Word-Frequency Distributions in Chinese- and English- Speaking Older Adults: An Analysis across Languages and Cognitive Statuses

Tongfu Yang[1], Lihe Huang[1], Tsy Yih[1#*]

[1] School of Foreign Studies, Tongji University, Shanghai, China
[#] Tsy Yih is the transliteration of the name of the first author in his mother tongue, Shanghai Wu Chinese. He is also known as Zi Ye in Mandarin pinyin.
[*] Corresponding author's email: yihtsy@outlook.com

DOI: https://doi.org/10.53482/2025_59_426

ABSTRACT

This study investigates how the word-frequency distributions in spoken language reflect cross-linguistic and cognitive differences in older adults. We analyzed *Cookie Theft* picture descriptions from 96 older adults: 48 Mandarin speakers (24 cognitively impaired and 24 cognitively normal) and 48 English speakers (24 cognitively impaired and 24 cognitively normal) and modeled their word frequency distributions using three functions: Zipf, Zipf-Mandelbrot, and Exponential model. All three models showed excellent goodness of fit at both group and individual levels, indicating that the basic Zipfian structure of lexical distributions is preserved in late life and is not disrupted by mild cognitive impairment. As for the fitting parameters, however, the decay parameter *a* in the Exponential and Zipf models consistently distinguished Mandarin from English, suggesting that language-specific lexical patterns are robustly encoded in the slope of the distribution but that adding a shift parameter can dampen how clearly *a* reflects them. By contrast, differences between cognitive groups were weak and inconsistent, implying that parameter *a* provides only a coarse and context-dependent reflection of cognitive status in short, constrained picture-description tasks.

Keywords: Zipf's law, Cognitive impairment, Cross-linguistic analysis.

1 Introduction

Human language can be viewed as a self-organizing complex system whose internal structural regularities, or potential regularities, can be effectively revealed by quantitative methods (Ferrer-i-Cancho 2018; Köhler 1987; Oudeyer 2006; Steels 2000). In the context of population aging, this quantitative perspective on language is particularly relevant. Age-related changes in cognition, especially those associated with Alzheimer's disease (AD) and its prodromal stages have been repeatedly shown to manifest systematically in spoken and written language (e.g., Liu et al. 2021; Sand Aronsson et al. 2021;

Orimaye et al. 2017; Gao and He 2025), making language an ecologically valid, multidimensional window into age-related cognitive decline and a promising tool for early detection and longitudinal monitoring. Therefore, investigating the systemic regularities of these linguistic changes holds significant clinical and theoretical value.

Among the various regularities observed in language, the power-law pattern exhibited in word-frequency distributions is one of the most robust phenomena across languages and levels of linguistic structure. It has been documented at multiple levels of analysis, from micro-level measures such as dependency distance and the length of linguistic units (e.g., Liu 2009; Sigurd et al. 2004) to more macro-level domains including language evolution (Bentz 2014) and acquisition (Ellis 2012). This regularity is most commonly captured by the classical Zipf's law (Zipf 1949), which succinctly characterizes the inverse relationship between a word's frequency and its rank.

$$f_r = Cr^{-a}$$

However, because the classical Zipf model has limitations in fitting empirical data, especially at the high-frequency end of the distribution, subsequent work has proposed refined models such as the Zipf-Mandelbrot (ZM) model (Mandelbrot 1966), which introduces an additional parameter to increase the flexibility of the fit.

$$f_r = C(r + b)^{-a}$$

As most previous studies have relied on Zipf and ZM power-law models to fit rank-ordered frequencies, Altmann (2018) introduced an Exponential function as a unified model for diversification phenomena, including rank-frequency distributions. Both families of models aim to capture the same basic empirical regularity, namely that frequency decreases monotonically as rank increases.

$$f_r = 1 + ae^{-br}$$

Taken together, these models provide powerful mathematical tools for quantifying the organizational structure of linguistic systems, yet empirical work has typically selected a single best-fitting model for a given dataset, with relatively little attention to systematic comparisons across models.

In research on older adults, most studies using Zipf's law have focused on specific characteristics of connected speech such as dependency-based measures (Liu et al. 2021; Sand Aronsson et al. 2021; Gao and He 2025) rather than on the rank-frequency distribution of words, even though works on aphasia have demonstrated that quantitative analyses of rank-frequency distributions can distinguish impaired speech from that of healthy controls (Neophytou et al. 2017; van Egmond et al. 2015). To our

knowledge, only one study has directly analyzed the rank-frequency distribution of words in the spontaneous speech of older adults (Abe and Otake-Matsuura, 2021).

Furthermore, a critical gap exists at the intersection of typological linguistics and clinical research. Quantitative studies have already established that rank-frequency distributions, differ significantly across typologically distinct languages (Popescu and Altmann 2008; Jiang and Liu 2015; Neophytou et al. 2017). Yet, this comparative cross-linguistic perspective has not been extended to aging populations. It remains unknown whether the linguistic manifestations of cognitive decline are largely universal, or whether they are modulated by the typological properties of a given language, such as Mandarin Chinese versus English.

Against this background, the present study applies three rank-frequency models: Zipf's law, the ZM model, and Altmann's Exponential diversification model to picture-description narratives produced by cognitively normal and cognitively impaired older adults in Mandarin Chinese and English. By jointly modeling word-frequency distributions across models, languages and cognitive groups, we aim to firstly assess how well each model captures the word-frequency structure of older adults' spoken language, secondly examine whether model parameters are sensitive to cross-linguistic differences between Mandarin and English, and lastly test whether these parameters can discriminate between cognitively normal and cognitively impaired speakers within each language. Specifically, we address the following research questions:

RQ1: Do word-frequency distributions of older adults' spoken language in Mandarin Chinese and English conform to Zipf's law, the ZM model, and Exponential diversification model?

RQ2: Do the parameters of these three models show cross-linguistic differences between Mandarin- and English-speaking older adults, and are these differences dependent on the choice of model?

RQ3: Do the model parameters distinguish between cognitively impaired and cognitively normal older adults within each language, and are such group differences robust across the three models?

2 Material

2.1 Participants

The speech data analyzed in this study were drawn from two existing corpora: the DementiaBank corpus (Becker et al. 1994) and the MCGD (Multimodal Corpus of Gerontic Discourse) (Zhou 2024). On the basis of language (Mandarin Chinese vs. English) and cognitive status (CI: cognitively impaired and CN: cognitively normal), we selected 96 speakers and divided them into four groups of equal size (24 speakers per group). For each participant, we extracted basic demographic information, including age, sex, years of education, and MMSE score. Descriptive statistics for age, sex, education, and MMSE are reported in **Table 1**.

Cognitive status (CI and CN) was determined using the Mini-Mental State Examination (MMSE) with education-specific cutoff points. In line with previous work on mild cognitive impairment, we adopted the following cut-offs for MMSE total scores: MMSE ≤ 19 for illiterate individuals, MMSE ≤ 22 for participants with elementary school education, MMSE ≤ 26 for those with middle school education and above (Jia et al. 2021).

Table 1: Group comparisons of demographic variables in the Chinese sample.

Language	Variable	CN	CI
Chinese	Number of participants	24	24
	Age (mean ± SD)	64.25 ± 9.40	73.33 ± 7.69
	Edu (mean ± SD)	10.62 ± 5.15	9.79 ± 4.51
	MMSE (mean ± SD)	27.88 ± 1.39	20.75 ± 4.58
	Sex (Female / Male) (n (%))	13 (54.2%) / 11 (45.8%)	15 (62.5%) / 9 (37.5%)
English	Number of participants	24	24
	Age (mean ± SD)	63.83 ± 7.70	73.67 ± 7.56
	Edu (mean ± SD)	13.58 ± 3.13	12.42 ± 2.34
	MMSE (mean ± SD)	28.21 ± 0.88	20.54 ± 3.98
	Sex (Female / Male) (n (%))	15 (62.5%) / 9 (37.5%)	17 (70.8%) / 7 (29.2%)

Notes: CI = cognitively impaired; CN = cognitively normal.

2.2 Corpus

All speech samples are based on the *Cookie Theft* picture descriptions. In the original corpora, the English data are provided in CHAT (.cha) format in DementiaBank, whereas the Mandarin Chinese data from MCGD are available as plain-text (.txt) transcripts. For the purposes of uniform processing and modelling, all files were converted into plain-text format after preprocessing.

Preprocessing included **speaker selection, data cleaning, and language-specific tokenization and normalization.** First, we retained only the participant's speech (the PAR: tier in the English CHAT files and the corresponding participant lines in the Chinese transcripts) and removed all utterances produced by examiners. During data cleaning, all non-lexical items were removed, including non-lexical filled pauses in the English transcripts (e.g., *&-uh*), as well as coding brackets and meta-linguistic annotations such as [+ gram] and [//]. Regarding lexical filled pauses, we strictly distinguished between interactional signals and production-related hesitation markers. English items such as *yeah* and *oh* and the Mandarin filler *en* "yes" were excluded as they functioned as backchannel responses to the examiner rather than as part of the picture description itself. The Mandarin token *zhège* "this" was retained in all cases. While *zhège* can function as a demonstrative, its non-demonstrative use typically serves as a filled pause indicating lexical retrieval difficulty during the description process. Unlike backchannels, these hesitation markers reflect the speaker's internal cognitive planning directly related to the task. Orthographic reconstruction marks like *spillin(g)* were changed to their full canonical form (e.g., *spilling*). At the same time, we preserved repetitions (e.g., *the the cookie*) and grammatically deviant forms whenever the lexical items were still identifiable, on the grounds that these features reflect genuine production

patterns and directly influence the resulting word frequency distributions. Finally, language-specific tokenization and normalization procedures were applied: for Mandarin Chinese, the cleaned transcripts were segmented into word-level tokens using the NLP tool THULAC (Sun et al. 2016); for English, participant speech was extracted from the .cha files and converted to plain text, tokenized based on space and punctuation, converted to lower case, with common contractions (e.g., *he's*) expanded (e.g., *he is*) and the resulting text further processed using the spaCy toolkit (Honnibal et al. 2020) for tokenization and lemmatization. All automated procedures were followed by manual verification. We inspected the transcripts of every participant on a case-by-case basis to ensure the highest level of accuracy.

After cleaning and tokenization, the corpus consisted of 96 *Cookie Theft* picture descriptions, corresponding to 96 processed text files (one per participant). For each file, we computed the total number of tokens (words) and types (distinct word forms) and derived a rank-frequency list in which word forms are ordered from the most to the least frequent. These rank-frequency tables constitute the primary input to the models fitted in the subsequent analyses. At the corpus level, we report the overall token and type counts separately for the Mandarin Chinese and English datasets (see **Table 2**). At the individual level, descriptive statistics regarding token counts, type counts, and Type-Token Ratio (TTR) are presented in

Table A2. To illustrate the structure of the derived frequency data, the first 5 tokens of each group are presented in **Table 3** (detailed data are available at https://github.com/toferyoung-wq/detailed-data).

Table 2: token, type, and TTR statistics by language and cognitive group.

Language	Group	Texts	Total tokens	Total types	Tokens (mean ± SD)	Types (mean ± SD)	TTR (mean ± SD)
Chinese	CI	24	2447	1242	101.96 (40.93)	51.75 (12.62)	0.54 (0.11)
Chinese	CN	24	2508	1401	104.50 (46.41)	58.38 (21.20)	0.59 (0.10)
English	CI	24	2139	1122	89.12 (39.90)	46.75 (12.52)	0.56 (0.10)
English	CN	24	2470	1379	102.92 (37.65)	57.46 (15.10)	0.58 (0.08)

Notes: CI = cognitively impaired; CN = cognitively normal.

Table 3: The first 5 tokens of each group

Language	Cognitive Status	Rank	Token	Frequency
Chinese	Cognitively Impaired	1	*zhège*	141
		2	*shì*	131
		3	*le*	104
		4	*zhè*	95
		5	*zài*	72
	Cognitively Normal	1	*zhège*	128
		2	*shì*	109
		3	*le*	81
		4	*de*	80
		5	*zài*	63
English	Cognitively Impaired	1	*the*	218
		2	*be*	196
		3	*and*	123
		4	*she*	51
		5	*i*	50
	Cognitively Normal	1	*the*	275
		2	*be*	233
		3	*and*	112
		4	*a*	69
		5	*to*	54

3 Methodology

3.1 Model Formulas and Fitting

Building on rank-frequency tables, we modeled the resulting word frequency distributions in Python (van Rossum and Drake 2009) at both group level (aggregated by language and cognitive status) and individual level (one model per participant). For each cleaned transcript and each aggregated corpus, the corresponding sequence (rank, frequency) was used as input to a series of non-linear fits implemented with SciPy (Virtanen et al. 2020).

At the modeling level, we considered three mathematical models of the rank-frequency distribution.

The Zipf model (1) assumes a power-law decay of frequency with rank, where C is a scale parameter and a is the Zipf exponent controlling the slope.

(1) $$f_r = Cr^{-a}$$

The ZM model (2) introduces *a* shift parameter *b,* which captures additional curvature in the high-frequency region.

(2) $$f_r = C(r + b)^{-a}$$

The Exponential model (3) includes *a* and *b*. *a* determines the height of the curve at low ranks and *b* controls the rate of exponential decay with rank. In the implementation, the Zipf and ZM models explicitly include the scale parameter *C,* whereas the Exponential model does not contain a separate *C* term[1].

(3) $$f_r = 1 + be^{-ar}$$

The parameter settings in the fitting procedure reflected the model structures. For each text, in the Zipf and ZM models the initial value of *C* was set to the frequency of the most frequent word in that text (i.e., the rank-1 token), and the remaining parameters were initialized at 1.0. Parameter bounds were imposed to avoid degenerate solutions: in the Zipf model, the exponent a was constrained to the interval [0, 10]; in the ZM model, *a* and *b* were constrained to [0, 10] and [0, 100], respectively. For the Exponential model, which does not contain a separate scale parameter *C*, the parameters *a* and *b* were both initialized at 1.0 and constrained to [0, 1000] and [0, 100]. These ranges were chosen to be wide enough to cover all empirically plausible values observed in preliminary fits and in previous work on Zipfian word-frequency distributions, but to exclude clearly implausible or numerically unstable solutions (e.g., negative exponents). Importantly, all best-fitting parameter estimates in our data fell well inside these bounds, indicating that the constraints served only to stabilize the optimization rather than to artificially restrict the models. The maximum number of function evaluations in curve_fit was set to 20,000 to improve convergence for long-tailed distributions.

For each model and each text, the optimization returned a set of best-fitting parameters and the corresponding predicted frequencies. Model performance was summarized using the coefficient of determination R^2 (4) which is computed from the squared deviations between the observed frequencies f_r and the model predictions $\hat{f}_r$ relative to the variance of f_r.

[1] The model was originally expressed as $f_r = 1 + ae^{-br}$. For comparability with the Zipf and ZM models, we simply swap the positions of *a* and *b,* using *a* denote the decay parameter.

(4) $$R^2 = 1 - \frac{\sum_{r=1}^{R}(f_r - \hat{f}_r)^2}{\sum_{r=1}^{R}(f_r - \bar{f})^2}$$

Thus, for every participant and for each of the three models we obtained a set of fitted parameters (*C*, *a*, *b*) together with an associated R^2 value. All individual-level parameter estimates and goodness-of-fit indices were exported as CSV files, along with detailed rank-wise tables that include, for every word, its rank, observed frequency, predicted frequency, and residual. These exported tables form the basis for the subsequent statistical analyses and visualizations.

3.2 Statistical Analysis and Visualization

All statistical analyses were carried out in R (R Core Team 2023). We used the packages car (Fox and Weisberg 2019) for analysis of variance, emmeans (Lenth 2020) for estimated marginal means and post-hoc contrasts, and dplyr (Wickham et al. 2023) for data handling. For each model parameter of interest, we analyzed the individual-level estimates obtained from the Python fitting procedure with Language (Chinese vs. English) and Cognitive Status (CI vs. CN) as between-subject factors.

Because the raw parameter distributions were typically right-skewed and strictly positive, we applied a natural-log transformation to the parameter estimates so that the data better met the assumptions of linear modelling and ANOVA. For each parameter, we then analyzed the effects of language (Chinese vs. English) and cognitive status (CI vs. CN) in a two-way framework. Whenever the language × cognitive status interaction was significant, we conducted planned post-hoc comparisons using the emmeans package. Specifically, we focused on four a priori contrasts: CI vs. CN within each language, and Chinese vs. English within each cognitive-status group. Statistical significance for these planned contrasts was determined on the basis of Holm-corrected *p*-values, in order to control the family-wise error rate within this set of comparisons.

The visualization was carried out in Python. Data were processed by Pandas (McKinney 2010) and NumPy (Harris et al. 2020) and visualized by Matplotlib (Hunter 2007) with significance markers manually added based on the results of the statistical tests.

4 Results and Discussion

4.1 Goodness of fit

The first aim of this study was to examine whether the rank-frequency distributions of older adults' *Cookie Theft* descriptions in English and Mandarin can be captured by three standard diversification models (Exponential, Zipf, ZM). At the group level, the R^2 values (see **Figure 1** and **Figure 2**) indicate excellent fits in both languages, with values ranging from 0.924 to 0.995 for Chinese and from 0.859 to 0.955 for English, and no clear differences in goodness of fit between cognitively impaired and cognitively normal speakers within each language. At the individual level (see **Figure 3**), the models also

performed well: most participants showed R^2 values above 0.800, with only a few exceptions (e.g., one Chinese CN speaker with R^2 = 0.769 under the Zipf model), and again no systematic differences in R^2 between CI and CN groups within language group for the three models.

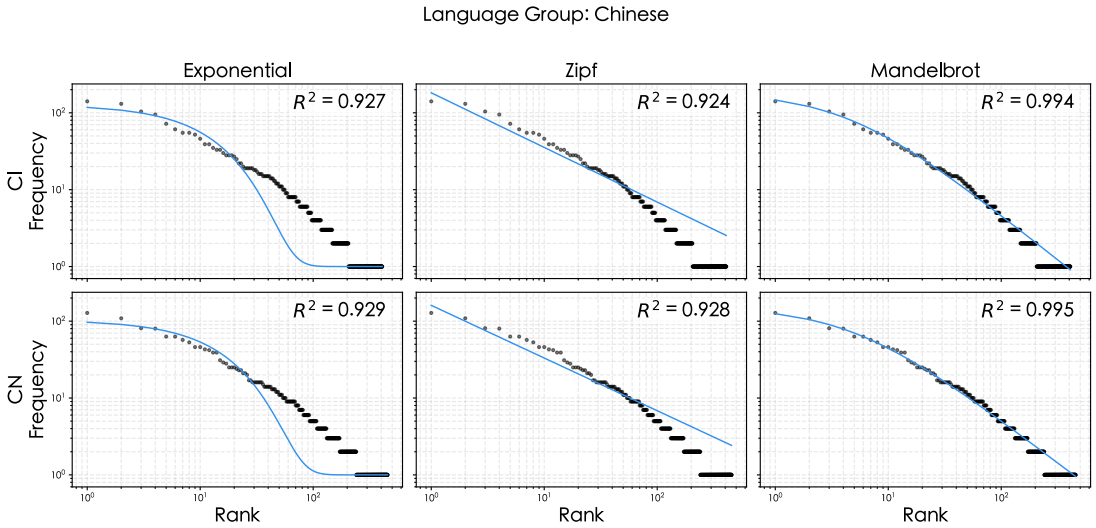

Figure 1: Log-log rank-frequency plots for Zipf, ZM and Exponential models of the Chinese group at group level, with R^2 values shown for each.

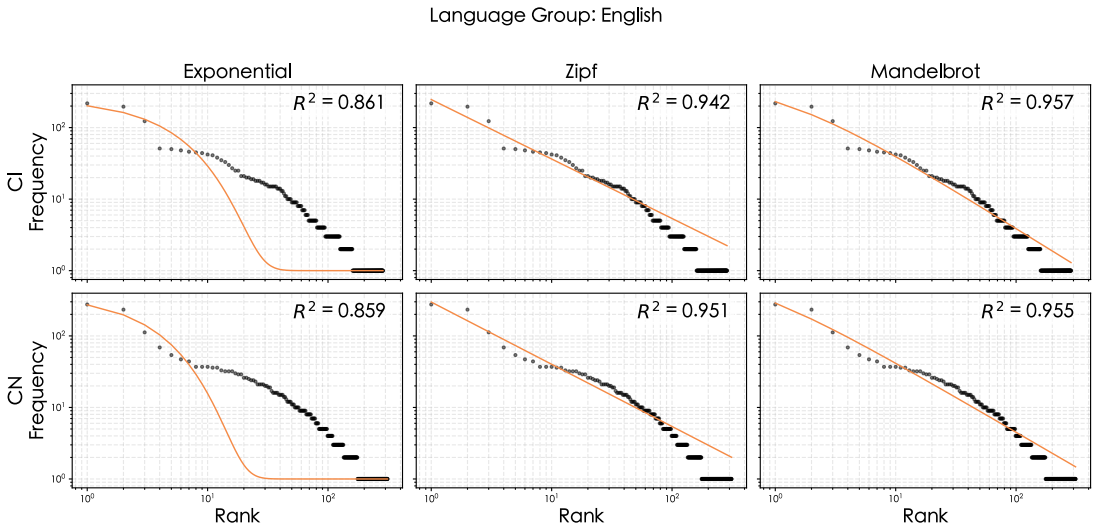

Figure 2: Log-log rank-frequency plots for Zipf, ZM and Exponential models of the English group at group level, with R^2 values shown for each.

More fine-grained inspection of the group-level fits reveals systematic language differences in how the three models capture the distributions. The Chinese data are better fitted by the Exponential and ZM models than the English data, whereas English achieves slightly higher R^2 than Chinese under the basic Zipf model. For the exponential model, this difference appears to be driven by the overall shape of the distributions: in Chinese, the high-frequency head and mid-frequency range closely follow the

exponential curve and the low-frequency tail deviates only mildly, whereas in English the mid and low ranks show larger deviations from the Exponential prediction. Comparing the Zipf and ZM fits leads to a similar conclusion. In Chinese, adding the ZM shift term (parameter *b*) yields a marked improvement over the basic Zipf model, with R^2 increasing from 0.924 to 0.994 (CI) and from 0.928 to 0.995, which points to a more pronounced high-frequency head in the Chinese distributions. In English, by contrast, the head correction improves the fit only marginally with R^2 increasing from 0.942 to 0.957 (CI) and from 0.951 to 0.955 (CN); the English curves tend to show a very steep drop at the top ranks followed by a relative flattening or slight rise in the mid-frequency range.

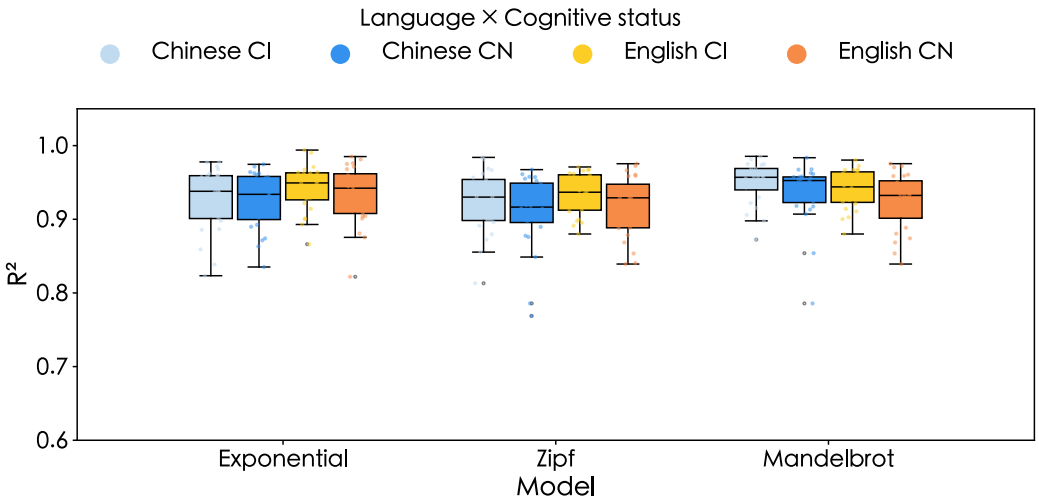

Figure 3: Individual-level R^2 values for the three models, with model names on the x-axis and colors indicating groups (from left to right: Chinese CI, Chinese CN, English CI, English CN).

Overall, these results suggest that, although Chinese and English differ slightly in the detailed shape of their word-frequency distributions, all three models capture the word-frequency structure of older adults' *Cookie Theft* descriptions well. Cognitive impairment does not appear to substantially alter the overall rank-frequency pattern.

4.2 Differences in Model Parameters

To address RQ 2 and 3, we conducted two-way ANOVAs on the log-transformed decay parameter *a* for all three models, with language and cognitive status as between-subject factors. The detailed results of two-way ANOVA are presented in **Table A1.** Significant language × cognitive status interactions were found in Exponential and Zipf, respectively F (1,92) = 6.210, $p = 0.015$, $\eta p^2 = 0.060$; F (1,92) = 7.820, $p = 0.006$, $\eta p^2 = 0.080$. We therefore focus on the corresponding simple effects based on post-hoc tests (see **Table 4** for details). For the ZM model, however, the log-transformed parameters still exhibited clear violations of the homogeneity-of-variance assumption (Levene's test for *a*: F (3, 92) = 7.310, $p < 0.001$). Moreover, neither the main effects ($p = 0.344$ for language and $p = 0.151$ for cognitive status)

nor their interaction (p = 0.296) reached statistical significance for the ZM parameters. Therefore, we do not pursue further inferential analyses for this model and treat its results as purely supplementary.

Table 4: Simple-effects analyses of parameters *a* (log-transformed) of Exponential and Zipf model across four groups.

Models	Contrast	Estimate (log scale)	SE	t (92)	*p* (Holm)
Exponential	Chinese_CI − Chinese_CN	0.213	0.129	1.654	0.130
	English_CI − English_CN	−0.241	0.129	1.869	0.130
	Chinese_CI − English_CI	−0.392	0.129	−3.038	**0.009****
	Chinese_CN − English_CN	−0.847	0.129	−6.561	**< 0.001*****
Zipf	Chinese_CI − Chinese_CN	0.113	0.048	2.338	**0.045***
	English_CI − English_CN	−0.078	0.048	−1.616	0.110
	Chinese_CI − English_CI	−0.120	0.048	−2.479	**0.045***
	Chinese_CN − English_CN	−0.311	0.048	−6.433	**< 0.001*****

Notes: CI = cognitively impaired; CN = cognitively normal. Estimates are contrasts on the log-transformed decay parameter a (first group minus second group), based on estimated marginal means from two-way ANOVAs with language (Chinese vs. English) and cognitive status (CI vs. CN) as between-subject factors. SE = standard error of the contrast; t(92) = t-statistic with 92 residual degrees of freedom; *p* (Holm) = Holm-adjusted *p*-value for the planned comparisons. * $p < .05$; ** $p < .01$; *** $p < .001$.

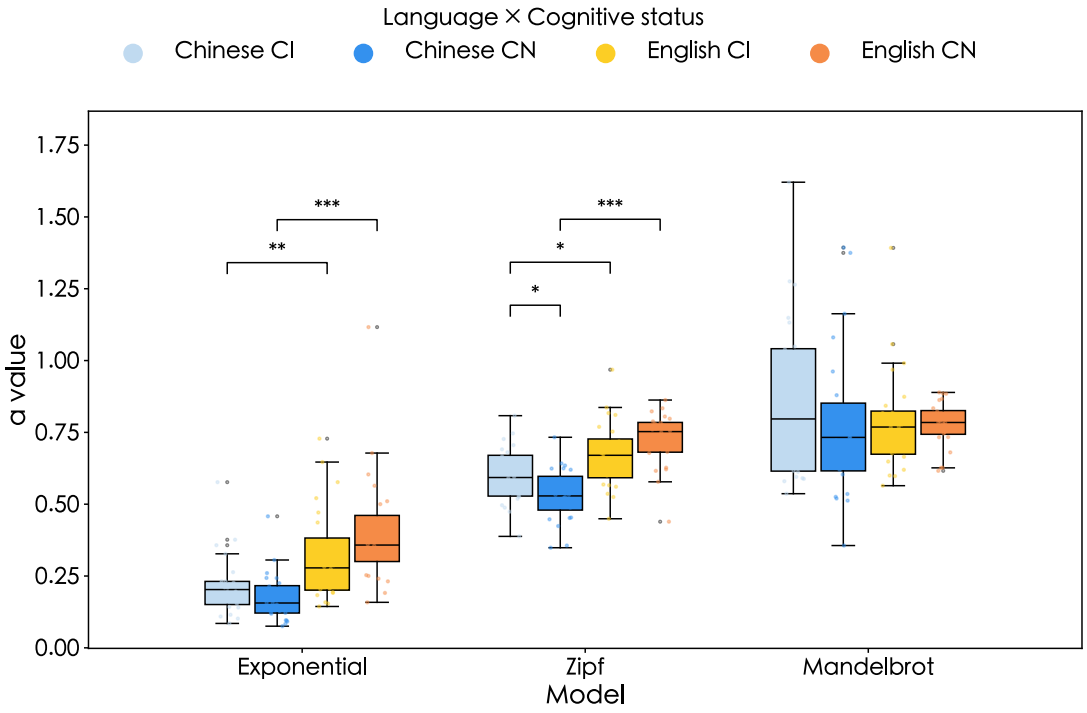

Figure 4: Individual-level values of the decay parameter *a* for each model. Asterisks indicate significance (*p < .05, **p < .01, *p < .001).

4.2.1 Changes in Parameter *a* across Languages

In both the Exponential and Zipf models, after controlling for cognitive status, both the boxplots (see **Figure 4**) and the simple-effects analyses (see **Table 4**) indicate that across all comparable contrasts, the English groups exhibited significantly larger decay parameters *a* (Exponential: $p = 0.009$ in CI and $p < 0.001$ in CN; Zipf: $p = 0.045$ in CI and $p < 0.001$), indicating steeper distributional slopes than the Chinese groups, with log-scale differences ranging from 0.120 to 0.847. This discrepancy was particularly pronounced for the Exponential model in the cognitively normal group (Chinese_CN − English_CN = − 0.847).

At first glance, this finding seems to contradict previous work. Earlier studies have reported that morphologically poorer languages show steeper Zipfian slopes: for instance, Bentz et al. (2014) found that Modern English has a steeper slope than Old English, and Neophytou et al. (2017) reported steeper slopes for English than for morphologically richer Greek. On this view, Mandarin Chinese which is typically described as an analytic language with virtually no inflectional morphology would be expected to have a steeper slope than English. Our data point in the opposite direction.

We argue that this apparent contradiction is unlikely to reflect macro-level typological differences, but rather arises from the specific task (*Cookie Theft* picture description) and the language-specific properties of the resulting corpora. Inspection of the data shows that, in the English group, the highest-frequency items are predominantly *be, and* and *the*, whereas in the Chinese group they are mainly *zhègè*

"this", *zhè* "this" and *shì* "is". At the task level, the *Cookie Theft* picture invites both object naming and event description. Differences in familiarity with the scene may lead Chinese- and English-speaking older adults to focus on different aspects: English speakers tend to attend more to the actions and frequently use *be + V-ing* constructions to describe ongoing events, which raises the frequency of the lemma *be*, whereas many Chinese speakers, who are less familiar with the depicted setting, place greater emphasis on object identification and often rely on *shì* "be"[2]. At the language level, English speakers also tend to use overt conjunctions such as *and* to maintain discourse coherence, while Mandarin speakers often do not employ an explicit coordinator in analogous contexts. Likewise, the rich article system of English routinely pushes the into the very top ranks, whereas the use of Chinese classifiers such as *yígè* "one" is more optional and more strongly influenced by individual stylistic preferences than by grammatical obligation. Together, these task- and corpus-specific factors lead to extremely high frequencies of *and*, *the* and *be* in the English data, and a much larger gap between the top ranks and the mid ranks, whereas in the Chinese data the contrast between the head and the middle of the distribution is less pronounced (can also be observed at group level, see **Figure 1**, **Figure 2** and **Table 3**). This offers a plausible explanation for why the English distributions in our data appear steeper than the Chinese ones, despite typological expectations based on morphological richness.

Taken together, these results suggest that the decay parameter a can indeed capture specific characteristics of the rank-frequency distribution. First, language-specific properties, especially at the lexical level, are clearly reflected in the shape of the distributions. At the same time, model properties also matter: for example, no significant effects were detected for the ZM model, indicating that introducing additional parameters may increase R^2 but at the cost of masking corpus-specific structure. This underscores the value of applying more than one model in future work rather than relying on a single specification. Finally, our conclusions are constrained by the limited size of the corpus and the relatively simple, picture-description task. Studies with larger samples and more diverse, cognitively demanding tasks may yield more robust and generalizable insights into how *a* relates to language use.

4.2.2 Changes in Parameter *a* across Cognitive Statuses

Across cognitive-status groups, the associations between the decay parameter *a* and cognition appear relatively weak. Compared with the highly pronounced differences across languages, the differences between CI and CN are much less distinct: simple-effects analyses show that only one significant effect was detected in the Chinese group under the Zipf model (p =0.045, see **Table 4**) and the cross-cognitive trends even go in opposite directions across languages (from CI to CN, *a* generally decreases in Chinese but increases in English, see **Figure 4**).

[2] In Mandarin, *zhègè* "this" does not always function as a genuine demonstrative; it is often used as a filled pause that buffers lexical retrieval difficulty. When *zhè* "this" and *shì* "is" occur together, the phrase typically has a clear demonstrative-copular function which means *this is* or *there is* in English, and in this sense the copula *shì* "is" is the item that more directly reflects Chinese older adults' focus on object naming in the task.

Several factors may underlie this pattern. First, the differences between cognitively impaired and normal group in our corpora are relatively subtle, making them difficult to capture with a single slope parameter. Previous work suggests that *a* has good discriminative power for severe language disorders, but is much less sensitive to finer-grained variation: van Egmond et al. (2015) showed that non-fluent aphasic speech has a reliably steeper Zipfian slope than healthy control speech, and Neophytou et al. (2017) replicated this pattern for both fluent and non-fluent aphasia in English, while finding no systematic differences between the two aphasic subtypes. In addition, Abe and Otake-Matsuura (2021) reported no robust association between Zipf's exponent and cognitive scores in cognitively screened Japanese elders. Second, the behavior of *a* is model-dependent: although the decay parameter plays a comparable conceptual role across the Exponential, Zipf and ZM models, the cross-model discrepancies indicate that each model emphasizes different aspects of the rank-frequency shape. Third, in relatively short texts the parameter *a* is easily distorted by idiosyncratic high-frequency patterns, so the way *a* changes across cognitive-status groups can differ substantially between languages. For example, in the English group, the higher *a* values observed in the CN speakers may stem from their longer descriptions (see **Table 2**, CN has longer TTR than CI) which naturally contain more instances of function words such as *and* and *the*, thereby raising the head of the distribution, whereas the higher *a* in Chinese CI speakers may be driven by frequent use of *zhège* "this" as a filled pause in contexts of word-finding difficulty.

Overall, the parameter *a* appears to distinguish impaired from unimpaired language only when the underlying differences are relatively pronounced. Moreover, given the limited text length and the complexity of the models, our findings suggest that the decay parameter *a* offers only a coarse, context-dependent reflection of cognitive status, and its usefulness as a marker of cognition in short, constrained tasks such as *Cookie Theft* descriptions requires further investigation.

5 Conclusion

In this study, we modeled the word-frequency distributions of *Cookie Theft* picture descriptions produced by older Mandarin and English speakers, with and without cognitive impairment, using three models: Zipf, ZM and Exponential model. All three models provided excellent goodness of fit at both the group and individual levels, indicating that Zipfian structure is preserved across languages and cognitive statuses in this task. At the parameter level, however, the models behaved somewhat differently and were only weakly sensitive to subtle contrasts. The decay parameter *a* in the exponential and Zipf models reliably distinguished Mandarin from English, whereas the Zipf-Mandelbrot model did not, suggesting that lexical-level language-typological differences are indeed captured in the shape of the distributions. By contrast, the ability of *a* to differentiate cognitive status was limited: only one significant difference was found for Mandarin under the Zipf model, which is likely due to the relatively mild degree of language impairment in our CI group.

However, this study has several limitations. First, the sample size is relatively small, which limits the statistical power of the analyses and reduces the robustness of group comparisons. With only 24 speakers per group, subtle effects may go undetected and parameter estimates may be more vulnerable to individual variability. Second, the *Cookie Theft* picture-description task is simple and may not elicit representative speech that fully captures the properties of participant's speech production, nor does it strongly amplify differences between cognitively healthy and impaired speakers. Third, cognitive status was dichotomized into CI and CN based solely on MMSE, which is a relatively coarse grouping. Future work should therefore use longer speech samples, adopt more fine-grained cognitive classifications, and employ more complex and varied elicitation tasks in order to better assess the extent to which the decay parameter a can discriminate between intact and impaired language.

Acknowledgements

This research was supported by the National Social Science Fund of China Grant: [24BYY120] awarded to Lihe Huang.

References

Abe, M. S., Otake-Matsuura, M. (2021). Scaling laws in natural conversations among elderly people. *PLOS ONE*, 16(2), e0246884. https://doi.org/10.1371/journal.pone.0246884.

Altmann, G. (2018). *Unified modeling of diversification in language*. Lüdenscheid: RAM-Verlag.

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., McGonigle, K. L. (1994). The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6), pp. 585-594.

Bentz, C., Kiela, D., Hill, F., Buttery, P. (2014). Zipf's law and the grammar of languages: A quantitative study of Old and Modern English parallel texts. *Corpus Linguistics and Linguistic Theory*, 10(2), 2014, pp. 175-211. https://doi.org/10.1515/cllt-2014-0009.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, 32, pp. 17-44. https://doi.org/10.1017/S0267190512000025.

Ferrer-i-Cancho, R. (2018). Optimization models of natural communication. *Journal of Quantitative Linguistics*, 25(3), pp. 207-237. https://doi.org/10.1080/09296174.2017.1366095.

Fox, J., Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks, CA: Sage.

Gao, N., He, Q. (2025). A corpus-based dependency study of the syntactic complexity in the connected speech of Alzheimer's disease. *Aphasiology*, 39(11), pp. 1456-1479. https://doi.org/10.1080/02687038.2024.2434858.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585, pp. 357-362.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), pp. 90-95.

Jia, X., Wang, Z., Huang, F., et al. (2021). A comparison of the Mini-Mental State Examination (MMSE) with the Montreal Cognitive Assessment (MoCA) for mild cognitive impairment screening in Chinese middle-aged and older population: A cross-sectional study. *BMC Psychiatry*, 21, 485. https://doi.org/10.1186/s12888-021-03495-6.

Jiang, J., Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction and the implications—Based on a parallel English Chinese dependency treebank. *Language Sciences*, 50, pp. 93-104. https://doi.org/10.1016/j.langsci.2015.04.002.

Köhler, R. (1987). System theoretical linguistics. *Theoretical Linguistics*, 14(2–3), pp. 241-247. https://doi.org/10.1515/thli.1987.14.2-3.241.

Lanzi, A. M., Saylor, A. K., Fromm, D., Liu, H., MacWhinney, B., Cohen, M. (2023). DementiaBank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2), pp. 426-438. https://doi.org/10.1044/2022_AJSLP-22-00281.

Lenth, R. V. (2020). *emmeans: Estimated marginal means, aka least-squares means* [R package]. Retrieved from https://CRAN.R-project.org/package=emmeans.

Liu, H. (2009). Probability distribution of dependencies based on a Chinese dependency treebank. *Journal of Quantitative Linguistics*, 16(3), pp. 256-273. https://doi.org/10.1080/09296170902975742.

Liu, J., Zhao, J., Bai, X. (2021). Syntactic impairments of Chinese Alzheimer's disease patients from a language dependency network perspective. *Journal of Quantitative Linguistics*, 28(3), pp. 253-281. https://doi.org/10.1080/09296174.2019.1703485.

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In: P. F. Lazarsfield, N. W. Henry (Eds.). *Readings in Mathematical Social Sciences* (pp. 350-368). Cambridge, MA: MIT Press.

McKinney, W. (2010). Data structures for statistical computing in Python. In: S. van der Walt, J. Millman (Eds.). *Proceedings of the 9th Python in Science Conference* (pp. 51-56). https://doi.org/10.25080/Majora-92bf1922-00a.

Neophytou, K., van Egmond, M., Avrutin, S. (2017). Zipf's law in aphasia across languages: A comparison of English, Hungarian and Greek. *Journal of Quantitative Linguistics*, 24(2-3), pp. 178-196. https://doi.org/10.1080/09296174.2016.1263786.

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. https://doi.org/10.1186/s12859-016-1456-0.

Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech*. Oxford: Oxford University Press.

Popescu, I. I., Altmann, G. (2008). Hapax legomena and language typology. *Journal of Quantitative Linguistics*, 15(4), pp. 370-378. https://doi.org/10.1080/09296170802326699.

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Sand Aronsson, F., 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), pp. 900-913. https://doi.org/10.1080/02687038.2020.1742282.

Sigurd, B., Eeg-Olofsson, M., Van Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. Studia Linguistica, 58(1), pp. 37-52. https://doi.org/10.1111/j.0039-3193.2004.00109.x.

Steels, L. (2000). Language as a Complex Adaptive System. In: Schoenauer, M. et al. (Eds.). *Parallel Problem Solving from Nature PPSN VI. Lecture Notes in Computer Science* (Vol. 1917), pp. 17-26. Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/3-540-45356-3_2.

van Egmond, M., van Ewijk, L., Avrutin, S. (2015). Zipf's law in non-fluent aphasia. Journal of Quantitative Linguistics, 22(3), pp. 233-249. https://doi.org/10.1080/09296174.2015.1037158.

van Rossum, G., Drake, F. L. (2009). *Python 3 reference manual*. Scotts Valley, CA: CreateSpace Independent Publishing Platform.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17, pp. 261-272.

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D. (2023). *dplyr: A grammar of data manipulation* [R package]. Retrieved from https://dplyr.tidyverse.org.

Zhou, D. (2024). Multimodal corpus of geronto discourse: Construction and reflection. *Linguistic Research*, 36, pp. 20-34.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Appendix

Table A1: Two-way ANOVA results for the log-transformed decay parameter *a* across models

Model	Effect	df_1	df_2	F	*p*	partial η^2
Exponential	Language	1	92	9.232	0.003**	0.091
	Cognitive Status	1	92	2.736	0.102	0.029
	Language × Cognitive Status	1	92	6.205	0.015**	0.063
Zipf	Language	1	92	6.144	0.015**	0.063
	Cognitive Status	1	92	5.465	0.022*	0.056
	Language × Cognitive Status	1	92	7.817	0.006**	0.078
Mandelbrot	Language	1	92	0.906	0.344	0.01
	Cognitive Status	1	92	2.1	0.151	0.022
	Language × Cognitive Status	1	92	1.103	0.296	0.012

Notes: *$p < .05$, **$p < .01$, *$p < .001$

Table A2: Token, type, and TTR statistics at individual level.

Language	Cognitive Status	Participants	Token	Type	TTR
Chinese	Cognitively Impaired	CHI001	84	44	0.52
Chinese	Cognitively Impaired	CHI002	64	44	0.69
Chinese	Cognitively Impaired	CHI003	150	77	0.51
Chinese	Cognitively Impaired	CHI004	64	39	0.61
Chinese	Cognitively Impaired	CHI005	180	77	0.43
Chinese	Cognitively Impaired	CHI006	97	49	0.51
Chinese	Cognitively Impaired	CHI007	84	55	0.65
Chinese	Cognitively Impaired	CHI008	96	43	0.45
Chinese	Cognitively Impaired	CHI009	149	62	0.42
Chinese	Cognitively Impaired	CHI010	85	58	0.68
Chinese	Cognitively Impaired	CHI011	78	53	0.68
Chinese	Cognitively Impaired	CHI012	85	44	0.52
Chinese	Cognitively Impaired	CHI013	65	34	0.52
Chinese	Cognitively Impaired	CHI014	75	48	0.64
Chinese	Cognitively Impaired	CHI015	95	56	0.59
Chinese	Cognitively Impaired	CHI016	221	75	0.34
Chinese	Cognitively Impaired	CHI017	69	32	0.46
Chinese	Cognitively Impaired	CHI018	88	56	0.64
Chinese	Cognitively Impaired	CHI019	114	46	0.40
Chinese	Cognitively Impaired	CHI020	74	49	0.66
Chinese	Cognitively Impaired	CHI021	141	67	0.48
Chinese	Cognitively Impaired	CHI022	84	45	0.54
Chinese	Cognitively Impaired	CHI023	140	51	0.36
Chinese	Cognitively Impaired	CHI024	65	38	0.58
Chinese	Cognitively Normal	CHI025	91	60	0.66
Chinese	Cognitively Normal	CHI026	139	71	0.51
Chinese	Cognitively Normal	CHI027	64	42	0.66
Chinese	Cognitively Normal	CHI028	174	92	0.53
Chinese	Cognitively Normal	CHI029	56	35	0.62
Chinese	Cognitively Normal	CHI030	101	65	0.64
Chinese	Cognitively Normal	CHI031	92	61	0.66
Chinese	Cognitively Normal	CHI032	61	28	0.46
Chinese	Cognitively Normal	CHI033	139	75	0.54
Chinese	Cognitively Normal	CHI034	107	50	0.47
Chinese	Cognitively Normal	CHI035	201	106	0.53
Chinese	Cognitively Normal	CHI036	124	61	0.49
Chinese	Cognitively Normal	CHI037	59	37	0.63
Chinese	Cognitively Normal	CHI038	116	68	0.59
Chinese	Cognitively Normal	CHI039	169	93	0.55

Chinese	Cognitively Normal	CHI040	108	58	0.54
Chinese	Cognitively Normal	CHI041	83	45	0.54
Chinese	Cognitively Normal	CHI042	36	31	0.86
Chinese	Cognitively Normal	CHI043	47	30	0.64
Chinese	Cognitively Normal	CHI044	187	79	0.42
Chinese	Cognitively Normal	CHI045	77	47	0.61
Chinese	Cognitively Normal	CHI046	68	49	0.72
Chinese	Cognitively Normal	CHI047	68	43	0.63
Chinese	Cognitively Normal	CHI048	141	75	0.53
English	Cognitively Impaired	ENG001	123	62	0.50
English	Cognitively Impaired	ENG002	109	44	0.40
English	Cognitively Impaired	ENG003	66	41	0.62
English	Cognitively Impaired	ENG004	139	68	0.49
English	Cognitively Impaired	ENG005	56	40	0.71
English	Cognitively Impaired	ENG006	123	52	0.42
English	Cognitively Impaired	ENG007	75	49	0.65
English	Cognitively Impaired	ENG008	51	33	0.65
English	Cognitively Impaired	ENG009	109	63	0.58
English	Cognitively Impaired	ENG010	49	28	0.57
English	Cognitively Impaired	ENG011	105	54	0.51
English	Cognitively Impaired	ENG012	68	45	0.66
English	Cognitively Impaired	ENG013	33	24	0.73
English	Cognitively Impaired	ENG014	97	54	0.56
English	Cognitively Impaired	ENG015	70	36	0.51
English	Cognitively Impaired	ENG016	99	50	0.51
English	Cognitively Impaired	ENG017	48	31	0.65
English	Cognitively Impaired	ENG018	73	43	0.59
English	Cognitively Impaired	ENG019	85	52	0.61
English	Cognitively Impaired	ENG020	189	68	0.36
English	Cognitively Impaired	ENG021	78	47	0.60
English	Cognitively Impaired	ENG022	51	34	0.67
English	Cognitively Impaired	ENG023	177	64	0.36
English	Cognitively Impaired	ENG024	66	40	0.61
English	Cognitively Normal	ENG025	113	56	0.50
English	Cognitively Normal	ENG026	97	55	0.57
English	Cognitively Normal	ENG027	67	43	0.64
English	Cognitively Normal	ENG028	143	71	0.50
English	Cognitively Normal	ENG029	60	41	0.68
English	Cognitively Normal	ENG030	78	46	0.59
English	Cognitively Normal	ENG031	114	68	0.60
English	Cognitively Normal	ENG032	95	52	0.55

English	Cognitively Normal	ENG033	153	79	0.52
English	Cognitively Normal	ENG034	89	56	0.63
English	Cognitively Normal	ENG035	77	49	0.64
English	Cognitively Normal	ENG036	78	49	0.63
English	Cognitively Normal	ENG037	52	38	0.73
English	Cognitively Normal	ENG038	197	91	0.46
English	Cognitively Normal	ENG039	116	71	0.61
English	Cognitively Normal	ENG040	109	59	0.54
English	Cognitively Normal	ENG041	157	76	0.48
English	Cognitively Normal	ENG042	100	55	0.55
English	Cognitively Normal	ENG043	121	69	0.57
English	Cognitively Normal	ENG044	158	77	0.49
English	Cognitively Normal	ENG045	42	28	0.67
English	Cognitively Normal	ENG046	80	58	0.72
English	Cognitively Normal	ENG047	98	48	0.49
English	Cognitively Normal	ENG048	76	44	0.58