firming the relative slowing of speech rate with age previously found by Horton and his
collaborators.

A further problem in the studies of Horton et al. (2010) and Meylan and Gahl (2014)
concerns their measure of lexical diversity. They measured lexical diversity using the
Uber Index (Dugast, 1980), a variation of the traditional type-token ratio, but claimed to
be less dependent on sample size. Unfortunately this index is far from independent from
sample size (cf., Tweedie & Baayen, 1998; see also Appendix S1). Aware of this poten-
tial problem, Meylan and Gahl assumed that, as the sample sizes for each speaker are
more or less similar across the Switchboard I corpus, they should not expect any system-
atric effects of sample size to affect the results (cf., Meylan & Gahl, 2014, p. 1007). This
is, however, a problematic assumption. First, as I will show, there is considerable vari-
ability in the sample sizes contributed by different speakers in that corpus (i.e., ranging
from 94 to almost 3,000 words). Second, and most important, this variability is system-
atic: The sample size (i.e., length of the contributions to a conversation) of each speaker
is significantly related to his/her age. Any effects found are therefore suspect of being just
the consequence of the differences in sample size (i.e., speakers of certain ages might just
happen to talk more than those of others, but there is no real change in the properties of
their speech beyond its quantity). Notice that this problem is not exclusive of the Uber
Index: All measures of lexical diversity that can be estimated from a corpus suffer from
sample size bias (cf., Tweedie & Baayen, 1998). It is therefore crucial that the bias of
sample size is explicitly considered when evaluating the change in measures of lexical
diversity.

Moscoso del Prado Martín (2014) introduces a uniform information-theoretical frame-
work for measuring the lexical, morphological (inflectional), and syntactic diversity from
corpora. In this approach, all diversities are quantified using the same measure, the
entropy (Shannon, 1948) of their distribution. This approach to measuring diversity,
which is dominant in biology (cf. Gotelli & Chao, 2013), presents several advantages: (a)
uniformity, all aspects of diversity are measured using the same standard tool, rather than
ad hoc measures for each level; (b) interpretability, in contrast with the often obscure val-
ues and units obtained from traditional lexical diversity measures (e.g., Uber Index, Her-
dan Index, etc.), entropy provides easily interpretable values measured in well-understood
standard units (i.e., bits, nats, etc.); (c) finiteness, entropy offers valid finite values even
for distributions with a potentially infinite number of types (as would be the case for syn-
tactic structures according to many linguistic theories); (d) the raw measures exhibit faster
convergence and higher consistency than type-token ratio variants and germane measures
(see Appendix S1); and (e) there are effective, well-studied methods for correcting the
sample size bias.

It is often overlooked that the evolution of cognitive abilities along the adult lifespan
is both nonlinear and multi-faceted. In a recent study, Hartshorne and Germine (2015)
report that cognitive abilities do not necessarily follow monotonically increasing/decreasing patterns. Rather, they often exhibit nonlinear trends, where a certain ability improves in the early stages of adulthood, reaches a peak, and decreases thereafter. Crucially, Hartshorne and Germine found remarkable variability in the ages at which different abilities peak, ranging from the early teens for several types of short-term memory tasks, to the fifties (or perhaps later), for vocabulary tasks. In turn, language itself is far from being a monolithic system, but it involves a wide range of types of knowledge and skills. The studies on the evolution of linguistic performance in natural dialog have focused mostly on properties of the speech signal, on the vocabulary, and on the disfluencies (with syntax being indirectly considered as well by Horton et al., 2010). All studies assume monotonic (linear) trends of the age of the speakers on vocabulary size (Horton et al., 2010; Meylan & Gahl, 2014). This might, however, be inappropriate. Several authors have reported a possible decline of vocabulary size with very old age (e.g., Hartshorne & Germine, 2015; Kavé et al., 2010; Kemper et al., 2001), which may arise from non-monotonic—peaking—trends (Hartshorne & Germine, 2015; Kavé et al., 2010). Furthermore, beyond vocabulary, it is important to assess the evolution of higher levels of linguistic processing with age. As mentioned above, experimental evidence seems to indicate a decrease in the ability to produce and comprehend complex syntactic structures (e.g., Kemper et al., 2001; Waters & Caplan, 2001). Higher levels of linguistic structure might be subject to a different set of cognitive constraints than those affecting the lower levels.

Numerous studies have documented that aging affects men and women differently. Both sexes age differently with respect to a wide range of biological markers (e.g., Nakamura & Miyao, 2008). Anatomical changes on the brain are reported to have different evolutions with age with respect to sex (e.g., Cowell et al., 1994; Gur & Gur, 2002). In turn, these physiological differences result in differences in the aging pattern of behavior and cognitive performance (e.g., Costa, Santos, Cunha, Palha, & Sousa, 2013; Gur & Gur, 2002). In particular, many studies report that men are affected earlier and more pronouncedly than women by both anatomical changes (e.g., Cowell et al., 1994) and decreases in cognitive performance across a wide set of domains (e.g., Costa et al., 2013; Gur & Gur, 2002). These sex-differentiated aging patterns raise the question of whether (and how) men differ from women in terms of how aging affects their linguistic abilities. Sex differences concerning language have been widely reported at anatomical, physiological, and behavioral levels. Cerebral areas of importance for language processing have been found to be of different sizes and neural densities between males and females (Harasty, Double, Halliday, Krill, & McRitchie, 1997; Sowell et al., 2003; Witeelson, Glezer, & Kilgar, 1995). Likewise, men and women differ in the brain areas they engage in language processing (see, e.g., Baxter et al., 2003, and references therein). There are significant differences in the linguistic behavior of boys and girls during language acquisition (Hartshorne & Ullman, 2006), and these differences extend to their behavior later in life (Ullman, Miranda, & Travers, 2007). Interestingly, it appears that these behavioral differences can be modulated by hormonal factors (Estabrooke, Mordecai, Maki, & Ullman, 2002; Ullman et al., 2002), and these are known to change with age. The aging literature on this question has focused almost exclusively on investigating whether men and women differ in terms of their lexical...
knowledge, whose changes with age do not appear to depend on the sex of the speaker (e.g., Gur & Gur, 2002). It remains, however, unclear whether sex differences could be observed at higher levels of linguistic structure, such as morphology and syntax. From the corpus analysis perspective, few studies have investigated the joint effects of sex an aging on linguistic performance. Analyzing the Switchboard I corpus, Shriberg (1996) noticed that men produced certain types of disfluencies more often than women did, but she found it difficult to disentangle these differences from socioeconomic properties of the speakers. On their part, in their analysis of task-oriented speech, Bortfeld et al. (2001) studied the effects of both the age and the sex on the speakers on the number (and type) of disfluencies they produced. Once more, they found that men produced more disfluencies than women did, and this effect could not be attributed to socioeconomic variables. Furthermore, the role that the speaker played in the joint activity (i.e., leading “director” vs. more subordinate “matcher”) increases the number of disfluencies produced by men, but not by women. Unfortunately, however, Bortfeld and her colleagues did not analyze whether the age-related patterns found in the disfluencies differed also in terms of sex.

In this study, I investigate how the age and sex of speakers affect the complexity of the lexicon (i.e., vocabulary), inflectional morphology (i.e., grammatical processes that change the form of words to fit them into particular contexts; e.g., “eat”—“eats”—“ate”—“eating”—“eaten,” “car”—“cars”), syntax (i.e., grammatical processes governing the ordering and grouping of words), and disfluencies (i.e., rephrasings, filled pauses, etc.) produced by speakers in telephone conversations. As several previous studies (Horton et al., 2010; Meylan & Gahl, 2014; Shriberg, 1996), I use the conversations in the Switchboard I Corpus combined with syntactic parses of those conversations (from the Penn Treebank; Marcus, Santorini, Marcinkiewicz, & Taylor, 1999). I follow Moscoso del Prado Martín (2014) in using entropy for characterizing all diversity measures. I improve on previous studies by applying sample size bias correction techniques when these are available, and in all cases explicitly discounting the confounds that might be caused by sample size differences. Similar to what Meylan and Gahl (2014) did, and in contrast with Horton et al. (2010), I use mixed-effects regression models to account for possible properties of the speaker. Crucially, unlike all previous dialog studies, the regressions include nonlinear terms, hence allowing to model any potential non-monotonicities in the relationships between the measures and the speaker ages. In addition, I also investigate how the sex of the speakers interacts with the evolution of their linguistic performance with age. Finally, the results are discussed in relation to previously reported behavioral and neurophysiological studies on the influence of sex and aging on language abilities.

2. Method

2.1. Materials

I used the Switchboard I Corpus (Godfrey et al., 1992), a collection of telephone conversations between previously unacquainted native speakers of American English of
diverse ages and backgrounds, triggered by a conversational prompt (i.e., an initial conversation topic was chosen by the speaker making the call, but both speakers were free to change topics during the conversation). I crossed the conversations with the syntactic parse trees provided in the Penn Treebank (Marcus et al., 1999) for a subset of the dialogs. This resulted in 650 conversations for which syntactic parse trees were available in the Treebank. In total, the subcorpus contained 1,023,832 words (excluding punctuation, digits, and non-alphabetic characters), that is, an average of 788 words per participant in each conversation. These words were grouped into 120,414 parse trees, corresponding to an average of 93 parse trees per participant in a conversation. In total, there were 359 distinct speakers (165 women and 194 men, all born between 1924 and 1972), some of which took part in more than one conversation (ranging from a single conversation for more than 25% of the speakers, to two speakers who took part in 12 conversations each; the median speaker took part in three conversations). As is shown in Fig. 1, the ages of the speakers were similarly distributed for men and women (i.e., the age distributions were not significantly different according to a two-sample Kolmogorov–Smirnov test: \( D = .100, p > .250 \)).

Each time one of the 359 speakers took part in a conversation, I attached to that conversation’s record his/her age in years (computed as the difference in days between the birth date\(^1\) and the date of the recording, divided by 365), sex (647 instances of women and 653 instances of men), level of education (“less than high school”: 14 cases, “less than college”: 58, “college”: 798, “more than college”: 417, and “unknown”: 13), the conversational topic chosen (with 64 different values), and the American English dialect area where the speaker resided (“New England”: 55 cases, “North Midland”: 165, “Northern”: 190, “New York City”: 76, “South Midland”: 427, “Southern”: 127, “Western”: 176, “mixed”: 83, and “unknown”: 1 case) as provided by the Switchboard I Corpus.

![Fig. 1. Distribution of distinct speakers by their sex and their age in 1991 (when Switchboard data collection began).](image-url)
2.2. Corpus processing and measurements

The words in each conversation were lemmatized (e.g., “eat,” “eats,” “ate,” “eating,” and “eaten” were all considered to be instances of the lemma EAT, and both “car” and “cars” were taken as instances of the lemma CAR) using the WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) automatic lemmatizer. The frequency distribution of the lemmas was used for computing the lexical diversity \( H[H[L]; Moscoso del Prado Martín, 2014], that is, the entropy (Shannon, 1948) of the frequency distribution of lemmas. One could compute the entropy using Shannon’s original expression,

\[
H[L] = - \sum_{\ell \in L} p(\ell) \log p(\ell)
\]  

where the \( p(\ell) \) correspond to the relative probabilities of the lemmas used by the speaker. However, using corpus counts directly into this expression—the maximum-likelihood estimator—results in substantial underestimation of the entropy (i.e., the estimator is biased; Miller, 1955). In order to attenuate this problem, the entropies were computed from the frequencies using the optimal reduced bias entropy estimator described by Chao, Wang, and Jost (2013).

The entropy of the frequency distribution of unlemmatized word—lemma pairs \( H[W, L] \) was also computed for each participant in each conversation using the method of Chao et al. (2013). The difference between the two entropies \( H[W, L] \) and \( H[L] \) is the inflectional diversity (Moscoso del Prado Martín, 2014),

\[
\]  

Inflectional diversity corresponds to the average inflectional entropy (Moscoso del Prado Martín, Kostić, & Baayen, 2004) of the lemmas used by each participant. The latter is a measure of the diversity of inflected variants for each lemma in the corpus. This measure has been shown to capture the cost of recognizing and acquiring different words (Baayen & Moscoso del Prado Martín, 2005; Moscoso del Prado Martín et al., 2004; Stoll et al., 2012). Therefore, as the average inflectional entropy, inflectional diversity captures the complexity of the morphological system of a person from an information-processing perspective. The values of this measure are an index of how many distinct inflected variants are used for the average word.

In order to measure the syntactic complexity, for each participant in each conversation, I extracted from the Penn Treebank (Marcus et al., 1999) the syntactic parse trees corresponding to all of the utterances produced by that participant. The parse trees were cleaned to remove all disfluencies marked in the tree (i.e., false starts, hesitations, “huh,” pauses, etc.). Punctuation nodes, and the tree leaves (i.e., the words themselves), were removed, so that the leaves of the new tree would be the part-of-speech tags (e.g., “singular noun,” “adverb,” “adjective,” . . .). Finally, to ensure that all trees had the same root node, a new node \( SS_0 \) was added to each tree directly dominating its root. This process
resulted, for each conversation participant, in a collection of parse trees like the one in Fig. 2a. From those trees, I extracted the phrase-structure production rules (see Fig. 2b). Using these productions and their frequencies of usage, for each speaker I induced a probabilistic context-free grammar (PCFG; Booth & Thompson, 1973) by maximum-likelihood estimation (i.e., using the raw frequencies of occurrence of the rules in each conversation participant’s sample). Finally, from the induced PCFG, I computed the entropy of the parse trees it generates (Chi, 1999; Grenander, 1976), the syntactic diversity (Moscoso del Prado Martín, 2014). This measures how many distinct parse trees (taking their probabilities into account) could be generated using the grammar rules provided, and it has been shown to be a relevant measure of processing difficulty (Hale, 2006). Given the small samples, this entropy is obviously an underestimate of the entropy of the trees that the speaker could have hypothetically produced (see Appendix S1). This is not, however, a problem, as I explicitly include the sample size as an independent predictor in the regression models. Therefore, any differences in entropy estimates due to sample size alone are accounted for.

In order to obtain controls for the biases that result from the entropy estimation procedures above (see Appendix S1), for each participant in each conversation, I recorded the mean length in words of the clauses (i.e., parse trees) he/she produced (after removing disfluencies) and the total summed length in words of all the utterances he or she produced in the conversation. Finally, for each speaker in each conversation I also recorded the average number of disfluencies per clause that were labeled in the corpus. A summary of the corpus measures is provided in Table 1. Fig. 3 summarizes the marginal distributions, correlations, and nonlinear relations between each of the numerical variables considered.

---

(a) (b)

Fig. 2. (a) Example of a syntactic parse tree (with disfluencies removed). The nodes in italic font are the terminals that are removed prior to rule extraction. (b) Phrase-structure rules extracted from the tree in (a).
2.3. Analyses

For each of the diversity measures (lexical, inflectional, and syntactic), and for the mean number of disfluencies per clause, I fitted a generalized additive mixed-effects model (GAMM; cf., Wood, 2006). All four models included random effects of conversation identity (with 650 possible values), topic of conversation (with 64 possible values), and dialect area of the speaker (“New England,” “North Midland,” “Northern,” “New York City,” “South Midland,” “Southern,” “Western,” “mixed,” or “unknown”). In each model, the random effect terms were deemed necessary using Wald tests on maximum-likelihood model fits with different random effect structures. After fitting each model, the model residuals were inspected. Those models whose residuals diverged substantially from normality were refit after transforming the dependent variable using its logarithm (giving rise to log-normal regressions), the resulting log-normal model residuals were again inspected for deviations from normality, but none was found. All regression models included fixed effects of speaker’s role in the conversation (caller vs. callee), speaker sex (woman vs. man), listener sex (woman vs. man), the interaction between both sexes, and level of education (“less than high school” vs. “less than college” vs. “college” vs. “more than college” vs. “unknown”). The fixed effects that did not reach significance were removed from the model fits.

On the one hand, in the models fitting the diversity measures, it is necessary to consider the total sample size (i.e., the summed lengths in words of the speaker’s utterances in the conversation), as this will be the main factor determining the negative bias of the entropy estimators (i.e., entropy estimates generally increase with increasing sample size; Miller, 1955). To account for these biases, those three models included a nonlinear effect of the total length of utterances (modeled using a thin plate regression spline with automatically determined dimension; cf., Wood, 2006). On the other hand, the number of disfluencies per clause does not depend on the total length of the utterances, but on the length of the clauses themselves: Longer clauses offer more opportunities for disfluencies to arise and, for this reason, disfluency counts are known to be linearly related to clause

Table 1

<table>
<thead>
<tr>
<th>Measure</th>
<th>Unit</th>
<th>Minimum</th>
<th>1st Quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>years</td>
<td>19.71</td>
<td>28.75</td>
<td>34.73</td>
<td>37.17</td>
<td>46.42</td>
<td>67.59</td>
</tr>
<tr>
<td>Age difference</td>
<td>years</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>11.79</td>
<td>18</td>
<td>40</td>
</tr>
<tr>
<td>Total length</td>
<td>words</td>
<td>94</td>
<td>513</td>
<td>696.5</td>
<td>787.6</td>
<td>980</td>
<td>2973</td>
</tr>
<tr>
<td>Clause length</td>
<td>words/clause</td>
<td>2.31</td>
<td>6.36</td>
<td>7.81</td>
<td>8.32</td>
<td>9.77</td>
<td>26.79</td>
</tr>
<tr>
<td>Lexical div.</td>
<td>nats/word</td>
<td>3.88</td>
<td>4.86</td>
<td>4.98</td>
<td>4.97</td>
<td>5.09</td>
<td>5.58</td>
</tr>
<tr>
<td>Inflectional div.</td>
<td>nats/word</td>
<td>.0199</td>
<td>.1112</td>
<td>.1358</td>
<td>.1358</td>
<td>.1627</td>
<td>.2620</td>
</tr>
<tr>
<td>Syntactic div.</td>
<td>nats/clause</td>
<td>3.90</td>
<td>14.40</td>
<td>17.88</td>
<td>18.91</td>
<td>22.55</td>
<td>62.68</td>
</tr>
<tr>
<td>Disfluencies</td>
<td>disfl./clause</td>
<td>1.46</td>
<td>2.33</td>
<td>2.70</td>
<td>2.92</td>
<td>3.32</td>
<td>12.58</td>
</tr>
</tbody>
</table>

*Note.* nats are \(\log_e\) information units, the same way that bits are \(\log_2\) based information units; for example, \(1 \text{ nat} = 1/\log(2) \approx 1.4427 \text{ bits.}\)
lengths (Oviatt, 1995; Shriberg, 1996). This was accounted for by including a nonlinear effect (also a thin plate spline) of mean clause length in the regression fitting the number of disfluencies.

To investigate the evolution of the measures with age, a nonlinear effect (thin plate spline) of age (in years) was included in the four regression models. In those models where sex was found to have a significant contribution (significantly higher or lower values for women than for men), I fitted an additional model with the same fixed and random effect structure, including a nonlinear interaction considering different effects of age
for women and men (instead of the nonlinear effect of age). Both models were compared using Wald log-likelihood tests, and only the better model was kept.

Finally, it is necessary to take into account that—on average—women and men tend to have different preferences on the topics about which they like to talk. Fig. 4 compares, for the 64 topics, the number of times each topic was chosen by a female or a male caller. It shows how the likelihood of a topic being chosen significantly depends on the sex of the caller ($\chi^2(63) = 127.6, p < .001$). If the conversation is about a topic one has not—or would have rather not—chosen, this might lead a speaker to have less to contribute to the conversation, and therefore use a poorer vocabulary or a less complex syntax. To account for this possibility, I included additional mixed-effects (i.e., random slopes) of speaker sex by topic and speaker role (caller/callee) by topic. According to Wald tests on maximum regressions, and including them anyway did not change the pattern of results. For these reasons, in what follows, I do not discuss these mixed effects any further.

3. Results

Table 2 summarizes the results of the GAMM regressions. Before moving into the specific effects found in each of the models, as a “sanity check,” it is worth examining whether the nonlinear terms succeeded in reconstructing the shape of the correction terms included in the models, that is, the sample size (i.e., fragment length) measures used for correcting the bias of the entropy estimators, and the clause length term included to account for the fact that longer clauses afford more disfluencies per clause. These are plotted in Fig. 5. It was noticed that the estimates become too noisy beyond sample sizes of 1,500 words or mean clause lengths of more than 15 words, as there are few data points with these characteristics. Panel (a) plots the effect of the sample size on the estimated lexical diversity. The monotonically increasing concave shape plotted in the graph, is the exact shape one should expect for the convergence of the Chao–Wang–Jost entropy.
Table 2
Effect significance in the four GAMM models on the untransformed dependent variables

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Lexical Diversity a</th>
<th>Inflectional Diversity a</th>
<th>Syntactic Diversity b</th>
<th>Number of Disfluencies b</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speaker’s role</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaker’s sex</td>
<td>$F (1, 1285.96) = 10.51$</td>
<td>$F &lt; 1$</td>
<td>$F (1, 1279.233) = 2.58$</td>
<td>$F (1, 1283.89) = 2.10$</td>
</tr>
<tr>
<td></td>
<td>$p &lt; .001$</td>
<td></td>
<td>$p = .108$</td>
<td>$p = .147$</td>
</tr>
<tr>
<td><strong>Listener’s sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listener’s sex</td>
<td>$F (1, 1282.88) = 1.55$</td>
<td>$F &lt; 1$</td>
<td>$F (1, 1279.84) = 44.93$</td>
<td>$F (1, 1285.04) = 45.31$</td>
</tr>
<tr>
<td></td>
<td>$p = .214$</td>
<td></td>
<td>$p &lt; .001$</td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td><strong>Sex interaction</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex interaction</td>
<td>$F (2, 1282.88) = 1.68$</td>
<td>$F (1, 1285.01) = 1.23$</td>
<td>$F (1, 1279.84) = 16.03$</td>
<td>$F &lt; 1$</td>
</tr>
<tr>
<td></td>
<td>$p = .195$</td>
<td></td>
<td>$p &lt; .001$</td>
<td></td>
</tr>
<tr>
<td><strong>Level of education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level of education</td>
<td>$F (4, 1285.96) = 6.14$</td>
<td>$F (1, 1289.09) = 3.26$</td>
<td>$F (1, 1279.84) = 3.08$</td>
<td>$F (4, 1285.04) = 10.34$</td>
</tr>
<tr>
<td></td>
<td>$p &lt; .001$</td>
<td></td>
<td>$p = .015$</td>
<td>$p &lt; .001$</td>
</tr>
<tr>
<td><strong>Length of utterances</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length of utterances</td>
<td>$F (7.04, 1285.96) = 42.27$</td>
<td>$F (2.27, 1289.09) = 71.94$</td>
<td>$F (6.92, 1279.84) = 224.34$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p &lt; .001$</td>
<td></td>
<td>$p &lt; .001$</td>
<td></td>
</tr>
<tr>
<td><strong>Mean length of clauses</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean length of clauses</td>
<td>–</td>
<td></td>
<td>–</td>
<td>$F (1.91, 1285.04) = 162.27$</td>
</tr>
<tr>
<td>Age</td>
<td>$F (1, 1285.96) = 7.96$</td>
<td>$F (3.64, 1289.09) = 3.82$</td>
<td>$F (2.42, 632.84) = 15.51$</td>
<td>$F (1.00, 638.04) = 5.97$</td>
</tr>
<tr>
<td></td>
<td>$p &lt; .001$</td>
<td></td>
<td>$p &lt; .001$</td>
<td>$p = .015$</td>
</tr>
<tr>
<td><strong>Age × Sex</strong></td>
<td>$\chi^2(2) = 1.557$</td>
<td>$\chi^2(2) = .018$</td>
<td>$\chi^2(2) = 5.970$</td>
<td>$\chi^2(2) = 24.869$</td>
</tr>
<tr>
<td>Interaction</td>
<td>$p &gt; .250$</td>
<td>$p &gt; .250$</td>
<td>$p = .050$</td>
<td>$p &lt; .001$</td>
</tr>
</tbody>
</table>

**Note.** The $p$-values and degrees of freedom of the $F$-tests are approximations.

*aUsing normal regression model.

*bUsing log-normal regression model.
estimator that were used (see Appendix S1). In contrast the inflectional diversity—plotted in panel (b)—shows a quasilinear, slightly concave increase. The inflectional diversity is the difference between two entropy estimates (see Eq. 2), the first of which is expected to be only slightly larger than the second. In such small magnitude of difference, the convergence is necessarily slow, hence the almost linear—but still concave—pattern. Panel (c), plotting the convergence of the grammatical diversity, also exhibits a concave convergence pattern. However, as these are fully uncorrected maximum-likelihood estimates (I do not know any method for correcting the bias of PCFG entropy estimates), their convergence should be expected to be extremely slow (see Appendix S1). Finally,
the number of disfluencies per clause is expected to be directly proportional to the average clause length (Oviatt, 1995; Shriberg, 1996), hence the linear pattern in panel (d).

The model fitting the lexical diversities did not reveal any effect of sex, either of the speaker or of the listener (or their interaction). However, it did show a significant effect of the speaker’s role in the conversation, indicating that the speaker making the call (the caller) and choosing the topic overall used a richer vocabulary than the speaker receiving the call (the callee). A significant main effect for the speaker’s level of education was also present, indicating that the lexical diversity was lowest for people with education below high school, slightly higher for education below college, higher for people educated at college level or more, and highest for the cases whose educational level was unknown (13 datapoints). After discounting the nonlinear effect of the length of utterances, there was a significant effect of the age of the participants. As plotted in Fig. 6a, this effect indicated that lexical diversity (i.e., vocabulary) of the utterances produced increases linearly with the speaker’s age. In other words, speakers enrich their vocabularies at a steady rate throughout their lives, with no suggestion of decline up to advanced ages. This pattern was not significantly different between men and women.

The model fitting inflectional diversities revealed no effects of the speaker’s role, or of the sex of neither speaker nor listener. It found, once again, an effect of the speaker’s level of education (i.e., the 13 speakers whose educational level was unknown exhibited richer inflection). After discounting the nonlinear effect of the length of the utterances, there was a nonlinear effect of age of the participants (which was not found to differ by sex). As shown in Fig. 6b, inflectional diversity evolves non-monotonically with speaker age.

Fig. 6. (a) Effect of speaker’s age on the lexical diversities. (b) Effect of the speaker’s age on the inflectional diversities. Note. The unreliable estimates for ages over 63 have been clipped from the graphs.
age irrespective of his/her sex, peaking at around 45 years of age, and decreasing thereafter.

The GAMM fit to the syntactic diversities revealed main effects of the speaker’s sex (i.e., overall men used a more diverse syntax than women did), the sex of the listener (i.e., speakers made use of a more diverse syntax when talking to a man than when talking to a woman), and level of education (i.e., speakers with education below high school used less varied syntactic constructions than those speakers who had education of high school or above, and the 13 speakers whose educational level was unknown exhibited the richest syntax). There was a nonlinear effect of the length of the utterances as before. Interestingly, the effect of age on the syntactic diversities was different for women and men and significant for both (the presence of the interaction was marginally significant; \( p = .050 \)).\(^4\) These effects are plotted in Fig. 7. On the one hand, the syntactic diversity of utterances produced by women (left-hand side panel) increases throughout their lives, with a slight attenuation in the late fifties. On the other hand, the syntactic diversity of utterances produced by men (right-hand side panel), although overall richer than that of women, peaks at around 45 years of age and clearly decreases thereafter. In addition, for men, there seems to be an acceleration of the increase in syntactic complexity starting around the mid thirties.

Finally, the fit to the mean number of disfluencies per clause revealed main effects for the speaker’s sex (overall, women produced less disfluencies per sentence than men did), and a significant interaction between the sex of the speaker and the sex of the listener (besides producing on average more disfluencies than women, men’s disfluencies were further increased when talking to other men, whereas for women the sex of the listener

---

**Fig. 7.** Effect of the speaker’s age on the syntactic diversities of female speakers (left panel) and male speakers (right panel). Note. The plots have been back-transformed from the logarithmic scale in which the regressions were fitted; the unreliable estimates for ages over 63 have been clipped from the graphs.
did not significantly affect their average number of disfluencies). As in the previous three models, there was also an effect of the level of education (the speakers with education below high school produced more disfluencies than those speakers who had education of high school or above and, as in the previous models, the 13 speakers with unknown educational level produced the lowest number of disfluencies per clause). The mean clause length also exhibited a significant nonlinear effect. The effect of age on the disfluencies was clearly different for women and men, and was significant for both. These effects are plotted in Fig. 8. As they age, women steadily produce less disfluencies, following a linear trend. In contrast, men appear to follow approximately the same pattern as women until they reach the age of 45, from where the number of disfluencies they produce markedly increases. The clause length measure employed has a very strong linear relationship with the syntactic diversity measure considered above (i.e., longer clauses require more syntax; Pearson’s $r = .98$, $t(1298) = 165.89$, $p < .001$). Therefore, by partialing out the effect of clause length, I have implicitly partialled out the syntactic diversity as well. In other words, the effect of aging on the number of disfluencies cannot be attributed to differences in syntactic complexity.⁵

As evidenced by the plots and correlations in Fig. 3, the lexical, inflectional, and syntactic diversity and disfluency measures studied above are far from independent from each other. It would be therefore desirable to investigate to what degree do the results obtained reflect genuinely independent components of the evolution of linguistic abilities along the lifespan. This would normally be achieved by including multiple diversity and disfluency measures in the same regression models. The strong relationships between the variables, compounded with the need to include clause length, and sample size predictors

![Fig. 8. Effect of the speaker’s age on the average number of disfluencies per clause produced by female speakers (left panel) and male speakers (right panel). Note. The plots have been back-transformed from the logarithmic scale in which the regressions were fitted; the unreliable estimates for ages over 63 have been clipped from the graphs.](image-url)
into the models, would lead to extremely high multicollinearity, rendering the resulting analyses virtually useless.

An alternative route to address the problem above is to consider whether those measures can be decomposed into a set of predictors that are uncorrelated to each other. One assumes that the measured variables are the result of a linear mixing between multiple originally independent source signals. In the case of linear relations between variables, this is typically achieved using principal component analysis (PCA). However, the relations between our variables of interest are often nonlinear. It is then more adequate to use independent component analysis (ICA; cf., Hyvarinen & Oja, 2000), which estimates a set of original source variables (i.e., “independent components”) that are not only uncorrelated, but also independent in a nonlinear information-theoretical sense. In order to evaluate how the age effects represent different aspects of linguistic abilities, I performed an ICA decomposition on the original dependent variables, and repeated the GAMM regressions above, using the independent components as the dependent variables, following the same methodology, fixed effect, and random effect structure that was used in the analysis of the original variables. The results of those analyses—fully reported in Appendix S2—confirm that the pattern of results reported here are not a side effect of the multi-collinearity between the four measures used: (a) there is a nonlinear pattern on the diversity of inflectional and grammatical constructions used by speakers of different ages, (b) how aging affects speakers is dependent on the sex, with men showing an earlier decay than women do, with an onset at around 45 years of age, and (c) men’s disfluencies appear to increase from age 45, whereas women’s do not.

4. Discussion

This study demonstrates that age-related changes in the linguistic structures produced by speakers in natural conversations are heterogeneous; lexical diversities improve throughout speakers’ lives, while grammatical (i.e., inflectional & syntactic) diversities and disfluencies exhibit nonlinear patterns. Furthermore, the aging patterns in language are differentiated with respect to the sex of the speakers. Whereas women’s performance steadily increases until ages beyond 60 (with perhaps some decrease in their use of inflectional morphology), men exhibit a clear decrease in the richness of the grammatical structures they produce from the age of 45. At this age, the complexity of the syntax and inflectional morphology of the utterances they produce begins to recess. After age 45, the reduction in grammatical complexity is accompanied by a sudden marked increase in the number of disfluencies produced by male speakers, but not by female speakers.

Importantly, in contrast with previous corpus studies on aging (Bortfeld et al., 2001; Horton et al., 2010; Meylan & Gahl, 2014; Shriberg, 1996), the use of generalized additive mixed-effect models including nonlinear terms has enabled the investigation of the peaking patterns exhibited by the different measures with age, while simultaneously taking into account multiple properties of the speakers and the conversations that could give rise to confounds. Furthermore, independent component analysis (see Appendix S2) was
used to argue that the patterns reported for different measures should not be attributed to effects of a single aspect generating what appear to be multiple patterns by spreading intercorrelations, a weakness shared by the previous studies. This is especially important as it underlines the polyhedral nature of human language: Aging affects different aspects of language in different ways, similarly to what has been observed for cognitive abilities in general (e.g., Hartshorne & Germine, 2015). This also addresses the limitation expressed by Horton et al. (2010, p. 713) that this is a “found” dataset—rather than one elicited under controlled experimental conditions—and is therefore subject to possible confounds arising from the properties of the speakers. This is reminiscent of the ever-present tension in biology between the complementary fields of ethology and behavioral experimentation, appearing in linguistics under the names of corpus linguistics and psycholinguistics. The importance of well-designed, controlled experiments is beyond doubt, but this needs to be complemented with observational data of linguistic behavior in natural contexts. As is the case in ethology, natural dialogs are often subject to additional constraints generally difficult to recreate in the laboratory (e.g., Adams et al., 2002; Stine-Morrow et al., 2006). In this respect, I think that Horton and his colleagues might have underestimated the possibilities of modern statistical modeling techniques for addressing the possible confounds that arise in observational data.

A related technical aspect advanced by this study is the demonstration that one can draw meaningful inferences from samples that are—in the scale of corpora—extremely small. I combined the information-theoretical framework developed in Moscoso del Prado Martín (2014) for studying diachronic aspects of language, with appropriate non-parametric corrections for the strong biases that arise in such small sample sizes. As I discussed, sample size effects were a problem in all previous corpus studies of the evolution of lexical and syntactic complexity with age (Horton et al., 2010; Meylan & Gahl, 2014). The results reported demonstrate that one can obtain reliable comparisons about the diversities represented by samples as small as the contributions of a single speaker to a short telephone conversation. As it is evidenced by the analyses in the text (and the simulations detailed in Appendix S1), such technique is able to recover even very succinct patterns that are initially swamped under much confounding noise and is able to discard them when they are just spurious by-effects of other nonlinear relations. I believe that this is a useful contribution to the study of language using corpora: For many populations and languages, obtaining sufficiently large corpora is often simply beyond reach.

The linear increase in vocabulary richness throughout the lifespan is consistent with previous research on natural conversations (Horton et al., 2010; Meylan & Gahl, 2014), as well as with the experimental literature (see Verhaeghen, 2003, for an extensive meta-analysis, and Hartshorne & Germine, 2015, for a recent view). This result also supports the argument that vocabulary learning is generally spared in aging, continuing up to an advanced age, and that the observed slowing down of older people in vocabulary tasks is probably a by-effect of their having to access a larger and more detailed lexicon (e.g., Kavé & Nussbaum, 2012; Ramscar et al., 2014). In this respect, my result should, however, be taken with some care. The sample analyzed lacks any data beyond the age of 67 years, and in fact only six conversations were included at this age, and just a single
woman and a single man (both aged 67) in the pool of speakers were older than this. This resulted in confidence intervals beyond age 63 too large to draw any meaningful inference. It remains therefore possible that a peak in lexical diversity might be reached much later in life (cf. Singer, Verhaeghen, Ghisletta, Lindenberger, & Baltes, 2003).

The results for the grammatical components of language (inflectional morphology and syntax) were rather different. Instead of the linear increase that was observed for the vocabulary, both of these components exhibited significantly non-monotonic patterns. The diversity of inflectional forms shows an increasing pattern up to the age of 45 and decreases thereafter for both sexes. The diversity of syntactic structures used by speakers shows a very similar pattern, peaking at 45 years of age for men and not showing clear evidence of decline for women at least into their early sixties. In contrast with the ever-increasing vocabulary, from 45 years of age, men use less and less diverse grammatical constructions. It is as if the language they produced were becoming more and more “ossified” with age, making use of a more limited and predictable set of constructions. The progressive decrease in the syntactic complexity of the utterances produced by men from age 45 onward is consistent with the behavioral literature, which indicates that there is a decrease in the performance of older speakers in producing (e.g., Kemper et al., 2001) and comprehending (e.g., Waters & Caplan, 2001) syntactically complex sentences (see Burke & Shafto, 2008, for a detailed review). In comprehension, Antonenko et al. (2013) report that a decrease in syntactic ability—as reflected in decreased ability to understand sentences with increasing numbers of syntactic embeddings—in older age is paired with reduced functional connectivity within “dedicated syntax networks” in the brain. Finally, neural atrophy in older age (i.e., loss of both gray and white matter) is well documented, and this neural deterioration requires older speakers to recruit additional brain resources for syntactic and semantic processing (cf., Tyler et al., 2010, and references therein). Importantly, the changes in white matter volume are reported to be nonlinear, increasing from ages 19 to 40, and decreasing thereafter (Sowell et al., 2003). In short, the decreased syntactic complexity of the utterances produced by older men is fully in line with what is reported from the behavioral and neurophysiological literature: Older persons have more difficulties in processing syntax, and this is due to both anatomical and functional differences in their brains, contradicting Ramscar et al. (2014)’s statement that cognitive decline in linguistic abilities is a “myth.” Interestingly, I find that only men manifest the properties of a decreasingly diverse syntax on their speech. To my knowledge, this has been reported neither in the behavioral nor in the neurophysiological literature. However, none of the studies that I have found on this topic analyzed whether the patterns of change differed by sex, and they might have therefore overlooked it.

The question arises as to what degree do the results on syntax depend on the use of the specific grammatical formalism from the Penn Treebank, to which I propose no theoretical commitment. Indeed, the specific values of the syntactic entropy for each participant in a conversation will—to some degree—depend on the grammatical theory used for constructing the parses. However, one should expect the relative values of those entropies (i.e., the shape of their relationship with age and other variables) to remain more or less unchanged. Different syntactic structures should, in the majority of cases, receive
different parses within the same grammatical formalism (even if the specific parses differ across formalisms), and it is precisely such variability that is measured by the entropy values (Chi, 1999; Grenander, 1976). In fact, the current data already demonstrate this point. The average clause lengths were almost perfectly correlated (Pearson’s $r = .98$) with the syntactic complexity measure (see Fig. 3), and one can replicate the analysis on syntactic diversity replacing it with the mean clause length and obtain the very same results. Crucially, average clause lengths are basically identical to the proxy for syntactic complexity—but remain completely independent of any grammatical paradigm—that is most often used in the field of language acquisition and often also in clinical studies, the mean length of utterance (MLU), dating as far back as Nice (1925). Although the current standard practice—following Brown (1973)—is to measure MLUs in morphemes rather than in words, this does not really make any difference (Parker & Brorson, 2005). In this respect, the current results indicate that MLUs are indeed reliable measures of the average syntactic complexity of the utterances of an individual: They correlate almost perfectly with the productivity of the grammar the individual is using. This finding further validates the use of MLU-like measures in studies using corpora for which syntactic parses are not available (e.g., mean clause length was one of the proxies for syntactic complexity used in the study by Horton et al., 2010).

The results regarding the disfluencies are also remarkable. There is disagreement in the literature as to whether aging affects the amount of disfluencies produced by speakers. Multiple authors (Duchin & Mysak, 1987; Juste & Furquim de Andrade, 2011; de Oliveira Martins & Furquim de Andrade, 2008; Shewan & Henderson, 1988) have failed to find any difference on the number of disfluencies produced by younger and older subjects, not even for centenarians (Caruso, McClowry, & Max, 1997; Searl, Gabel, & Fulks, 2002). In contrast, others report higher disfluency rates among older speakers (Bortfeld et al., 2001; Horton et al., 2010; Schow, Christensen, Hutchinson, & Nerbonne, 1978). It was noticed that, for women, there is actually little difference in the number of disfluencies produced by younger and older speakers (and none if one takes the result of the ICA analyses into account), but, for men, there is a marked increase in the production of disfluencies from age 45 onwards. Crucially, the results found for men mimic those reported by Bortfeld et al. (2001); older speakers produce more disfluencies than middle-aged ones, which are themselves in this respect no different from the younger speakers. With respect to sex, Furquim de Andrade and Martins (2011) failed to find sex differences on the number of disfluencies, whereas such differences are reported by other studies (Bortfeld et al., 2001; Shriberg, 1996). My results suggest that the disagreements in the literature stem from failing to jointly considered sex and age as interacting variables.

One possibility is that the changes in the use of complex grammar in older ages are, per se, not an indication of cognitive decline, but rather reflect a change in speaking styles as one matures or some sociolectal differences across generations. This explanation would need to account for both the marked increase in the number of disfluencies produced by men above the age of 45 and the differences between the sexes. It could perhaps be that the increasing use of disfluencies is an effect of sociolect. For instance, if a speaker’s dialect includes words and/or constructions that are outdated or rare today, the
speaker might hesitate in using those with younger speakers, leading to increased disfluencies and complex syntax due to circumlocutions. Given the history of cultural gender differences, it may not be surprising to find a lag between men and women (e.g., due to slower entry into the workplace for older women). In such case, one would expect to find that the age difference between speakers influences the grammatical diversity, which was not present in our data. Furthermore, one would expect variables such as the degree of acquaintance of the speakers to play a role in these factors. However, after controlling for such degree of acquaintance, Bortfeld et al. (2001) found a pattern of increase in disfluencies very much like that reported here for men. Together with the evidence for neurophysiological effects of aging in areas relevant to language, and even patterns of decay also beginning at around age 40 (Sowell et al., 2003), it seems more parsimonious to attribute the differences observed here to actual effects of aging.

The results indicate that men and women exhibit different patterns of aging with regard to their linguistic performance. Sex differences in language processing have often been reported in the literature (see, Ullman et al., 2007, for a review). In this respect, the findings that women use more inflectional morphology and show fewer disfluencies than men are perhaps not very surprising. Women are known to outperform men in both of these, even from an early age (e.g., Hartshorne & Ullman, 2006). It is more surprising that men show an overall increased diversity over women in the use of syntax itself. Although most studies on language abilities have found—when anything—higher performance for women, some theories have proposed that men should in fact be better at tasks involving “procedural” processes, such as those necessary for processing syntactic regularities (Hartshorne & Ullman, 2006; Ullman et al., 2002). Most surprising of all is the accelerated pattern of change found for men. It seems that men’s inflectional and syntactic abilities and fluency peak at around age 45, decreasing from there on. One would think that this could point toward an earlier onset of dementias for men than for women. The literature, however, indicates that—if anything—it is women who show a higher incidence of dementias (cf. Ruitenberg, Ott, van Swieten, Hofman, & Breteler, 2001). The fact that the reduction in overall fluency (i.e., increased pauses, false starts, self-corrections, etc.) in men seems to be very strong over and above any effects of syntax suggests that there are general cognitive, not purely linguistic, mechanisms of importance for language performance, playing an important role in the breakdown of linguistic skills. The procedural/declarative distinction drawn by Ullman and his colleagues offers one tentative explanation for these patterns. However the declarative aspects of language (i.e., vocabulary and knowledge) improve throughout the lifespan for both sexes, the procedural abilities subserving grammatical processing begin to decay in middle age, with a later onset of this decay for women than for men.

Finally, a note of caution is owed here. Anyone who has talked to—otherwise healthy—older men knows that they often have no difficulties in either speaking or understanding. The changes reported in this study do not necessarily constitute deficits. They are relatively small-scale differences in performance. Even the most marked of these, the disfluencies (i.e., from Fig. 8 one can deduce that 60-year old men produce on average 42% more disfluencies per sentence than women do, and 27% more than 45-year old men.
do), may not be an indication of poor performance. Some authors have even found that such disfluencies might actually be beneficial for listeners, who may be able to compensate for them and actually facilitate their understanding (e.g., Brennan & Schober, 2001; Lau & Ferreira, 2005). In sum, the differences reported here are important in offering insights into the processes involved in mental aging, but whether they point to any form of impairment remains an open question.

Acknowledgments

I thank John W. Du Bois, Roger Levy, Michael Ramscar, Petar Milin, and two anonymous reviewers for helpful suggestions on this paper, even if disagreement remains on some aspects.

Notes

1. As only the year of birth was available, all birth dates were set to July 1 of the corresponding year.
2. I also considered a possible random effect of the speaker’s identity. However, considering this effect was problematic: The number of conversations per speaker was generally small, with many speakers participating in a single conversation. This is compounded by the overwhelming majority of the age variance being between speakers (rather than within speakers). In such a situation, it is sometimes not advisable to include a random effect (e.g., Clark & Linzer, 2015). In GAMMs, this situation is aggravated, resulting in considerable shrinkage on the nonlinear effect estimates that is not easily detected by correlations. Much of the systematic nonlinear effects of age is erroneously attributed to the (non-systematic) speaker random effect coefficients, to the point that one could then analyze those random effect adjustments as a function of speaker sex and age and obtain the very same pattern of results reported here. Therefore, this random effect was discarded from the regressions.
3. Notice, however, that speaking less does not necessarily imply using a poorer vocabulary or syntax, especially when—as is the case in this study—the amount of speech produced by each speaker is specifically considered as a nonlinear predictor separately from any other predictors under consideration.
4. The estimates for significance of nonlinear interactions in GAMM are estimated by model comparison of the random effects part of the model and are only very rough approximations testing only the difference in degrees of freedom of the smoothers and not their difference in shape. Even if the effect is marginally significant, I decided to keep this interaction as the models with the interaction definitely improved on the models without it. According to Akaike’s Information Criteria (AIC), the model with the interaction was indeed better. The AIC difference
approached two units (1.96), generally interpreted as only weak support for the alternative model without the interaction. Further support for this choice is the clearly different shapes (which the \( p \)-value does not test) between the effects for each sex.

5. One obtains identical results—with the evident rescaling—if one analyzes the number of disfluencies per word instead of the number of disfluencies per clause.

6. All analyses were also conducted removing the six points with ages above 63. This did not substantially change the results.

7. Additional nonlinear effects of the age difference between speakers or its absolute value were added in the regressions, none approached significance (\( F < 1 \) in all cases but the absolute value of the age difference’s effect on the number of disfluencies, for which \( F(1, 1284.01) = 1.673, p = .196 \)).

References


---

**Supporting Information**

Additional Supporting Information may be found in the online version of this article:

**Appendix S1.** Validity of the entropy bias adjustment method.

**Appendix S2.** Independent component analysis.