Dispersion Measures as Predictors of Lexical Decision Time, Word Familiarity, and Lexical Complexity

Adam Nohejl Taro Watanabe
Nara Institute of Science and Technology
{nohejl.adam.mt3, taro}@is.naist.jp

Abstract

Various measures of dispersion have been proposed to paint a fuller picture of a word's distribution in a corpus, but only little has been done to validate them externally. We evaluate a wide range of dispersion measures as predictors of lexical decision time, word familiarity, and lexical complexity in five diverse languages. We find that the logarithm of range is not only a better predictor than log-frequency across all tasks and languages, but that it is also the most powerful additional variable to log-frequency, consistently outperforming the more complex dispersion measures. We discuss the effects of corpus part granularity and logarithmic transformation, shedding light on contradictory results of previous studies.

1 Introduction

Measures of dispersion have been proposed in corpus linguistics to complement frequency, a measure of central tendency. While a word's frequency tells us how common the word is in the whole corpus, its dispersion tells us how evenly it is spread. For instance, the words *very* and *yeah*, or *came* and *data* may have similar overall frequencies, but *yeah* and *data* would likely have lower dispersions, as they are specific to a certain register or domain.

The conceptually simplest dispersion measure is the range: the number of corpus parts in which a word occurs. The parts may be of different granularity and function, e.g. individual texts, authors, domains, or registers. In the TUBELEX corpus [1], which is based on YouTube videos, our example words have the following frequencies (in thousands of occurrences) and ranges (in thousands of YouTube channels): *very*: 332 and 35; *yeah*: 333 and 19; *came*: 64 and 17; *data*: 64 and 7, confirming our expectations.

The number of texts in which a word appears was used to organize pedagogical word lists as early as in 1920 by Keniston [2] and mentioned as "range" by Thorndike (2021) [3]. Gries (2008) [4] lists thirteen more advanced dispersion measures that have been proposed over the decades, often theoretically motivated or considering intuitive interpretability. What is critically missing, as Gries [4] also argues, is external validation.

We aim to bridge this gap between theoretical corpus dispersion research, psycholinguistics, and NLP applications, with a comprehensive evaluation of dispersion measures on five languages, three tasks, and three levels of corpus part granularity. Two of the tasks predict psycholinguistic data, lexical decision time (LDT) and word familiarity, and one is an NLP task, lexical complexity prediction.

2 Related Research

Adelman et al. (2006) [5] evaluated log-range on English word naming and LDT, concluding that log-range[1] is a better predictor than log-frequency.

Brysbaert and New (2009) [8] replicated the results of Adelman et al. on the SUBTLEX-US subtitle corpus. Most later studies on film subtitles reached similar conclusions [9, 10, 11, 12, 13], but a few of them did not find a statistical significant difference on individual datasets [14, 15, 16]. All of these studies investigated only log-range across subtitle files (typically thousands of files corresponding to films or show episodes) as a dispersion measure, and evaluated it on LDT or word naming times, essentially replicating Brysbaert and New's study [8] on other languages.

Gries (2010) [17] evaluated multiple dispersion measures on English word naming and LDT data. The study did not reach a conclusive result on the two datasets.

Gries (2021) [18] experimented with log-frequency,

[1] Adelman et al. call range "contextual diversity", avoiding the terms "dispersion" and "range", which is arguably confusing [6, 7], but does not detract anything from the practical value of their study.

word length, and multiple dispersion measures as features for a random forest model of English auditory LDT.

3 Examined Measures

As we have noted, both frequency and range are commonly log-transformed to achieve better correlation with psycholinguistic variables. Since our evaluation will include words not present in the corpus, which would result in undefined values ($\log 0$), we take the following steps to examine the log-transformation of all examined measures and frequency: Using $n =$ number of corpus parts, we apply smoothing in the form $(dn + 1)/(n + 1)$ to each dispersion measure d if we log-transform it. Therefore, with a slight abuse of notation, we always use "$\log d$" in the following text to refer to $\log((dn + 1)/(n + 1))$. For (log-)frequency, we always use Laplace-smoothed frequency [19].

We use all measures in forms appropriate for unequally sized corpus parts and normalized to the range $[0, 1]$, adapting Gini index and Gries's DP to $d = 1 - d^*$ from their original formulas d^*, so that high values indicate high dispersion. We use the following variables, given a word w: n is the number of corpus parts; v_i is the number of occurrences of w in part i; k_i is the number of tokens in part i; $s_i = k_i / \sum \mathbf{k}$ is the proportion of part i; $r_i = v_i / \sum \mathbf{v}$ is the proportion of occurrences of w in part i; $p_i = v_i / k_i$ is the relative frequency of w in part i, i.e. frequency normalized per part; $q_i = p_i / \sum \mathbf{p}$ is the frequency normalized per part and per word. For each variable x_i indexed by corpus part, we understand $\mathbf{x} = (x_i)_{i=1}^n$ as the corresponding vector with sum $\sum \mathbf{x} = \sum_{i=1}^n x_i$, mean $\mu_\mathbf{x} = \sum \mathbf{x}/n$, and standard deviation $\sigma_\mathbf{x} = \sqrt{\sum_{i=1}^n (x_i - \mu_\mathbf{x})^2 / n}$.

We examine the following dispersion measures:

$$\text{Range } R = \frac{\sum_{i=1}^n [v_i > 0]}{n} \tag{1}$$

$$\text{Gini index } G = 1 - \frac{\sum_{i=1}^n \sum_{j=1}^n \left| q_i - q_j \right|}{2n} \tag{2}$$

$$\text{Juilland's [20] } D = 1 - \frac{\sigma_\mathbf{p}}{\mu_\mathbf{p}\sqrt{n-1}} \tag{3}$$

$$\text{Lyne's [21] } D_3 = 1 - \frac{\sum_{i=1}^n (r_i - s_i)^2}{4} \tag{4}$$

$$\text{Gries's [4] DP} = 1 - \frac{\sum_{i=1}^n |r_i - s_i|}{2} \tag{5}$$

$$\text{Rosengren's [22] } S = \frac{\left(\sum_{i=1}^n \sqrt{q_i}\right)^2}{n} \tag{6}$$

$$\text{Carroll's [23] } D_2 = \frac{-\sum_{i=1}^n q_i \log q_i}{\log n} \tag{7}$$

Regardless of the formulas above, we define each measure as 0 for words missing from the corpus.[2]

Gini index is the discrete variant of the well-know index of inequality. It was proposed as a dispersion measure independently by Murayama et al. (2018) [24] (as Word GINI $= -\log G$) and Burch et al. (2017) [25] (as D_A, later adjusted to unequally sized parts, with a slightly different normalization from our G [26]). We investigate G and $\log G$ using the formula given by Glasser (1962) [27], which reduces computation time to $O(n \log n)$.[3]

As far as we can tell, "distributional consistency" proposed by Zhang et al. (2004) [28], is simply equal to Rosengren's S (6).[4]

Finally, we observe that inverse document frequency (idf) and variation coefficient (vc) can be expressed as linear functions of log-range (1) and Juilland's D (3), respectively, and therefore do not need to be examined separately in terms of linear correlation:

$$\text{idf} = \log \frac{1}{R} s = -\log R \tag{8}$$

$$\text{vc} = \frac{\sigma_\mathbf{p}}{\mu_\mathbf{p}} = \sqrt{n-1}\,(1 - D) \tag{9}$$

4 Evaluation

We evaluate the measures on TUBELEX [1], a large YouTube subtitle corpus for English, Chinese, Spanish, Indonesian, and Japanese. Word frequency in TUBELEX was already demonstrated to achieve correlation with psycholinguistic variables on par with or superior to film subtitle corpora [1]. TUBELEX also provides three levels of linguistically valid corpus parts: videos, channels, and categories (tens of thousands, thousands, and 15 parts respectively). We use TUBELEX in its default tokenization.

For evaluation, we use the same datasets for LDT (3 languages), word familiarity (5 languages), and lexical complexity (3 languages) as were used for extrinsic evaluation of TUBELEX log-frequency by Nohejl et al. (2024) [1]. Word familiarity and lexical complexity differ from the commonly employed LDT or word naming tasks by being

[2] This is in line with the formulas for range, S, and D_2, which would give 0 for a zero frequency word. D, D_3, and DP would be undefined, and Gini index would be 1.

[3] We also use sparse arrays to represent $\mathbf{q}$, resulting in $O(m \log m)$ time, where m is the number of non-zero elements of $\mathbf{q}$. The sparseness of frequency vectors $\mathbf{q}$ grows with the number of corpus parts, keeping computation time reasonable.

[4] This seems to have escaped the attention of Zhang et al. [28] and Gries. Gries only noted that it gives the same numerical result in an example scenario [4] and appears similar in cluster analysis [17].

Table 1 Mean improvement in R_a^2 of the log-transformed measure over the non-log-transformed (number of datasets of total 11 with positive improvement, if any, in parentheses). Cases where logarithm improves R_a^2 by at least 0.001 are printed in bold.

(a) Dispersion measures as single predictors.

Dispersion Measure d	$\overline{\Delta R_a^2}$ of $\log d$ vs. d (#Datasets: $\Delta R_a^2 > 0$)		
	Videos	Channels	Categories
Range	**0.356** (11)	**0.329** (11)	−0.060 (1)
Gini Index	**0.361** (11)	**0.340** (11)	−0.000 (5)
Juilland's D	−0.230	−0.213	−0.202
Gries's DP	−0.095	−0.102	−0.169
Rosengren's S	**0.362** (11)	**0.341** (11)	−0.273
Carroll's D_2	−0.218	−0.176 (1)	−0.241
Lyne's D_3	−0.112	−0.089 (1)	−0.043
Frequency (for comparison): **0.389** (11)			

(b) Two predictors: dispersion measure and log-frequency.

Dispersion Measure d	$\overline{\Delta R_a^2}$ of $\log d$ vs. d (#Datasets: $\Delta R_a^2 > 0$)		
	Videos	Channels	Categories
Range	**0.016** (8)	**0.023** (10)	−0.010 (3)
Gini Index	**0.003** (4)	**0.002** (5)	**0.002** (6)
Juilland's D	−0.006 (3)	−0.006 (3)	−0.010 (1)
Gries's DP	−0.002 (3)	0.000 (7)	−0.005 (3)
Rosengren's S	**0.009** (5)	**0.009** (6)	−0.018 (1)
Carroll's D_2	−0.010 (5)	−0.006 (4)	−0.014 (1)
Lyne's D_3	−0.002 (3)	−0.004 (2)	−0.001 (4)

based on subjective ratings as opposed to reaction time, while the lexical complexity used in this case differs from the other data by being rated by non-native speakers or a mix of natives and non-natives.

We evaluate the dispersions in two scenarios: as single predictors, and as one of two predictors, the other one being log-frequency. In both cases, we measure adjusted R^2 [29]:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{10}$$

where n is the number of examples (dataset size) and p is the number of variables (1 or 2). We compute R^2 (coefficient of determination) for linear least squares (multiple) regression fitted to the whole dataset, which allows us to interpret it as measure of (multiple) correlation strength.[5)]

We predict mean LDT from three studies: the English Lexicon Project [30], restricted to lower-case words following the approach of Brysbaert and New [8]; the MELD-SCH database [31] of simplified Chinese words; and SPALEX [32] for Spanish. For English and Chinese, we use the published mean LDT. SPALEX only provides raw participant data, which we process by removing times out of the range [200 ms, 2000 ms] [32], and computing the means. We predict mean word familiarity from five databases: Chinese familiarity ratings [33], English MRC lexical database [34, 35], Indonesian lexical norms [36], Japanese word familiarity ratings for reception [37], and Spanish lexical norms [38]. Lastly, we predict lexical complexity for English, Spanish and Japanese using the evaluation sets of the MultiLS dataset [39]. In total, we are evaluating on 11 datasets (task-language combinations).

4.1 To Log or Not to Log

In Table 1, we compare each dispersion measure with its log-transformed version. Perhaps surprisingly, which one is a better predictor does not depend solely on the measure, but also on corpus part granularity, and whether the measure is used as a single predictor or with log-frequency.

When used as single predictors (Table 1a), the logarithmic transformation benefits range, Gini index, and Rosengren's S, resulting in stronger correlations on all 11 datasets – but only if videos and channels are used as parts. For all three measures, the difference between using and not using log-transformation is critical (0.340 to 0.362), comparable to that between log-frequency and frequency (0.389).

When dispersion measures are employed along with log-frequency (Table 1b), applying logarithm is moderately beneficial for the same measures as above and for Gini index for categories, but the improvements are not robust across datasets.

We will report and discuss each measure with logarithm applied or not applied according to these results.

4.2 Results

As shown in Figure 1 (solid bars), the only dispersion measure robustly stronger than log-frequency as predictors of LDT, word familiarity, and lexical complexity is log-range for channels and videos. Although it does not come near in correlation strength, Gries's DP is worth noting as the only measure performing the best with categories (the coarsest part granularity).

When used along with log-frequency, the following measures result in particularly robust improvements (in decreasing order): log-range for channels, log-range for videos, range for categories, and Rosengren's S for cate-

[5)] Using R^2 instead of Pearson's (multiple) correlation coefficient r (R) allows us to ignore the different polarity of the tasks (rare words have low familiarity but high complexity). The adjustment is appropriate for comparing different numbers of independent variables.

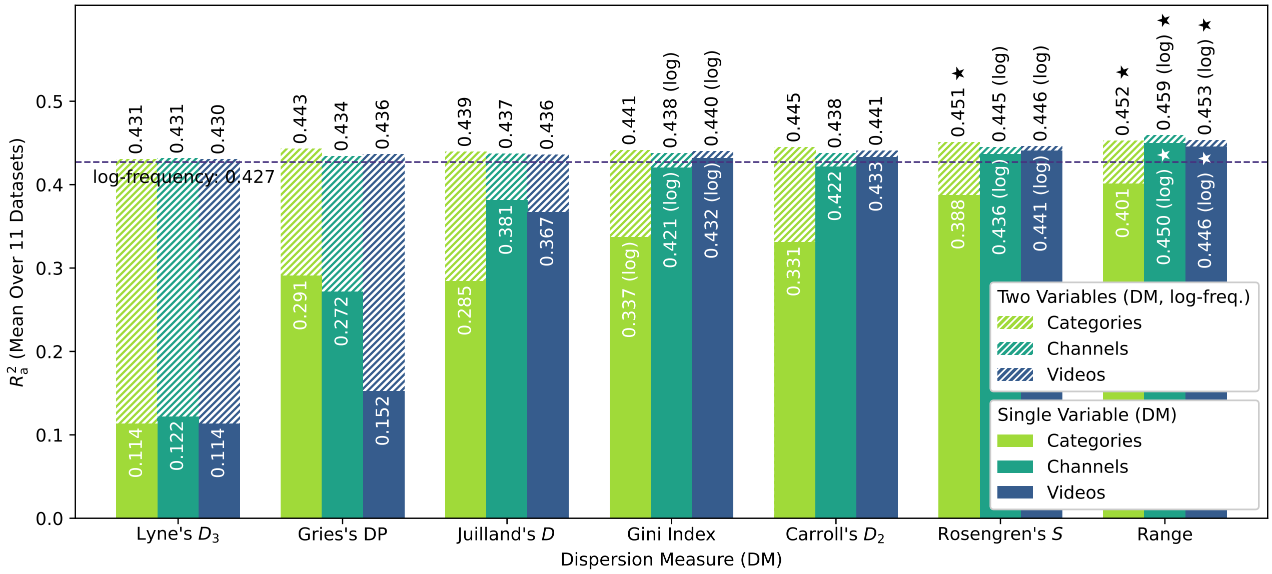

Figure 1 Mean R^2_a computed over 11 datasets for each dispersion measure, part granularity, and prediction with/without log-frequency as a second variable, where "(log)" indicates log-transformed measures. Stars indicate robust predictors, namely: $\boxed{\star}$ single predictors that were not significantly ($p < 0.001$) worse than log-frequency for any dataset, and ★ predictors that, when used with log-frequency, improved the prediction by $\Delta R^2_a \geq 0.01$ for at least 8 of 11 datasets.

gories, as shown in Figure 1 (hatched bars).

5 Discussion

We extended the previous results of Adelman et al. (2006) [5] and film subtitle studies (e.g. [8]), which showed that log-range predicts LDT better than log-frequency, to word familiarity and lexical complexity prediction. More importantly, we found that the viability of range as a single predictor depends on (1) a fine corpus part granularity, i.e. channels or videos in the case of TUBELEX, and (2) the log-transformation. This explains the low correlation achieved using only non-log-transformed range in some studies, e.g. Baayen (2010) [40]. When dispersion is used along with log-frequency, log-range for videos and channels (fine parts) are still the best choices, followed by range and Rosenberg's S, both non-log-transformed and based on categories (coarse parts).

These finding offer a guideline for choosing dispersion measures as model variables, based on the corpus parts are available. The previous studies that we know of have not compared multiple part granularities of a single corpus.

Besides three levels of granularity, our evaluation encompassed 11 datasets (task-language combinations). As the results generally agreed across datasets, we have not reported them individually. For instance, the robust single predictors (marked $\boxed{\star}$ in Figure 1) were significantly better than log-frequency on most datasets and not significantly different on two to three of them. We focused on the general, not the insignificant exceptions. This highlights the importance of evaluation on multiple datasets and puts into perspective the insignificant differences between log-range and log-frequencies on individual datasets reported in a few previous studies [14, 15, 16].

We believe that linear regression, which we used for analysis, gives more widely applicable and interpretable results than rank correlation (Gries, 2010 [17])[6] or random forests (Gries, 2021 [18])[7]. We hope that future investigations of what we have called "exceptions" or less "widely applicable" bring deeper insights into specific use cases and interactions with different data and granularities.

Our results are immediately applicable to NLP tasks that have relied on frequencies for modeling words perceived as common or simple, such as language learning applications or lexical simplification. Range data for TUBELEX, which we have used, is readily available as `channels` and `videos` in its word lists, and for most SUBTLEX language mutations as "contextual diversity" (CD), all based on corpus parts of comparable granularity.

[6] Rank correlation is an appropriate evaluation method for applications that require only ranking, but it obscures the different "shapes" of dispersion metrics.

[7] With enough training data a random forest may be more fitting for a practical application, but caution is needed when using it as an evaluation tool. The experiment in [18] used the full data for both training and testing. Moreover, the features optimal for a random forest and large training data may not perform well in other scenarios.

Acknowledgments

We are grateful to an anonymous reviewer of [1] for encouraging us to explore other dispersion metrics, given the curious result achieved by Gini index in our evaluation.

References

[1] Adam Nohejl, Frederikus Hudi, Eunike Andriani Kardinata, Shintaro Ozaki, Maria Angelica Riera Machin, Hongyu Sun, Justin Vasselli, and Taro Watanabe. Beyond Film Subtitles: Is YouTube the Best Approximation of Spoken Vocabulary? *ArXiv preprint*, Vol. arXiv:2410.03240v1 [cs], , October 2024.

[2] Hayward Keniston. Common Words in Spanish. *Hispania*, Vol. 3, No. 2, pp. 85–96, 1920.

[3] Edward Lee Thorndike. *The Teacher's Word Book*. Teachers College, Columbia University, 1921.

[4] Stefan Th Gries. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, Vol. 13, No. 4, pp. 403–437, January 2008.

[5] James S. Adelman, Gordon D.A. Brown, and José F. Quesada. Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, Vol. 17, No. 9, pp. 814–823, September 2006.

[6] Geoff Hollis. Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language*, Vol. 114, p. 104146, October 2020.

[7] Stefan Th. Gries. Analyzing Dispersion. In Magali Paquot and Stefan Th. Gries, editors, *A Practical Handbook of Corpus Linguistics*, pp. 99–118. Springer International Publishing, Cham, 2020.

[8] Marc Brysbaert and Boris New. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, Vol. 41, No. 4, pp. 977–990, November 2009.

[9] Qing Cai and Marc Brysbaert. SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles. *PLoS ONE*, Vol. 5, No. 6, p. e10729, June 2010.

[10] Maria Dimitropoulou, Jon Andoni Duñabeitia, Alberto Avilés, José Corral, and Manuel Carreiras. Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in psychology*, Vol. 1, p. 218, 2010.

[11] Emmanuel Keuleers, Marc Brysbaert, and Boris New. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, Vol. 42, No. 3, pp. 643–650, August 2010.

[12] Roger Boada, Marc Guasch, Juan Haro, Josep Demestre, and Pilar Ferré. SUBTLEX-CAT: Subtitle word frequencies and contextual diversity for Catalan. *Behavior Research Methods*, Vol. 52, No. 1, pp. 360–375, February 2020.

[13] Hien Pham, Benjamin V. Tucker, and R. Harald Baayen. Constructing two Vietnamese corpora and building a lexical database. *Language Resources and Evaluation*, Vol. 53, No. 3, pp. 465–498, September 2019.

[14] Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, Vol. 67, No. 6, pp. 1176–1190, 2014.

[15] Pawe l Mandera, Emmanuel Keuleers, Zofia Wodniecka, and Marc Brysbaert. Subtlex-pl: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, Vol. 47, No. 2, pp. 471–483, June 2015.

[16] Walter JB van Heuven, Joshua S Payne, and Manon W Jones. SUBTLEX-CY: A new word frequency database for Welsh. *Quarterly Journal of Experimental Psychology*, pp. 1052–1067, August 2023.

[17] Stefan Th Gries. Dispersions and adjusted frequencies in corpora: Further explorations. In *Corpus-Linguistic Applications*, pp. 197–212. Brill, January 2010.

[18] Stefan Th Gries. What do (most of) our dispersion measures measure (most)? Dispersion? *Journal of Second Language Studies*, Vol. 5, No. 2, pp. 171–205, November 2021.

[19] Marc Brysbaert and Kevin Diependaele. Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, Vol. 45, No. 2, pp. 422–430, June 2013.

[20] A.G. Juilland, D.R. Brodin, and C. Davidovitch. *Frequency Dictionary of French Words*. Romance Languages and Their Structures. Mouton, 1971.

[21] Anthony A. Lyne. *The Vocabulary of French Business Correspondence: Word Frequencies, Collocations, and Problems of Lexicometric Method*. Travaux de Linguistique Quantitative. Slatkine, 1985.

[22] Inger Rosengren. The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée*, Vol. 1, p. 103, 1971.

[23] John B. Carroll. An Alternative to Juilland's Usage Coefficient for Lexical Frequencies. *ETS Research Bulletin Series*, Vol. 1970, No. 2, pp. i–15, 1970.

[24] Taichi Murayama, Shoko Wakamiya, and Eiji Aramaki. WORD GINI: A proposal and application of an index to capture word usage bias [WORD GINI: Go no shiyō no katayori wo tsukamaeru shihyō no teian to sono ōyō] (in Japanese). *The 24th Annual Conference of the Association for Natural Language Processing [Gengoshori gakkai dai 24 kai nenji taikai]*, pp. 698–701, 2018.

[25] Brent Burch, Jesse Egbert, and Douglas Biber. Measuring and interpreting lexical dispersion in corpus linguistics. *Journal of Research Design and Statistics in Linguistics and Communication Science*, Vol. 3, No. 2, pp. 189–216, October 2017.

[26] Jesse Egbert, Brent Burch, and Douglas Biber. Lexical dispersion and corpus design. *International Journal of Corpus Linguistics*, Vol. 25, No. 1, pp. 89–115, April 2020.

[27] Gerald J. Glasser. Variance Formulas for the Mean Difference and Coefficient of Concentration. *Journal of the American Statistical Association*, Vol. 57, No. 299, pp. 648–654, September 1962.

[28] Huarui Zhang, Churen Huang, and Shiwen Yu. Distributional Consistency: As a General Method for Defining a Core Lexicon. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[29] Mordecai Ezekiel. *Methods of Correlation Analysis*. Wiley, Oxford, England, 1930.

[30] David A. Balota, Melvin J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. The English Lexicon Project. *Behavior Research Methods*, Vol. 39, No. 3, pp. 445–459, August 2007.

[31] Yiu-Kei Tsang, Jian Huang, Ming Lui, Mingfeng Xue, Yin-Wah Fiona Chan, Suiping Wang, and Hsuan-Chih Chen. MELD-SCH: A megastudy of lexical decision in simplified Chinese. *Behavior Research Methods*, Vol. 50, No. 5, pp. 1763–1777, October 2018.

[32] Jose Armando Aguasvivas, Manuel Carreiras, Marc Brysbaert, Pawe l Mandera, Emmanuel Keuleers, and Jon Andoni Duñabeitia. SPALEX: A Spanish Lexical Decision Database From a Massive Online Data Collection. *Frontiers in Psychology*, Vol. 9, , November 2018.

[33] Yongqiang Su, Yixun Li, and Hong Li. Familiarity ratings for 24,325 simplified Chinese words. *Behavior Research Methods*, Vol. 55, No. 3, pp. 1496–1509, April 2023.

[34] Max Coltheart. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, Vol. 33A, No. 4, pp. 497–505, 1981.

[35] M. (Max) Coltheart and Michael John Wilson. MRC Psycholinguistic Database Machine Usable Dictionary : Expanded Shorter Oxford English Dictionary entries / Max Coltheart and Michael Wilson. *Oxford Text Archive*, March 1987.

[36] Agnes Sianipar, Pieter van Groenestijn, and Ton Dijkstra. Affective Meaning, Concreteness, and Subjective Frequency Norms for Indonesian Words. *Frontiers in Psychology*, Vol. 7, , December 2016.

[37] Masayuki Asahara. Word Familiarity Rate Estimation Using a Bayesian Linear Mixed Model. In Silviu Paun and Dirk Hovy, editors, *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pp. 6–14, Hong Kong, November 2019. Association for Computational Linguistics.

[38] Marc Guasch, Pilar Ferré, and Isabel Fraga. Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behavior Research Methods*, Vol. 48, No. 4, pp. 1358–1369, December 2016.

[39] Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In Ekaterina Kochmar, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan, editors, *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 571–589, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[40] R. H. Baayen. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, Vol. 5, No. 3, pp. 436–461, January 2010.