Measuring Individual Differences in Word Recognition: The Role of Individual Lexical Behaviors

林欣霓　Hsin-Ni Lin

國立臺灣師範大學英語所語言組

Department of English, Linguistics Division

National Taiwan Normal University

hnlin98@std.ntnu.edu.tw

謝舒凱　Shu-Kai Hsieh

國立臺灣大學語言所

Graduate Institute of Linguistics

National Taiwan University

shukaihsieh@ntu.edu.tw

詹曉蕙　Shiao-Hui Chan

國立臺灣師範大學英語所語言組

Department of English, Linguistics Division

National Taiwan Normal University
shiaohui@ntnu.edu.tw

Abstract

This study adopts a corpus-based computational linguistic approach to measure individual differences (IDs) in visual word recognition. Word recognition has been a cardinal issue in the field of psycholinguistics. Previous studies examined the IDs by resorting to test-based or questionnaire-based measures. Those measures, however, confined the research within the scope where they can evaluate. To extend the research to approximate to IDs in real life, the present study undertakes the issue from the observations of experiment participants' daily-life lexical behaviors. Based on participants' Facebook posts, two types of personal lexical behaviors are computed, including *the frequency index of personal word usage* and *personal word frequency*. It is investigated that to what extent each of them accounts for participants' variances in Chinese word recognition. The data analyses are carried out by mixed-effects models, which can precisely estimate by-subject differences. Results showed that the effects of personal word frequency reached significance; participants responded themselves more rapidly when encountering more frequently used words. People with lower frequency indices of personal word usage had a lower accuracy rates than others, which was contrary to our prediction. Comparison and discussion of the results also reveal methodology issues that can provide noteworthy suggestions for future research on measuring personal lexical behaviors.

Keywords: individual differences, lexical behaviors, word recognition, computational linguistic approach, naturalistic data

1. Introduction

In the field of psycholinguistics, a major research interest is to investigate how people recognize written words or access the corresponding word representations stored in their mental lexicon. Psycholinguists usually undertake the investigation starting from isolated words since less factors are involved, compared to words within sentences. Therefore, research on the isolated word recognition is fundamental for understanding how lexical access takes places. In general, the term 'visual word recognition' is used to simply address the recognition of isolated written words.

Research of word recognition traditionally have concentrated on how characteristics of words *per se* (e.g. word length, word frequency, or neighborhood size) affected the procedure of recognition [1] [2] [3] [4] [5], taking the discrepancies between participants' performance as merely statistical deviation. Recently, however, there has been a growing interest in the individual differences (IDs, henceforth) of experiment participants. Results of the ID studies showed that the issue was noteworthy because personal experiences and knowledge of words (e.g. print-exposure experience [6] [7], reading skills [8], or vocabulary knowledge [9] [10] [11]) accounted for systematic variances between participants in word recognition. Even when participants were homogeneous in their educational level, their IDs sufficiently resulted in distinct performance in word recognition. Furthermore, [8] provided compelling evidence that conflicting results of regularity effects[1] in the literature were attributable to lacking control over participants' IDs of reading skills.

To date, the studies of IDs, however, have focused on test-measured or self-rated ID variables. In such approaches, the observed IDs were confined in the boundary of a test or questionnaire design, and the uniqueness of each individual in real life was neglected. In an attempt to examine the approximate real-life IDs, this research measures and analyzes IDs based on each participant's own lexical behaviors. Lexical behaviors here refer to a person's word usage and preference in his/her daily life. Intuitively, language usage reveals one's vocabulary knowledge, such as the words the person knows and how to use those words within context. Vocabulary knowledge was proved relating to word recognition [9] [10] [11]; hence, it is highly possible that IDs of lexical behaviors can explain the disparity of participants' performance in word recognition. The lexical behaviors mainly have two merits over the measure of vocabulary tests. First, people's lexical knowledge will be evaluated not by a small set of vocabularies in a given test, but by the words used by themselves. In this case, a variable's value assigned to a given participant is personalized and not confined to the scale or the total score of a test. The other merit resides in that the data of language usage can provide a deeper insight into a person's lexical knowledge, compared with a vocabulary test. If a person is able to use or produce a given word naturally (and frequently), it suggests that the word's representation has been firmly established in his/her mental lexicon.

Besides, it is worth noting that the stance we take in measuring the 'individuality' is **naturalistic** rather than **natural**, in that the lexical behaviors we describe are assumedly anchored in the interaction as naturalistic situated interactions, rather than natural ones (like using camera to collect data). A pitfall of the natural ones is that when observers and/or cameras are present those interactions are not quite what they would be in our absence.

---

[1] Regularity denotes that the extent to which the spelling-to-sound correspondence in words are invariant. The effects of regularity are that a response is made slower to less 'regular' words (e.g. pint) than to 'regular' words (e.g. name).

Therefore, the present study begins with a preliminary survey on the lexical behaviors of participants' naturalistic data on Facebook[2] Walls (Figure 1).

Figure 1. A snapshot of Facebook Wall

Our attention for lexical behaviors computed from participants' Facebook data is fastened upon *the frequency index of personal word usage* and the *personal word frequency* calculated from participants' language data. Whether the two variables are associated with participant's performance in a lexical decision task[3] will be explored respectively in two experiments. More important, as a pioneer study on lexical behaviors and word recognition, the other main objective of this research is to preliminarily explore its computational methodology.

The rest of this paper is organized as follows: Section 2 presents the procedure of our data collection, including conducting a lexical decision experiment and extracting the experiment participants' language usage data from the Facebook. Section 3 demonstrates the methods and results of two experiments, each of which computed a lexical behavior variable and further examined the relationships between participants' IDs of lexical behaviors and lexical-decision responses. Section 4 concludes this study by giving a summary and contributions of the current study. Section 5 provides potential research directions for future work.

2. Data collection

2.1 Lexical decision task

2.1.1 Participants

Sixteen Chinese native speakers (10 females and 6 males; ages ranging from 21 to 29 years old) consented to participant in the task and were offered participant fees. For the purpose of augmenting the possibility of finding individual differences (IDs) of personal lexical behaviors, the participants were recruited from diverse backgrounds. They should be right-handed, which was examined via a self-report handedness inventory [12].

2.1.2 Materials

Experiment materials included 456 Chinese words and 456 non-words. The word stimuli

---

[2] http://www.facebook.com/
[3] The lexical decision task is an extensively-used experiment of visual word recognition.

were nouns selected from the Chinese Lexicon Profile (CLP) [4], comprising 152 high-frequency, 152 mid-frequency, and 152 low-frequency words. In addition to word frequency, the number of characters, the number of senses, and the neighborhood size of words were collected from the CLP and will be treated as covariates at the stage of statistical analysis because we intended to disentangle their impacts on the lexical-decision responses.

To equalize yes and no stimuli, 456 non-words were also subsumed into the stimuli. These non-words were randomly generated by using characters of existing nouns in Chinese. Take two-character non-words for example. The procedure of random generation is illustrated in Figure 2. The first and second characters of existing nominal words were separately stored into two vectors. Next, the first and second characters of a non-word were randomly selected from the two vectors respectively and then combined altogether. If an automatically generated non-word sounded like an existing word, it would be removed from the non-word list.

The task is a within-subjects design; that is, a participant saw all of the 912 stimuli. The non-words, high-, mid-, and low-frequency words were evenly divided into four blocks. The order of four blocks was counterbalanced across 16 participants. Within a block, experimental stimuli were administered in a random order.
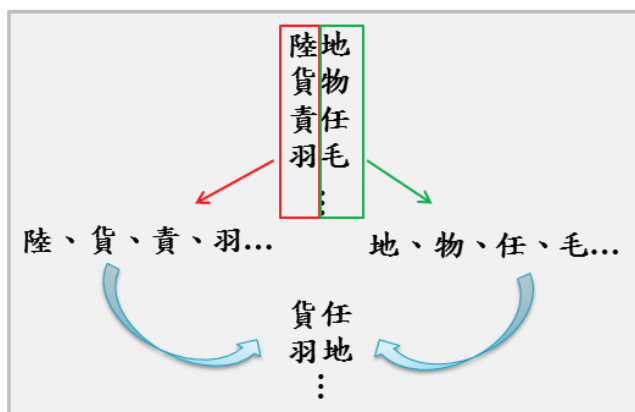

Figure 2. The procedure for random generation of two-character non-word stimuli in the visual lexical decision task

2.1.3 Procedure

Each participant was tested individually in a quiet room. The experiment was conducted and presented on a laptop via *E-prime 2.0 professional*. Participants were instructed to judge whether a visually presented stimulus was a meaningful word in Mandarin Chinese. They were required to respond as quickly as possible but without expense of accuracy, and their judgment were recorded as soon as they pressed the 'yes' or 'no' response button.

The procedure of a trial was initiated with a fixation sign (+) appearing in the center of the monitor for 1000 ms. Next, a stimulus was presented. The presentation would be terminated immediately when a participant responded. If no response was detected in 4000 ms, the given stimulus would be removed from the monitor. After termination of the stimulus

---

[4] The Chinese Lexicon Profile (CLP) is a research project launched at LOPE lab at National Taiwan University. The project purports to build up a large-scaled open lexical database platform for Chinese mono-syllabic to tri-syllabic words used in Taiwan. With its incorporation of behavioral and normative data in the long term, the CLP would allow researchers across various disciplines to explore different statistical models in search for the determinant variables that influence lexical processing tasks, as well as the training and verification of computational simulation studies. The number of Chinese words in CLP has been accumulated up to 204,922 so far.

presentation, a feedback was provided on the monitor for 750 ms, along with the participant's accumulated accuracy rate in a block.

The entire experiment included four blocks and lasted approximately one hour. Prior to the experiment, a practice session was given to familiarize participants with the experimental procedure. The session contained 4 words and 4 non-words, none of which appeared in the formal experiment.

2.2 Facebook data

The Facebook module in i-Corpus[5] was employed to gathering participants' data of language usage and preferences. The procedure is presented beneath. For the module was in its rudimentary stage of development, it was still semi-autonomous; more specifically, the initial steps in the procedure were manually accomplished.

[Step one]   Log in an APP to get a user's access token to Facebook

[Step two]   Paste the access token in the i-Corpus program

[Step three] Type in a participant's Facebook ID

[Step four]   Save the data on the participant's Facebook Wall (JSON format)

[Step five]   Extract each message in categories of post, photo, comment, and other users' walls (One message was saved as a text.) In this study, the quantification of participants' lexical behaviors is based on only the category of posts given that other categories of messages have context which is not shown in themselves.

[Step six]   Pre-process the 'post' messages by the CKIP Chinese Word Segmentation System[6]. After the segmentation, we obtained the token number in each participant' data of language usage (see Table 1).

Table 1. The token numbers in participants' Facebook posts

Subject	Chinese Token Number	Subject	Chinese Token Number
Subject01	12506	Subject09	7487
Subject02	2765	Subject10	7690
Subject03	2144	Subject11	4727
Subject04	3590	Subject12	4389
Subject05	8251	Subject13	5908
Subject06	3442	Subject14	18636
Subject07	4293	Subject15	985
Subject08	2960	Subject16	2260

---

[5] i-Corpus is an on-going NSC-granted research project conducted at the LOPE lab, National Taiwan University. This project envisions an effort to construct i-corpora so as to obtain and analyze a wide spectrum of individual linguistic and extra-linguistic data. Considering the collected material is restricted by some copyright issues, a set of iCorpus toolkits is proposed which performs the tasks of autonomous corpus data collection and exploitation (by running an integrated software package) to extract, analyze huge volumes of individual language usage data, and automatically provide an idiolect sketch with quantitative information for the benefits of linguistic and above all, sociolinguistic studies.

[6] http://ckipsvr.iis.sinica.edu.tw/

Results of the automatic segmentation were not further checked and corrected by human labor because the present study purports to explore and develop a methodology that is not labor-consuming and rather feasible for future research to compute and control the IDs of lexical behaviors. The segmented words from participants' Facebook posts were prepared for the computation of personal lexical behaviors proposed in the subsequent section.

3. Experiments on the individual differences of lexical behaviors

3.1 Experiment 1: The role of the frequency index of personal word usage in visual word recognition

Word frequency in corpora was attested to have a high negative correlation with word difficulty [13]. In this experiment, the Academia Sinica Balanced Corpus[7] frequency of a word was analogously taken as the possibility that the word is generally acquired and used by native speakers, thus being referred to for computing *the frequency index of personal word usage*. A lower frequency index of word usage indicates that a person was apt to use low-frequency words, which was preliminarily assumed to imply a person's relatively broader vocabulary knowledge. It was concerned that whether IDs of the frequency indices across participants were capable of explaining their differences in response latencies and accuracies.

3.1.1 Method

There were four steps to compute the frequency index per person, as shown in the following.

[Step one] Produce a list per participant which contained all of the words he/she used and the occurrence frequency of those words in his/her segmented Facebook data. Examples are shown in the first and second columns of Table 2.

[Step two] Gather from the CLP the corresponding word frequency in Sinica Corpus of each word on the list, as exemplified in the third column of Table 2. Note that a few words were assigned a missing value "NA" in the column since they did not appear in the Sinica Corpus. Those words, which possessed no Sinica frequencies, would be excluded from the calculation of participants' frequency indices. Given that some of them were a string that was erroneously grouped as a word by the automatic segmentation program (e.g. *zai4 wuo3 nao3* (在我腦) 'in my brain'), the exclusion enabled this experiment to filter out the data noise procured by automatic segmentation, thus diminishing the impact of segmentation errors on the calculation of individual lexical behaviors.

[Step three] Compute the frequency index of personal word usage U_j of the participant j by (1), where P_{ij} was the participant's personal frequency of the ith word, and S_i was the word's frequency in the Sinica Corpus. In this equation, U_j can be interpreted as the mean Sinica frequency of words used by the participant j on the Facebook. The lower the index was, the more rarely-seen words used by the participant were, which assumedly meant the person had broader word knowledge.

---

[7] http://db1x.sinica.edu.tw/kiwi/mkiwi/

$$U_j = \frac{\sum_{i=1}^{n} P_{ij} S_i}{\sum_{i=1}^{n} P_{ij}} \tag{1}$$

[Step four] The U_j index of each participant was put along with his/her response latencies and accuracies in the lexical decision task for analysis.

The steps of computation introduced above applied to the complete word list of each participant (called as "the Intact word list" hereafter). In addition to the list, this experiment also made the other word list for each participant to calculate another index. This word list (called as "the NV word list" hereafter) comprised only multi-character words tagged as nouns and verbs by CKIP Segmentation System and was preliminarily considered to be less affected by segmentation errors, compared with the Intact list.

Table 2. An example of a portion of one participant's word list

Word	Personal word frequency	Sinica word frequency
就	12	48749
這樣	4	7582
完全	2	3280
桔茶	1	NA
咒怨	1	NA
在我腦	1	NA

3.1.2 Results and Discussion

The data analyses were conducted by mixed-effects models in the *lme4* package of R[8] since the models can precisely estimate by-subject differences. In both the latency and accuracy analyses, experiment stimuli and participants were treated as random factors in the models. Procedure variables (i.e. block number and trial number) as well as word variables including types of word frequency, sense number, character number, and neighborhood size were taken as covariates. The inclusion of covariates was intended to disentangle their independent influences on the reaction latencies and accuracies. Provided that any covariate did not reach significance, it would be dropped out of the analysis; afterwards, the other variables would refit the models.

Ahead of the analysis of response latencies, incorrect responses (2.57%) were discarded at first. Two frequency indices of personal word usage respectively fitted mixed-effects models together with the above-mentioned random factors and covariates. Besides, note that the response latencies put into statistical analyses were log-transformed so as to reduce skewed distribution of reaction time. Inspection of the residuals of the models found notable non-normality, as shown in the upper right panel of Figure 3[9]. To improve the goodness of fit, we removed outliers with standardized residuals outside the interval (-2.5, 2.5) [14, 15], which were 2.54% of the correct-response data set in models of the Intact list and the NV list. After the removal, the models were refitted; the residuals of the refitted models are displayed in the lower right panel in the figure. As can be seen, the non-normality of the residuals was

---

[8] http://www.r-project.org/

[9] Figure 3 displays the residuals of the model fitted by the values computed from the Intact word list. The plot of residuals in the NV list model is not demonstrated because it was the same as Figure 3.

attenuated. In the final models, statistical results showed that the frequency indices from the Intact list ($p = .3638$) and NV list ($p = .4926$) both did not significantly vary with participants' response.
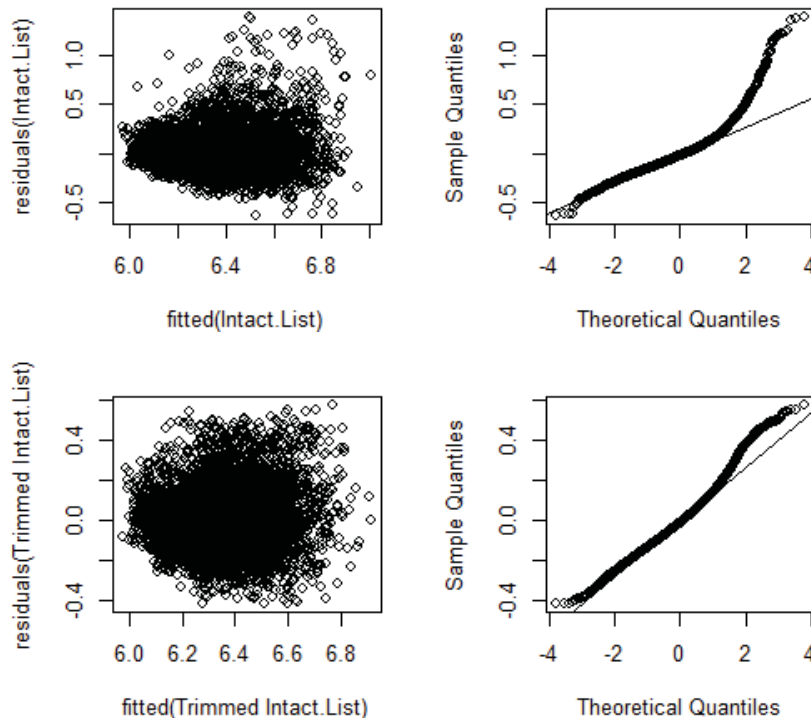

Figure 3. Residual diagnostics for the models of the Intact list before (upper panels) and after (lower panels) removal of outliers

Concerning the analysis of response accuracies, responses to all of the word stimuli in the task were taken into the analysis. Correct responses were coded as ones, and incorrect response as zeros. Seeing that the accuracy values were binomial, the analysis was carried out by the logistic mixed-effect models. Results suggested that the index computed from participants' NV lists was found to affect response accuracies ($p < .001$). Its effect on the accuracy, however, was opposite to our preliminary prediction that lower indices should suggest a person had broader lexical knowledge, thus relating to higher accuracy rates. Experimental results revealed that people with lower indices responded less accurately than those with higher indices. The counter-prediction may be ascribed to our methodology of computing the frequency index in two aspects.

The first aspect resides in that the personal indices were calculated by referring to an external lexical resource (i.e. the Academia Sinica Balance Corpus), where word frequency counts mainly came from written data rather than spoken data. When observing the calculation, we found that low-frequency words in the Sinica corpus encompassed not only rarely-used words but also words that were commonly used in daily-life conversation. Under the circumstances, a participant might receive a low frequency index from our computation because he/she utilized a number of 'low-frequency' words that are ubiquitous in spoken data, which are certainly not associated with broad lexical knowledge. This problem would become apparent when the frequency index was computed from the NV list of personal word usage. Unlike the NV list, the Intact list contained function words in addition to nouns and verbs. Function words, such as pronouns or conjunctions, are words that express grammatical

relations between sentences and other words, so their occurrence in both written and spoken data must be high. With the involvement of function words, the Intact list could relieve the computation problem which was yielded by the huge discrepancy of word frequencies between written and spoken data. This is the possible reason why our results depending on the NV list showed that people with lower frequency indices had lower response accuracies but the results relying on the Intact list did not.

The second aspect is that participants posted messages on their own Facebook Wall for diverse main purposes. Facebook is a social network designed for users to convey themselves and communicate with friends. Users can freely post any kind of messages they would like to share on their own Facebook Walls. Some users favored confiding their feelings at one moment; some preferred sharing anecdotes they experienced on a day; others often made serious comments on news and social events to evoke friends' or even the public's awareness. A skim over the Facebook data we collected could detect that the phenomena happened to users participating in this study. Accordingly, modes of the collected personal language data varied over a continuum illustrated in Figure 4. For instance, participants who were used to casually express their feelings in the data would be closer to the "informal" and "spoken" end of the continuum. A concern is raised about those who tended to take the Facebook Wall as the space to share informal messages. Even if a person has broad vocabulary knowledge and would use rarely-seen words when writing formal messages or articles, the possibility that he/she uses those words in the informal/spoken mode might decrease. Furthermore, due to the inconsistent modes across participants' Facebook data, the seriousness of the problem caused by the Sinica Corpus word frequency might vary from person to person. As mentioned above, various commonly-used spoken or informal words were shown as low-frequency words in the Sinica corpus. Those spoken vocabularies were the sources from which our computed frequency indices were distorted. Consequently, if one's Facebook posts were generally close to the informal end of the mode continuum, his/her index would be largely affected by the problem originated from the Sinica word frequency.

Figure 4. Continuum of modes in Facebook posts

According to the two forgoing aspects, our counter-hypothesis findings were predominately accredited to the Sinica word frequencies. Thereupon, it is suggested that the computation of frequency indices in future research should take a spoken corpus as the reference of general word frequencies. With respect to the concern that people with broad lexical knowledge may use informal register and extensively-used vocabularies on the Facebook, it is a reflection we had when looking at the Facebook data. The extent to which it impacted on the index computation was unsure. A future research may probe into the extent by comparing the frequency indices calculated from people's Facebook posts with those form their compositions in an academic exam. The compositions in an exam would be scored. In that case, people must write in the formal mode to show their competence as they can as possible. Via a comparison with this formal data of language usage, the influence of the

informal Facebook posts on the frequency index can be known.

3.2 Experiment 2: The role of personal word frequency in visual word recognition

This experiment investigates whether a subject's personal word frequency of a certain LDT[10] stimulus would influence his/her corresponding reaction latency. It was preliminarily hypothesized that if he/she used a word more frequently than other words, the response to the word would be more rapid. Besides, as shown in Table 1, each participant's data differ in length; to render frequency counts across the data sets comparable, two kinds of normalization were conducted. A comparison on the effectiveness of the normalization methods is also provided in the discussion on experiment results.

3.2.1 Method

The *personal word frequency* referred to the relative degrees to which a given LDT occurred in one's Facebook posts. Steps for its calculation are as follows:

[Step one] All of 16 participants' Facebook data were joined altogether into a file at first. If an LDT word stimulus appeared at least once in the file, it was chosen to be examined in this experiment. In total, there were 218 LDT stimuli conforming to the criterion, thus taken as the stimuli in this experiment.

[Step two] Personal word frequencies of the 218 stimuli were automatically counted.

[Step three] Two distinct methods were utilized to normalize the frequency counts. The first method was to divide the each subject's word frequencies by his/her own summed token numbers (see (2)). In the equation, F_{ij} was the participant j's frequency count of the ith word; the i was limited between 1 to 218 since only 218 words were selected as stimuli in this experiment. However, note that the i in the denominator was not limited within the range, but by n instead. The n was the number of word types in a participant's Facebook data. In other words, the denominator added up word frequencies of all word types, thus representing the participant's total token number. Consequently, the output of the equation, R_{ij}, was the participant j's frequency ratio of the ith stimulus.

$$R_{ij} = \frac{F_{ij}}{\sum_{i=1}^{n} F_{ij}} \qquad (2)$$

A potential problem of (2) was that the normalized figures were affected by each participant's token number. The token number was calculated according to the results of automatic segmentation, so it certainly would be contaminated by segmentation errors. Therefore, the other approach (i.e. the z-score approach) was also adopted. Like the previous equation, F_{ij} in (3) was the participant j's frequency count of the ith word. $\overline{F_i}$ was the mean of the participant's 218 word frequency counts, and S_{F_i} was the standard deviation of those frequency counts.

$$Z_{ij} = \frac{F_{ij} - \overline{F_i}}{S_{F_i}} \qquad (3)$$

---

[10] LDT refers to the lexical decision task in this paper.

[Step four] The two types of personal word frequency were respectively put along with his/her response latencies in the lexical decision task for analysis.[11]

3.2.2 Results and Discussion

Response errors in the lexical decision task (approximately 0.06% of the data set) were first screened. Two types of normalized personal word frequencies (i.e. ratio and z-score) were analyzed by mixed-effects models. Like the analysis in Experiment 1, in both models, two random factors and six covariates were also included. Random factors encompassed experiment stimuli and participants. Covariates were procedure variables (i.e. block number and trial number) and word variables (i.e. types of word frequency, sense number, character number, and neighborhood size). The covariates were subsumed in order to avoid mis-attributing the variances caused by procedure and word variables to the effect of personal word frequency. If there was any covariate not reaching significance, which meant it statistically did not affect the lexical-decision responses, it would be removed from the analysis and the other variables refitted the mixed models.

The residuals of the two models, however, showed marked non-normality, especially at the end of long response latencies (see the upper right panel in Figure 5)[12]. To attenuate the unfitness, outliers with standardized residuals outside the interval (-2.5, 2.5) were removed. The removed data in both the ratio and z-score models were 2.48% of the data set. After trimming the outliers, we refitted the models. The residuals in the trimmed models were close to normality, as shown in the lower right panel of Figure 5.

Statistical results showed that personal word frequency significantly accounted for response latencies in both the analyses of frequency ratio ($p < .001$) and z-score ($p < .05$). The estimates of them were negative, which are visualized in Figure 6. According to the figures, the negative estimates indicated that participants responded faster to stimuli with higher personal word frequencies. The experimental results revealed that IDs of frequencies of stimuli could explain individual variances between participants in lexical decision.

Words that frequency occurred in one's Facebook data revealed the things or issues he/she paid closer attention, the words he/she got accustomed to use but was unaware of, or his/her daily-life surroundings. Therefore, the effect of personal word frequencies in this experiment was considered to result from people's conscious or subconscious familiarity with words or concepts. The familiarity with word form and meaning facilitated the access to corresponding underlying lexical representations in the participants' mental lexicon.

Another discussion brought up in this experiment is a methodological issue of computing personal lexical behaviors. Among two types of normalization of personal word frequency counts, the ratio method was assumed to be possibly problematic since segmentation errors were involved, and the z-score method was hypothesized to be a better one. Nevertheless, the analyses of word frequency ratio and z-score both reached significance. This indicated that normalizing frequency counts by the token number in each personal corpus is feasible even though there are segmentation errors and noise among the tokens. Evidence can be found when we compare each participant's total token number, which includes segmentation errors, with his token number summed from the 218 stimuli in Experiment 2, which includes no errors. The two categories of token numbers are highly correlated ($r = .95$). The correlation suggests that although segmentation errors make the total token numbers of Facebook data imprecise and inaccurate, the numbers still generally reflect

---

[11] Unlike Experiment 1, the response accuracies were not analyzed in this experiment. It was because the accuracy of the 218 stimuli here was extremely high (99.4%).

[12] Figure 5 is the residuals of the model fitted by the personal word frequency ratios. The residuals of the z-score model are the same as those of the ratio model, so its residual plot is not given here.

the comparative differences between participants' genuine token numbers.
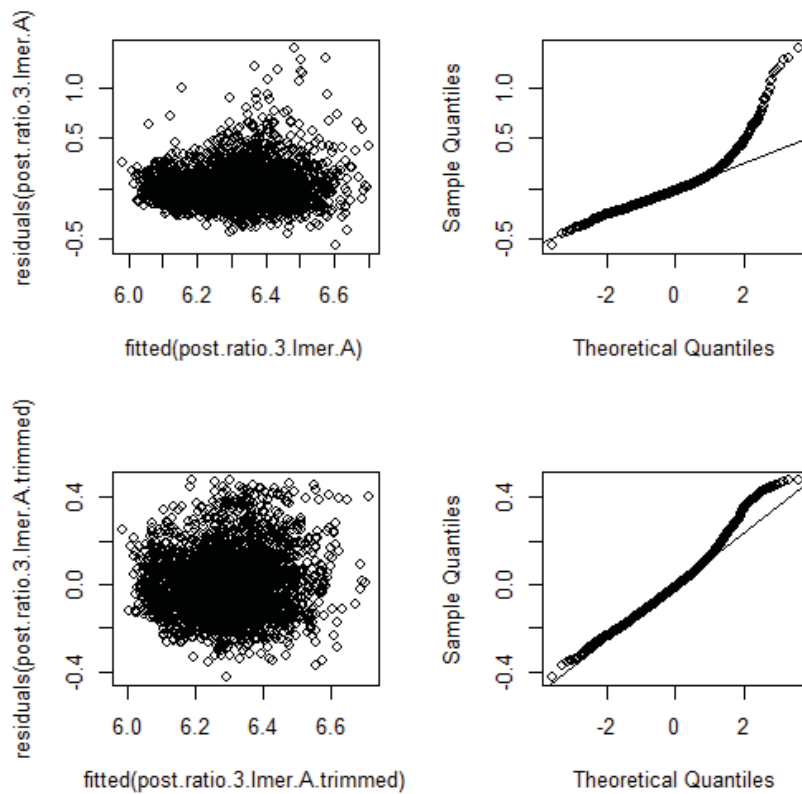

Figure 5. Residual diagnostics for the model of personal word frequency ratios before (upper panels) and after (lower panels) removal of outliers
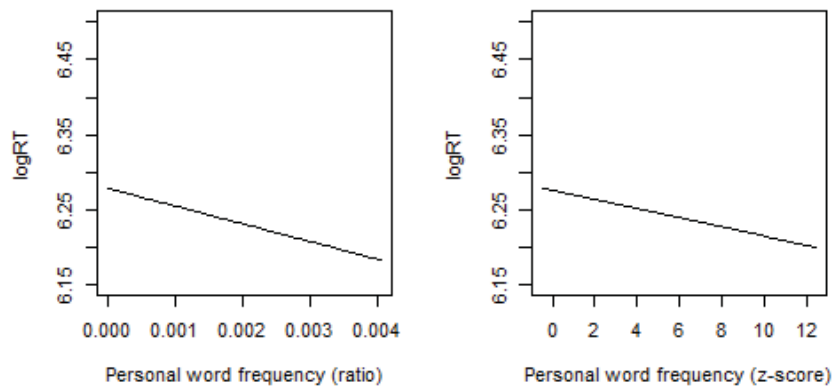

Figure 6. Partial effects of personal word frequency (ratio and z-score) in the analysis of Experiment 2

4. Conclusion

By integrating the approach of computational linguistics into a psycholinguistic experiment, the current study sheds a new light on methods of capturing the nature of IDs in word recognition. The interdisciplinary effort testified that the quantified personal lexical

behaviors were associated with word recognition, thus uncovering a territory to be explored. One promising prospect of this study is that as the methodology of measuring lexical behaviors grows mature in the future, the readily available data of language usage, like Facebook posts, can function as convenient and valid resources for researchers to control the participant factors.

Furthermore, through the comparison of experimental results, the present study made a preliminary exploration on the methodology of measuring lexical behaviors and suggests the relatively appropriate methods. The counter-prediction finding in the *frequency index* experiment was possibly attributed to that the Sinica Corpus mainly consists of written data; therefore, it is suggested that similar experiments in future research resort to the frequency counts in a spoken corpus. Additionally, according to our examination, a person's total token number is feasible for normalizing his/her frequency counts even though word segmentation errors were contained within the tokens. Finally, when naturalistic data like the Facebook posts are utilized for the measurement, it is recommend basing the computation on personal preference or pattern of lexical usage (e.g. Experiment 2), instead of on every single word in one's language usage data (Experiment 1).

5. Future Work

The present study examines word recognition by only concentrating on the lexical decision task. To obtain a clearer picture of the IDs in recognition, the future work can collect converging evidence from other types of extensively-used tasks, such as the naming task [16, 17]. Besides, this preliminary research recruited 16 participants. It is expected that when the number of participants increases in future research, it might give us other or deeper insight into the issue of individual differences (IDs). Moreover, in the Chinese Lexicon Profile (CLP) corpus mentioned in Section 2.1.2, there provides a great number of characteristics of words *per se*. Researchers may try to compute and explore individual lexical behaviors from the available characteristics, aside from the word frequency which is utilized in this study. In the respect of personal language usage data, we are constructing i-Corpus, which will comprise individualized corpora. A corpus per person will include various types of his/her language usage data, which can be looked into in the future so as to uncover multiple facets of personal language usage.

References

[1]   S. Andrews, "Frequency and neighborhood effects on lexical access: Activation or search?," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* vol. 15, pp. 802-814, 1989.

[2]   K. I. Forster and S. M. Chambers, "Lexical access and naming time," *Journal of Verbal Learning and Verbal Behavior,* vol. 12, pp. 627-635, 1973.

[3]   J. Grainger, "Word frequency and neighborhood frequency effects in lexical decision and naming," *Journal of Memory and Language,* vol. 29, pp. 228-244, 1990.

[4]   B. New*, et al.*, "Reexamining word length effects in visual word recognition: New evidence from the English Lexicon Project," *Psychonomic Bulletin & Review,* vol. 13, pp. 45-52, 2006.

[5]   C. P. Whaley, "Word—nonword classification time," *Journal of Verbal Learning and*

Verbal Behavior, vol. 17, pp. 143-154, 1978.

[6]   D. Chateau and D. Jared, "Exposure to print and word recognition processes," *Memory & Cognition,* vol. 28, pp. 143-153, 2000.

[7]   C. Sears*, et al.*, "Is there an effect of print exposure on the word frequency effect and the neighborhood size effect?," *Journal of Psycholinguistic Research,* vol. 37, pp. 269-291, 2008.

[8]   S. J. Unsworth and P. M. Pexman, "The impact of reader skill on phonological processing in visual word recognition," *Quarterly Journal of Experimental Psychology,* vol. 56A, pp. 63-81, 2003.

[9]   M. J. Lewellen*, et al.*, "Lexical familiarity and processing efficiency: Individual differences in naming, lexical decision, and semantic categorization," *Journal of Experimental Psychology: General,* vol. 122, pp. 316-330, 1993.

[10]   L. Katz*, et al.*, "What lexical decision and naming tell us about reading," *Reading and Writing,* in press.

[11]   M. J. Yap*, et al.*, "Individual differences in visual word recognition: Insights from the English Lexicon Project," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 38, pp. 53-79, 2012.

[12]   R. C. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory," *Neuropsychologia,* vol. 9, pp. 97-113, 1971.

[13]   H. M. Breland, "Word frequency and word difficulty: A comparison of counts in four corpora," *Psychological Science,* vol. 7, pp. 96-99, 1996.

[14]   M. J. Crawley, *Statistical computing: An introdution to data analysis using S-plus*. Chichester: Wiley, 2002.

[15]   R. H. Baayen, *Analyzing linguistic data : A practical introduction to statistics using R*. Cambridge: Cambridge University Press, 2008.

[16]   D. A. Balota and J. I. Chumbley, "Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 10, pp. 340-357, 1984.

[17]   D. A. Balota and J. I. Chumbley, "The locus of word frequency effects in the pronunciation task: Lexical access and/or production?," *Journal of Memory and Language,* vol. 24, pp. 89-106, 1985.