A corpus based analysis of lexical richness of Beijing Mandarin speakers: variable identification and model construction

Yanhui Zhang*

Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Shatin, Hong Kong

Abstract

This work concerns the lexical richness of Beijing Mandarin speakers measured by entropy. The data used for the study are the Beijing Mandarin Spoken Corpora, a conversational and spontaneous speech corpus of contemporary Beijing Mandarin speakers. Based on the sociovariational linguistic hypotheses and data analysis, the study attempts to identify and explain the key demographical and socioeconomic parameters that impact the entropy of each subject’s spoken texts. Both one-dimensional and multi-dimensional statistical models are proposed to quantify the relationships between the pertinent measure of lexical richness and the prominent indicative variables, including age, level of education, and profession premium. A multi-dimensional nonlinear model encompassing these findings is designed and calibrated with statistical estimation methods. Possible future directions and applications in relevant field of applied linguistic are provided.

Keywords: Lexical richness, Sociovariational analysis, Entropy, Beijing Mandarin, Corpus linguistics, Statistical modeling

1. Introduction

Chinese character is the basic linguistic unit carrying semantic contents in Chinese language. The mastery of high number of characters were once emblematic of a high degree in classic education and a core demand for elite intellectuals in traditional Chinese society where the culture was quite homogeneous. Although the contemporary society has become multidimensional and more diversified, speaking standard Mandarin with sound verbal richness still attaches symbolic expressions of power and socioeconomic status to the sayers (Dwyer, 1998). Indeed, according to Giles et al. (1981), lexical richness is one of the determining factors affecting the impression formulation and perceived professional competence in candidate assessment in the business world. Bourdieu (1977, 1991) also states that linguistic competence, in terms of lexical richness, is pursued by the general public because of the promised socioeconomic growth. It is probably even more so for Mandarin Chinese since the income disparity in urban cities and economically backward regions is huge, which has fueled the pursuance of stylish speaking of metropolitan Mandarin (Zhang, 2005).

But corpus-based quantitative analysis of lexical richness of spoken Mandarin Chinese is not easy. One of the main difficulties is the scarcity of comprehensively constructed spoken corpus. Apparently the construction of spoken Chinese corpus is more costly than written Chinese. In addition, a corpus-based investigation of any interested linguistic phenomenon often demands prudent systematic structuring and computing techniques where there could exist multiple and interrelated contributing factors. They are not only for the sake of model construction and variable detection, but more importantly, for integrating the isolated individual variables into an adaptive, dynamic, and ever-evolving system, which is fundamentally required by the modern language acquisition theories, such as the Unified Model (MacWhinney, 2007). For further discuss-
tion of challenges to construction of spoken Chinese corpus, one may refer to Huang (2000) and Yang (2013), for instance. In this work, we focus on the lexical richness of contemporary Beijing Mandarin speakers and leading contributing socioeconomic factors accounting for the lexical richness differences profiled by different sayers. Based on a spoken corpus of Beijing Mandarin recently released by Beijing Language and Culture University, we study the distributional patterns of the Chinese characters measured in average entropy as a function of several sociolinguistic variables pertaining to the speakers.

The overall objective of the current work is to construct a statistical linguistic model to explain the relationship between lexical richness of the Beijing Mandarin speakers’ discourse, calculated in terms of average entropy, and their demographical and socioeconomic variables, including age, education level and profession. The multi-dimensional model is constructed on the basis of two preliminary one-dimensional models to explain the relations between average entropy, age and education level respectively. Both one-dimensional relations are nonlinear concave curves, although they are not necessarily the strict quadratic functions as currently proposed. And the multi-dimensional model is proposed to be a second degree nonlinear surface. Technically there is potential for improvement in terms of goodness of data fitting by fine tuning the degree of the polynomial and the number of interaction terms, etc. But the models proposed in the current study have solidly captured the prominent features of the linguistic properties under review. The current model can be used to predict the lexical richness of a Beijing Mandarin speaker’s speech from a given set of sociolinguistic indicators. Or it can be used conversely, i.e., to assess an interested sociolinguistic variable pertaining to a subject, if it constitutes a working variable in the proposed optimized model, when other variables are known.

It is believed that the contributions of the current work are not only the regression models designed to explain the lexical richness of Chinese in the target corpus, but also the exemplification of applying statistical hypothesis and analytical methods to fathom linguistic variables. In particular, the analytical procedure used in this study to explain the dependence of spoken texts’ entropy on sociolinguistic variables may hint new directions and provide useful methodology reference for researches in first and second language acquisition.

The rest of the paper unfolds as follows. Section two discusses the existing measures of lexical richness, explaining why entropy is preferred over TTR even though the length of texts is fixed, and justifying why it is plausible to use Chinese characters instead of words to compute entropies. Section three introduces the Beijing Mandarin corpus used in this study, followed by some discussions of the significance of Beijing Mandarin in the Chinese lingua franca. Section four provides one-dimensional regression analysis with numerical simulations after a brief introduction of the sociolinguistic variables and their distributions in the corpus. Section five proposes a multi-dimensional model to accommodate all prominent variables and explains the underlying relationships. Finally, section six provides concluding remarks and proposals of possible future research.

2. Measures of lexical richness for Mandarin Chinese

Lexical richness of a speaker, in quantitative sense, refers to the amount of vocabulary that the speaker freely uses in discourse. It reflects the speaker’s ability and skills in maneuvering the basic units of speech. The richer lexicons used in the discourse, the higher degree of variations and sophistications perceived. Lexical richness is often used equivalently as lexical diversity in literature although lexical richness may entail a broader scope of meaning such as lexical economy as pointed by Johansson (2008), for example. This study does not differentiate the two terms, because the study’s fundamental concern is the quantifiable measurement contained in either of the two terms describing the words frequency and distributions in a discourse.

Traditionally, lexical richness is an indication applied for a number of purposes, including authorship detection (Smith and Kelly, 2002; Juola, 2006), writing quality measurement (Laufer and Nation, 1995; Mellor, 2011), and language proficiency assessment (Johansson, 2008). It is to be noted, however, that the accuracy of using lexical richness to determine the quality of writing or speaking, say, is sometimes questionable, especially when human’s judgment is, by default, ultimately trusted (Yu, 2009). This is mainly because what constitute a good writing or speaking is fairly subjective judgment in itself. Unquantifiable criteria such as attractiveness of topic, reference of examples, presentation of arguments, etc. can affect a rater’s judgment. In spite of the difficulty in defining the quality of a discourse, it is no question that lexical richness is one of the essential indicators of language proficiency (Laufer and Nation, 1995; Zareva et al., 2005; Yu, 2009).

2.1. TTR measure of lexical richness

There are a variety of measures to quantify lexical richness, among which TTR (type to token ratio) is probably the most widely known. Examples of using TTR for measuring relative lexical richness can be found in Guiraud (1960), Gross (1979), Smith and Kelly (2002), and Johansson (2008). But because TTR is sample size dependent, alternative measures have been proposed to neutralize the impact of text length. Examples of such modified measures, though more or less related to TTR, include root TTR (Guiraud, 1960), logarithm TTR (Herdan, 1960), and D (Malvern et al., 2004). Quantifying lexical richness remains an unsettled problem since it is theoretically improbable for any algebraic form to accommodate the infinite complexity of linguistic texts.

On the other hand, the imperfectness of existing measures does not hinder the current study because of the following reasons. First, although there are many measures for lexical richness, fundamentally they are all constructed on the basis of frequency of words. Second, the main drawback of the traditional measures is the dependence on text length. In another
word, many enhanced versions of lexical richness measures aim to mitigate the text length effect. This sort of rectification would be crucial for studies of a single but very long text—which is typical the case for authorship identification or genre analysis, for instance. But text length or the inconsistency between different measures is not the main concern of our study. Rather, the major challenge is the size of text samples and the high dimension nature of the embedded variables.

2.2. Entropy measure of lexical richness

While TTR is a legitimate measure to assess the variability when the text length is fixed, this study uses entropy, a measure originated from thermodynamics and information theory, to compute the lexical richness. In physics, entropy is a dimensionless quantity measuring the degree of being ordered in a collection of particles or objects in a given space. In corpus linguistics, entropy measures the variations of words and whether constituent words in a text are all equally frequent or some words are less frequent relative to others. To be specific, for a given Chinese corpus text \( T \), if \( T \) has \( n \) different characters indexed with 1, 2,...,\( n \), assuming that the probabilities for all the \( n \) characters appearing in the corpus are \( p_1, p_2,...,p_n \), then the entropy of the Chinese text is defined as

\[
\text{Entropy}(T) = - \sum_{i=1}^{n} p_i \ln(p_i)
\]

Without prior knowledge of the probability distribution, the percentage of occurrence of each character is used in practice for computing entropies. For example, let’s consider a text with 100 characters in total. When all the characters are different from one another, as an extreme case, the entropy of the text will be calculated as

\[
\text{Entropy}(T) = - \sum_{i=1}^{100} \frac{1}{100} \ln\left(\frac{1}{100}\right) = 4.6052
\]

On the contrary, when all characters are the same, the entropy will be simply calculated as

\[
\text{Entropy}(T) = - \sum_{i=1}^{100} \frac{100}{100} \ln\left(\frac{100}{100}\right) = 0
\]

As shown by the above two examples, the lower the entropy of a text, the lower its lexical diversity, where the second case reports the lowest bound of the lexical diversity a text may reach since it contains only one unique character.

Although a unified measure to perfectly fit for all aspects of inquiries is not at all possible, for the current research purpose, entropy seems to be more suitable than others. Statistically it not only provides the first mode of type token ratio, but also contains the dispersion information all across the interested text, and thus provides more complete distributional pattern about a given text. In terms of Chinese corpus, it includes not only frequency information of each individual character, but also relative frequency differences between characters. In another word, it incorporates more statistical moments than TTR or D. A tabulated combinatory example can be found in Mason (2000) for illustration of the advantage of entropy over TTR.

2.3. Entropy of Chinese text

When assessing the lexical richness of English texts, the common practice is to use the frequency of words. But for Chinese texts, the more appropriate and realistic scheme is to use the frequency of Chinese characters instead of multisyllabic words. Among other things, word segmentation of Chinese texts in itself is commonly known as a very challenging issue because there is no white space between words. “An improper segmentation may cause insurmountable problems for later prediction phases” (Peng et al., 2003). Due to the difficulties in Chinese word parsing, researchers often use Chinese characters to measure verbal richness in relevant studies, including authorship identification (Peng et al., 2003; Stamatatos, 2009; Zheng et al., 2006), and Chinese lexical richness analysis (Huang, 2000, for instance).

Technical challenge is not the sole reason for using Chinese characters instead of words. It is theoretically justifiable to use Chinese characters instead of multisyllabic words to calculate the lexical richness of Chinese texts in consideration of the following reasons. First, monosyllabic words occur most frequently in Chinese texts. Although there are debates about the monosyllabic nature of modern Chinese, it is at least generally agreed that the classical Chinese is a highly monosyllabic language, with each character carrying sufficient amount of semantic properties. Even in contemporary Chinese the monosyllabic words still account for a prominent percentage. For instance, based on a comprehensive corpus study, Huang (2000) reports that token frequency of monosyllabic Chinese words is as high as 45.83%. Moreover, it is reported in the same study that “monosyllabic nouns have an average frequency ten times bigger than the average frequency of all nouns.” and “monosyllabic verbs are more than four times as likely to be used than an average verb”. Another theoretical foundation making the use of single characters a plausible idea to quantify the lexical richness in Chinese corpus is the agreement between the occurrence of the component of the multisyllabic words and the lexical richness in Chinese texts. This is because that all disyllabic and trisyllabic words are also made up of monosyllabic characters. A substantial portion of such monosyllabic components actually also carries semantic meanings. This is even more so for quadrasyllabic words such as 阳春白雪 (yang2 chun1 bai2 xue3, white snow in early spring), and 冰清玉洁 (bing1 qing1 yu4 jie2, as pure as ice and as noble as jade).
3. Beijing Mandarin and the corpus for the study

In the current study, Beijing Mandarin is profiled as a subcategory and dialect of Mandarin Chinese, one prominent dialect throughout China’s recent history (Gordon and Grimes, 2005). The importance of Beijing Mandarin is further enhanced in recent years by non-mandatory social and economic factors. According to Bourdieu (1977, 1991), a dialect is pursued by population because the sayers of the dialect foresee favorable socioeconomic returns. Bai (1994) identifies five sociocultural factors, including the feeling of prestige and general approval, affecting people’s attitude towards learning standard Mandarin. The geoeconomic importance of Beijing metropolitan city reinforces the popularity of Beijing Mandarin, compared to other dialects.

The current study is based on a spoken corpus of Beijing Mandarin Chinese (Beijing Mandarin Spoken Corpora, BJKY) developed and maintained by the Research Institute of Linguistics, Beijing Language and Culture University. The corpus was constructed through in vivo interviews of about 380 individuals in Beijing metropolitan region. The interviewees are randomly selected with a broad range of demographical characteristics and socioeconomic dimensions. Each interview, lasted about one hour on average, covers a very large spectrum of sociocultural topics such as everyday life of Beijing, housing, education, movie and art, transportation, economics, and generational changes. All topics are easily accessible and familiar for general population.

These interviews were transcribed into texts by trained transcribers. In transcription, the original spoken Chinese features were kept with earnest efforts. All fillers, such as “a”, “en”, “zhei4ge zhei4ge”, “bu4 bu4 bu4” etc. are factually recorded. The same rubrics apply to the slip of tongues, repetitions, and broken sentences. For some Beijing local colloquialisms which are not supported by Chinese fonts and not substitutable by homophones, a blank square is used followed by Pinyin transcription of the colloquialism, e.g., “bia1 jiu4 gei3 yi2xia4zi”, in which “bia1” is an onomatopoeia featuring the sound of hand clapping. In extreme cases where the speaker’s phonological productions are actually intelligible, parentheses containing ellipsis are inserted at corresponding positions.

The online BJKY interface has capped the length of the transcribed speech discourse that a single inquiry can attain when a keyword is entered. Due to this restriction, the average length of each transcribed text in the current study is set as 500 characters or slightly less. On the other hand, the effectiveness of entropy method is still retained even though the texts are very short compared to the general notion of corpus analysis. To empirically validate such a claim, a text of 484 characters is randomly excerpted and a series of iterative computations are conducted to examine the trend of entropy as a function of text length. The following Fig. 1 is the plot of the computed entropy against the text length, where the sampled sequence of text is partitioned into 54 subtexts. The partitioning points are located at the end of, respectively, the 1st, 2nd, 3rd, 5th, 8th, 10th, 20th, ..., 470th, 480th, and 484th characters. The lengths of the first few subtexts are kept small to capture the subtle near-zero characteristics of the curve. It is evident from the plot that the entropy quickly flattens for the text containing 200 characters onward. Similar patterns are observed with different random selections of texts.

4. One-dimensional models

In the corpus under review, each subject participating in the experiments was tagged with seven variables. Two of the tags are demographical indicators, namely, age, and gender; four are socioeconomic indicators, namely, profession, ethnic, district of residence, and level of education; and one is topical indicator. To give a brief commentary on these indicators, it is noticed that district of residence is reasonably the most neutral factor with least causal-effect on the character richness in the study. This is not to claim that place of residence does not differentiate dialects at geographical scale. But in a metropolitan city like Beijing with very high internal population liquidity and homogeneity in metropolitan culture, the impact of subregional residence distribution in a relatively small radius would have little effect in shaping linguistics variations. The following table gives a summary of the rest of the variables (excluding gender and residence of district) and the description of the range of the sample space.

![Fig. 1. Plot of entropy against text length.](image-url)
For a heuristic principal factor analysis, sample inquiries from the BJKY corpus can be obtained using candidate keywords. For instance, both (xiao3 nü3 hai2) and (ya1 tou pian4 zi) mean little girl in Chinese. But the latter is often used by older generations of local Beijing residents, often attached with a negative sentiment. Consider the following discourse given by a male born in 1947.

我妈说男孩儿女孩儿啊？从厨房跑出来。我说丫头片子。

Wo3 ma1 shuo1 nan2 hair2 nü3 hair2 a? Cong2 chu2 fang2 pao3 chu1 lai2.
Wo3 shuo1 ya1 tou pian4 zi.

My mother came out of the kitchen and asked whether it is a boy or a girl. I said it is a girl.

Or consider another example where an unobtrusive corner or a remote place is referred. While an educated speaker often says 角落 (jiao3 luo4), many local people with lower education or unmindful of courteous speech often say 旮旯 (ga1 la1). The following excerpt is spoken by a nurse with a junior middle school education.

山旮旯儿里头，比较艰苦。

Shan1 ga1 lar1 li3 tou, bi3 jiao4 jian1 ku3.

(The life) in the remote mountainous region is difficult.

To give a clearer description of age and education, the key variables used in the subsequent analysis, 206 subjects are randomly sampled. Fig. 2 gives the histogram distributions of age and level of education of these subjects. Histograms for other variables can be readily obtained, but omitted here.

For each variable, the histogram and the probability distribution can be drawn. Qualitative analysis shows that all distributions of the above six variables pertaining to the subjects can be roughly modeled by normal distributions, unless one is keen to find perfect functional matches for each distribution, in which case more complicated functions such as gamma distribution will be appealed. Also, the subtle features such as the bimodal pattern shown in the age distribution will possibly need combination of more than one type of functions. After the distributions are identified in terms of specific functions, the parameters appearing in the distributions can be determined, for example, by least square regression routines. Since this is not the major goal of the current paper, the detailed procedures of this respect will be omitted.

4.1. Entropy as a function of age

To our best knowledge, there are not many previous studies concerning the age effect on language acquisition in general from a psycholinguistic perspective. The literature concerning the role of age in the dynamics of lexical richness of sayers is even scarce. There is an early literature by Sankoff and Lessard (1975), attempting to assess the impact of social stratifica-
tions on verbal richness from a sociolinguistic point of view. But they reported “a continuing enrichment of productive vocabulary with increasing age”, which is fundamentally different from both the hypothesis and analytical results of the current study. Also, theory and study from cognition of memory (Craik and Jennings, 1992) and second language acquisitions (Lasagabaster and Doiz, 2003; Munoz, 2008) generally tend to support the notion that the lexical learning trend will eventually bend down in age since the overall cognitive functions in humans will eventually decrease after certain ages, although the clear cutoff times may vary. This concave shaped lexical performance curve is verified by the data simulations using our texts samples. Taking into account the sporadic previous studies and the heuristic results from our data analysis, we propose our hypothesis as follows:

Hypothesis 1. Entropy of the spoken text of a contemporary Beijing Mandarin sayer is positively correlated with the sayer’s age before a certain threshold of age. After such a threshold of age, the associated entropy is flattened or negatively correlated with age.

To validate Hypothesis 1, we search for a model encompassing the quantitative properties characterized in the hypothesis. If such a model exists with sufficient explaining power on the data collected from the corpus under review, the hypothesis is validated. Otherwise, the hypothesis should be rejected. Linear models will not be appropriate for explaining the average entropy of the Chinese corpus in the current study because the entropy trend is monotonically increasing with age. To consolidate, this study proposes a second degree polynomial to describe the relationship between the average entropy and age:

\[ \text{Entropy} = C + a_1 \text{Age} + a_2 \text{Age}^2 \]

Using the corpus data sample explained in the previous section and run a least square regression routine in SPSS, the following plot of average entropy against age class is obtained. Here ages are grouped into seven evenly spaced classes from 25 up to 85, with a subinterval of 10 between any two adjacent classes.

A nonlinear regression procedure using SPSS yields the following estimation of the parameters for the quadratic model in ages:

- \( C = 4.4315 \)
- \( a_1 = 0.0112 \)
- \( a_2 = -0.0001 \)

The goodness of fit statistics, expressed in terms of adjusted \( R^2 \) is 0.916. And the standard error of the estimates is recorded as 0.013. The overall explaining power of the model is strong (\( F = 44.355, p = 0.002 \)).

It is noted that if the order of the polynomial is increased from two to three, i.e., third degree instead of second degree polynomial is used for the entropy functional in ages, the adjusted \( R^2 \) values can be increased to 0.960 with about equal overall significance of the modeling (\( F = 22.833, p = 0.014 \)). See the dash-dot curve in Fig. 3. However, it seems too hypothetical at this stage to introduce any polynomial with degree higher than two. It not only complicates the expression of the model, which is something model builders try to avoid, but also contravenes the concavity requirement in extended space of ages (although in the interested interval, say for ages 10–90, the concavity is still conserved).

4.2. Entropy as a function of education

The second sociolinguistic model to explore is the relationship between the average entropy and the level of education. Again, literature pertaining to this specific topic is rather limited. Intuitively, one would expect that the higher the education level of a sayer, the higher the character richness in speaking. But our data analysis departs from such a popular notion. Although the calculated entropy from the Beijing Mandarin corpus does show strong positive correlation with level of education when the education is lower or equal to university degree, further advanced degrees do not appear to further enhance the sayer’s lexical richness. By the heuristic analysis from the corpus data, and the initial trend obtained therein, we propose our second hypothesis of this study as follows:

Hypothesis 2. Entropy of the spoken text of a contemporary Beijing Mandarin sayer is positively correlated with the sayer’s education level. But advanced degrees beyond college education do not necessarily add the sayer’s spoken text entropy.

Using the same sample of data, going through a similar postulation as for the relation between entropy and age, a similar concave curve for entropy against level education is derived, where \( \text{Edu} \) denotes the level of education starting from 1 for illiterate to 7 for advanced degrees above bachelor’s degree:

\[ \text{Entropy} = D + b_1 \text{Edu} + b_2 \text{Edu}^2 \]

The validity of Hypothesis 2 depends on the calibration of the model defined by the above equation and the quality of the data fitting. Similar regression procedure yields the following estimation of the parameters for the quadratic model in education levels:
The goodness of fit statistics, expressed in terms of adjusted $R^2$, is 0.946. And the standard error of the estimates is recorded as 0.014. The overall explaining power of the model is strong ($F = 53.860, p = 0.001$). If allowing higher degree polynomials such as cubic polynomial instead of quadratic, the adjusted $R^2$ value can be increased to 0.957. Again, the slight improvement in curve fitting does not quite set off the burden to complicate the model formula. The following Fig. 4 is the numerical plot of average entropy against level of education.

One possible explanation to be proposed for explaining the negative effect of excessive advanced education on a sayers lexical richness is the superfluousness of education, which basically means that social performances do not always favor highly educated people. Sayers with advanced degrees, in discourse pertaining to topics of their specialty, will very likely exhibit high verbal richness, driven by at least the wide spectrum of their professional vocabulary and linguistic terminologies. But in daily life and usual social cultural settings in Beijing city, for instance, the lexical capacity empowered by their professional trainings will not be applicable, thus the character richness may not linearly increase with the level of education, as shown by our current corpus study. This interesting finding deserves further and deeper study.

4.3. Influence of profession and gender

Next variables to be investigated are profession and gender. Since both are categorical random variables, it is appropriate not to presume linear or nonlinear relations in the first place. Because no linear relation is assumed, correlation test does not
apply. The most appropriate test is t-test. With t-test, one can determine whether the mean average entropies for two types of profession, doctor and ticket seller, for instance, are statistically different.

To run the t-test, we first need to transform and scale the categorical values into arithmetic values. The Chinese corpus for this study has 15 profession tags after some very equivalent professions, such as high school teachers and zhongzhuang (intermediate specialty educational) teachers are grouped together. The list of these 15 professions, ranked in descending order of average entropy, is as follows: primary school teacher, doctors, jie1dao4 gan4bu (community officials), small self business owner, governmental official, university teacher, nurse, driver, middle school teacher, salesperson, policeman, unemployed, industry worker, student, and housewife. To incorporate profession into our multi-factor model to be introduced in the next section, we assign each profession with a numerical value called profession premium starting from 1 to 10, where 1 corresponds to the profession with lowest average entropy, which is the profession tagged with housewife in the corpus; and 10 the highest, which is the one tagged with primary school teacher. For those professions falling somewhere in between, we assign a profession premium using the linear interpolation between the profession premiums of housewives and primary school teachers, depending on how close these professions are relative to the two ends on the ranking list.

From a random sample consisting of the speaking texts tagged by 20 primary school teachers and 18 industry workers we run the two sample student’s t-test to test the hypothesis that the resulting average entropies of the texts from the two professions were equal. The results show the mean average entropies for the two professions were significantly different, $t = 2.224, p = 0.033$, for equal variances assumed; and $t = 2.265, p = 0.030$ for equal variances not assumed (for completeness, the Levene’s Test for Equality of Variances yields $F = 0.565, p = 0.465$). To be specific, the average entropy for the samples from school teachers is 4.7224, with SD = 0.1715; and the average entropy for the samples from industry workers is 4.6144, with SD = 0.1203. A 95% confidence interval on the difference between the two sample means is (0.0095, 0.2066) for equal variances assumed, and (0.0111, 0.2050) for equal variances not assumed.

Similarly, a two sample t-test was preformed to test the hypothesis that the resulting average entropies of the speech texts from male and female Beijing Mandarin speakers were equal. The results, however, does not reject the null hypothesis. That is, the mean entropies for the two samples, 68 tagged male and 138 tagged female, were not significantly different from each other, $t = −0.092, p = 0.927$, for equal variances assumed; and $t = −0.089, p = 0.929$ for equal variances not assumed (for completeness, the Levene’s Test for Equality of Variances yields $F = 0.851, p = 0.357$). To be specific, the average entropy for the samples of male Beijing Mandarin speakers is 4.671888, with SD = 0.1555; and the average entropy of female speakers is 4.673897, with SD = 0.1406. A 95% confidence interval on the difference between the two sample means is (−0.0450, 0.0401) for equal variances assumed, and (−0.0464, 0.0424) for equal variances not assumed.

In summary of the one factor modeling, three sociolinguistic factors, age, level of education, and profession are identified and tested as significant variables to be included in the modeling of average entropies of Beijing Mandarin speakers. Among these three pairwise one factor regressions, two of them (age to entropy, education to entropy) can be modeled by quadratic polynomials with negative leading terms, and the third one can be separately modeled by adding professional premiums.

5. Multi-dimensional model

After hypothesis tests and data calibrations for individual variables, it is possible and plausible to generalize the findings obtained in the previous one factor models into a multi-factor model encompassing all three sociolinguistic factors. Based on the pairwise one factor analysis, we would like to propose a multi-dimensional model of the following form:

$$
Ent_{\text{entropy}} = K + a_1 \text{Age} + a_2 \text{Age}^2 + b_1 \text{Edu} + b_2 \text{Edu}^2 + m \text{AgeEdu} + n \text{Pro}
$$

where the AgeEdu term models the interactions between age and level of education and Pro denotes the profession premium. Using the corpus data sample explained in the previous section and running a least square regression routine in SPSS, we obtain the following estimation of the parameters for the unified model:

$$
K = 4.122
a_1 = 0.006
a_2 = −3.2E − 05
b_1 = 0.128
b_2 = −0.011
m = −4.5E − 04
n = 0.022
$$

Indeed, the above model consolidates the age effect, the effect of education, and the influence of profession to explain the spoken lexical richness of Beijing Mandarin speakers. The goodness of fit statistics, expressed in terms of adjusted $R^2$ value, is 0.521. The overall explaining power of the model is significant ($F = 3.120, p = 0.019$). Although the $R^2$ value is not significantly high, we should keep in mind first that the $R^2$ value is not a centrally relevant measure in deciding the
validity of a highly nonlinear model. Second, there are possibly other hidden affecting variables that are not provided by the corpus. We do agree that further investigations will be required to thoroughly answer these concerns and conjectures.

6. Concluding remarks and future direction

The current work presents a corpus-based statistical modeling of the lexical richness of contemporary Beijing Mandarin speakers. Three sociolinguistic variables, namely, age, level of education, and profession, are identified as key factors explaining the lexical richness of the speakers measured in terms of entropy of the spoken discourse. In addition to one-dimensional models separately addressing one factor at a time, a unified three-dimensional model is designed and empirically tested for incorporating the overall findings. Since literature of sociovariational analysis based on large corpus is scarce, particularly for Chinese spoken corpus, our methods and results are relatively new and significant. The methodology and results of the current corpus-based study not only have sociolinguistic implications, but also provide insightful hints for applied researches in first or second language acquisitions.

There exists room for technical improvement for the multi-factor modeling and analysis. Possible aspects for improvement may include adding more interaction terms in the model, allowing fractional order factors in each term, and using exponential class of functions or other special functions instead of simple polynomials. This further investigation will be contingent on large quantity of simulations based on theories in neighborhood research such as cognitive science and memory mechanism, etc., since random trials are generally not plausible choices. Another technical point is about Chinese word parsing. Although it is sound enough at this stage, as explained in section two from linguistic theoretical point of view, to use Chinese characters instead of words in computing entropy of corpus, the equivalence of the two methods remains to be empirically verified in future when computerized Chinese word parsing becomes mature enough.

One major limitation of the current study relevant to the format of the current corpus is probably the lacking of phonological production data. The corpus we used consists of only transcribed texts of discourses of the experiment participants. Thus, interesting factors such as punctuality, tones variations etc. that may differentiate different speakers cannot be included. It would be definitely a worthwhile attempt to address these variables when audio texts of this particular corpus or other similar corpuses become available in future. Another corpus-related limitation we felt is that the corpus texts are more of one-way speech in nature (although in casual and relaxing cognitive conditions) rather than in vivo social talks. In other words, the corpus is not as conversational as it might be if directly recorded from spontaneous social exchange settings. This is probably due to the procedure of the experiment. In the current experiment, the interviewer does not seek to participate in the conversational. Rather, he or she plays a role of listener and recorder. After introducing a topic such as "please introduce your daily activities and entertainments", the interviewer pretty much lets the interviewee speak alone. A similar approach of study would be also interesting if applied to real social exchanges where the discourses are more spontaneous and the turns of speeches are changed more frequently.

The last weakness is that the sample text is not long enough. Because of the default setting of the system, each inquiry does not produce a text more than 50 characters. Similar procedure of modeling and analysis for large-length texts would be desired for comparative studies.

Appendix A. Statistics of selected samples from the corpus under study

<table>
<thead>
<tr>
<th>Ranking no.</th>
<th>Total Chinese characters</th>
<th>Lexical density (TTR)</th>
<th>Lexical richness (Entropy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>703</td>
<td>0.403877221</td>
<td>5.1442</td>
</tr>
<tr>
<td>2</td>
<td>551</td>
<td>0.475890985</td>
<td>5.112</td>
</tr>
<tr>
<td>3</td>
<td>601</td>
<td>0.240776699</td>
<td>5.0659</td>
</tr>
<tr>
<td>4</td>
<td>551</td>
<td>0.426439232</td>
<td>4.9687</td>
</tr>
<tr>
<td>5</td>
<td>601</td>
<td>0.401559454</td>
<td>4.9538</td>
</tr>
<tr>
<td>6</td>
<td>751</td>
<td>0.346875</td>
<td>4.9372</td>
</tr>
<tr>
<td>7</td>
<td>551</td>
<td>0.445859873</td>
<td>4.9348</td>
</tr>
<tr>
<td>8</td>
<td>651</td>
<td>0.379496403</td>
<td>4.9226</td>
</tr>
<tr>
<td>9</td>
<td>551</td>
<td>0.409185804</td>
<td>4.9124</td>
</tr>
<tr>
<td>10</td>
<td>551</td>
<td>0.399159664</td>
<td>4.9117</td>
</tr>
<tr>
<td>11</td>
<td>602</td>
<td>0.366729679</td>
<td>4.8856</td>
</tr>
<tr>
<td>12</td>
<td>551</td>
<td>0.413721414</td>
<td>4.8826</td>
</tr>
<tr>
<td>13</td>
<td>651</td>
<td>0.380434783</td>
<td>4.8778</td>
</tr>
<tr>
<td>14</td>
<td>551</td>
<td>0.388429752</td>
<td>4.8617</td>
</tr>
<tr>
<td>15</td>
<td>551</td>
<td>0.435416667</td>
<td>4.8602</td>
</tr>
<tr>
<td>Ranking no.</td>
<td>Total Chinese characters</td>
<td>Lexical density (TTR)</td>
<td>Lexical richness (Entropy)</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------</td>
<td>----------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>16</td>
<td>552</td>
<td>0.379876797</td>
<td>4.852</td>
</tr>
<tr>
<td>17</td>
<td>551</td>
<td>0.384615385</td>
<td>4.8499</td>
</tr>
<tr>
<td>18</td>
<td>551</td>
<td>0.392781316</td>
<td>4.8475</td>
</tr>
<tr>
<td>19</td>
<td>601</td>
<td>0.384912959</td>
<td>4.8453</td>
</tr>
<tr>
<td>20</td>
<td>651</td>
<td>0.345744681</td>
<td>4.8446</td>
</tr>
<tr>
<td>21</td>
<td>551</td>
<td>0.396624473</td>
<td>4.8409</td>
</tr>
<tr>
<td>22</td>
<td>601</td>
<td>0.361376673</td>
<td>4.8408</td>
</tr>
<tr>
<td>23</td>
<td>601</td>
<td>0.375241779</td>
<td>4.8324</td>
</tr>
<tr>
<td>24</td>
<td>593</td>
<td>0.370588235</td>
<td>4.8312</td>
</tr>
<tr>
<td>25</td>
<td>601</td>
<td>0.423228346</td>
<td>4.8284</td>
</tr>
<tr>
<td>26</td>
<td>619</td>
<td>0.362453532</td>
<td>4.825</td>
</tr>
<tr>
<td>27</td>
<td>601</td>
<td>0.375984252</td>
<td>4.823</td>
</tr>
<tr>
<td>28</td>
<td>601</td>
<td>0.404809619</td>
<td>4.821</td>
</tr>
<tr>
<td>29</td>
<td>601</td>
<td>0.35770751</td>
<td>4.8188</td>
</tr>
<tr>
<td>30</td>
<td>551</td>
<td>0.41322314</td>
<td>4.8177</td>
</tr>
</tbody>
</table>

Ranked by lexical richness measured in entropy, from high to low.

References


