Data preprocessing methods for second language acquisition in mixed effects models

Qingsong Peng
College of Communication and Information Engineering, Shanghai Technical Institute of Electronics & Information, Shanghai 201411. China
peng.qingsong@qq.com

Hui Yuan
College of Foreign Languages
Shanghai Ocean University
Shanghai 201306, China
hyuan@shou.edu.cn

***Abstract*—The investigation into lexical and conceptual representation in second language (L2) acquisition has garnered significant attention in linguistics and cognitive psychology as it sheds light on the intricate mechanisms underlying the development of language proficiency in non-native speakers. The issue of handling outlier data in response time (RT) measurements, a cornerstone in many psycholinguistic experiments, is meticulously examined. The paper argues that variance normalization as a preprocessing step offers a robust method for mitigating the influence of extreme values that can potentially distort the overall interpretation of the data. Variance normalization involves transforming the RT data such that each observation is expressed in terms of its deviation from the mean scaled by the standard deviation of the dataset, thereby ensuring that outliers do not disproportionately affect the statistical analyses.**

***Keywords*—Lexical representation, Data processing, Variance normalization, Outlier handling**

I. INTRODUCTION

The intricate interplay between linguistic and conceptual representation in second language acquisition (SLA) has long been a focal point of research in linguistics, psychology, and cognitive science. As learners progress through various stages of SLA, their linguistic systems undergo profound transformations, manifesting in how they encode, retrieve, and manipulate linguistic units such as words and their associated meanings. This complex process of integrating novel lexical items into an existing or emerging conceptual framework has been the subject of numerous empirical investigations, yielding insights into the nature of linguistic plasticity, cognitive flexibility, and the intricate neural networks underpinning language learning. This paper aims to provide a comprehensive overview of the current landscape of SLA research about lexical and conceptual representation, delving into the fundamental methods employed in this field and offering a nuanced discussion on data processing techniques, particularly about the treatment of outlier response times.

Lexical representation refers to the mental storage and organization of words within an individual's linguistic system. Learners in SLA must understand new words' phonological and orthographic forms and their syntactic and semantic properties. This process involves establishing connections between the novel lexical items and their conceptual representations, which are the mental constructs that embody the meanings and associations of these words. As learners progress, their lexical networks expand, facilitating more efficient access to lexical information and more sophisticated linguistic productions.

On the other hand, Conceptual representation concerns the mental structures that underlie our understanding of the world. These structures are abstract, domain-general, and cross-linguistic, allowing people to categorize, reason, and communicate their experiences. Learners in SLA must map the meanings of new words onto their existing conceptual frameworks, often adjusting these frameworks as they encounter linguistic and cultural nuances that challenge existing conceptual boundaries.

The interface between lexical and conceptual representation is dynamic and bidirectional. Lexical knowledge facilitates access to conceptual representations, enabling learners to comprehend and express meaning. Conversely, conceptual knowledge influences lexical processing, guiding the selection and interpretation of words based on their semantic and pragmatic fit within a given context. This interplay is crucial for successful language learning as it allows learners to build rich, nuanced representations of language that are deeply integrated with their cognitive systems.

Response times in reaction time tasks that deviate significantly from the overall distribution pose a unique challenge for data analysis. These outliers can arise due to various factors, including inattention, misinterpretation of instructions or exceptional cognitive processing. While excluding outliers can improve the statistical rigor of the analysis, it can also risk discarding valuable information about individual differences or rare but meaningful cognitive processes. Nation's (2001) [1] review of vocabulary learning strategies highlights the importance of utilizing the relationships between words to facilitate vocabulary growth.

In the context of SLA research on lexical and conceptual representation, the variance-based preprocessing method offers a nuanced approach to handling outlier response times. This method examines the variance of response times within individual participants or experimental conditions rather than simply discarding outliers based on absolute thresholds. By identifying and adjusting for within-subject variability, the

variance-based preprocessing method can help normalize the data while preserving information about the underlying cognitive processes.

II. DATA PREPROCESSING PROCEDURE

Data preprocessing is a crucial step before data analysis, ensuring the quality and reliability of the subsequent analytical procedures. It involves a series of systematic processes aimed at cleaning, organizing, and transforming raw data into a format suitable for statistical analysis or machine learning models. This process is particularly important for mixed effects models, which often deal with complex datasets containing fixed and random effects. We will delve into the key steps of data preprocessing, principles and methods for removing outliers (i.e., excessively large or small data points), and the consequences of inadequate outlier cleaning. Additionally, we will cite relevant literature to support our arguments.

The data preprocessing typically encompasses the following essential steps:

Data Collection involves gathering all relevant data sources for analysis. In mixed effects models, data may originate from multiple observational units (e.g., individuals, organizations, or regions) and contain multiple measurements or observations per unit.

Data Cleaning involves identifying and correcting errors, inconsistencies, and inaccuracies in the data. It includes handling missing values, correcting typos, and ensuring consistent data types.

Outliers are extreme values that deviate significantly from the rest of the data, potentially skewing the results of statistical analyses. Identifying and removing or adjusting these outliers is vital to ensure the accuracy of the analysis.

Depending on the nature of the data and the analysis objectives, data transformation techniques (e.g., normalization, standardization, log transformation) may be applied to improve the distribution and comparability of the variables.

If the analysis involves multiple datasets, data integration involves combining these datasets into a single, cohesive format while ensuring consistency and accuracy.

In cases where the dataset is huge, data reduction techniques (e.g., sampling, feature selection) may be employed to simplify the data without compromising its informative value.

Removing outliers is a delicate process that requires careful consideration to avoid inadvertently excluding valid data points or failing to identify genuine outliers.

III. EXAMPLE OF DATA PREPROCESSING BY MIXED EFFECTS MODELS

The paper by Wu (2017) [2] delves into the intricate processes underlying the development of lexical and conceptual representations among Chinese learners of English as a Foreign Language (EFL). Drawing upon the theoretical framework of the Revised Hierarchical Model (RHM), the study aims to uncover the patterns and mechanisms that shape how these learners acquire and represent words and their associated meanings in a second language. By employing a mixed-effects linear modelling approach, Wu [6] sheds light on the nuanced differences in these representations across varying proficiency levels, as evidenced through a translation judgment task from English to Chinese.

The RHM, an extension of the Hierarchical Model of Language Acquisition (Kroll & Stewart, 1994) [3] posits that language learners' lexical and conceptual representations evolve through distinct stages. Learners initially rely heavily on their native language (L1) for comprehension and production in the second language (L2), leading to a strong L1-L2 link. As proficiency increases, learners gradually develop more autonomous L2 representations, weakening the reliance on L1 and fostering direct connections between L2 lexical items and conceptual representations. This transition is marked by the emergence of a shared conceptual space where both L1 and L2 representations converge.

The study of Wu (2017) [2,4,5] recruited a diverse group of Chinese EFL learners, stratified into three proficiency levels (beginner, intermediate, and advanced) based on standardized English proficiency tests and self-reported language use. Participants were given a translation judgment task, presented with English words and asked to indicate whether each word had a direct Chinese equivalent. This task aimed to elicit evidence of the learners' lexical and conceptual representations by assessing their ability to link English words to their corresponding concepts independent of their L1 translations.

Data from the translation judgment task were analyzed using mixed-effects linear modelling (MELM), a statistical technique that accounts for both fixed effects (e.g., proficiency level) and random effects (e.g., individual differences among learners).[7-11] MELM allowed Wu to model the complex interactions between proficiency, task performance, and individual variation, providing a nuanced understanding of the developmental trajectories.

IV. EXPERIMENTAL RESULTS

In exploring the developmental relationship between lexical and conceptual representations, the methodology employed to preprocess and cleanse data plays a pivotal role in ensuring the validity and reliability of subsequent analyses. Wu's original approach, which involved the elimination of outlier reaction times (RTs) below 150 milliseconds (ms) and those exceeding three standard deviations (SDs) from the mean, has established a baseline for data integrity. However, the study aims to delve deeper into the nuances of the data by exploring the impact of adopting less stringent thresholds for outlier removal, specifically targeting RTs deviating by two and one SDs from the mean. The adjustment in data cleaning strategies resulting in distinct datasets of 1784 and 1571 records offers valuable insights into how varying levels of

data strictness can inform our understanding of the intricate link between lexical and conceptual processing.

In data preprocessing, the meticulous handling of outliers is a crucial step that significantly impacts the accuracy and reliability of subsequent analytical outcomes. The decision to remove data points that deviate considerably from the norm, whether excessively large or unusually small, is a delicate balance between preserving the integrity of the dataset and eliminating potential distortions that could skew results. The approach outlined in the scenario—removing values greater than three standard deviations from the mean for large outliers and similarly values less than three standard deviations for small outliers—demonstrates a nuanced understanding of data distribution and the need for tailored strategies in different contexts. Compared to a more arbitrary threshold, such as removing all values below 150 milliseconds, as mentioned in Wu [2,4,5], the practice underscores the importance of conducting a thorough data analysis before applying any preprocessing techniques.

TABLE 1. METHODS FOR REMOVING OUTLIERS FOR REACTION TIME

Methods for removing an abnormal response	Number of records after removing abnormal data
None	2000
<150, and >mean(x)+3*σ	1826
<mean(x)-2*σ, and mean(x)+2*σ	1784
<mean(x)-σ, and mean(x)+σ	1571

Standard deviation quantifies the extent to which individual data points vary from the mean of the dataset. By setting thresholds based on multiples of the standard deviation, researchers and analysts can identify data points unusually distant from the central tendency of the data. The choice of three standard deviations as a cutoff point is often justified by the empirical rule (also known as the 68-95-99.7 rule), which states that approximately 99.7% of the data in a normal distribution fall within three standard deviations of the mean. Thus, values beyond this threshold are considered highly unusual and potentially indicative of errors, measurement issues, or genuine but rare occurrences that may not represent the overall population.

While standard deviation thresholds are a widely accepted method for outlier detection, it is essential to recognize that the same approach may not be equally effective for all outliers. The distinction in the given example between handling large and small outliers differently highlights the need for a nuanced approach. Large outliers, often resulting from data entry errors, measurement anomalies, or extreme but legitimate events, can significantly skew mean and variance calculations, distorting the overall picture of the data. Similarly, small outliers can also introduce bias or misinterpretation, especially when they fall below a certain threshold deemed unrealistic or unphysical for the context (e.g., response times below a certain minimum in a reaction time study).

The careful handling of outliers during data preprocessing has far-reaching implications for the quality and validity of subsequent analyses. Failing to identify and appropriately address outliers can lead to misleading conclusions, biased estimates and reduced statistical power. Conversely, overly aggressive outlier removal can result in losing valuable information and introducing bias by excluding legitimate but unusual observations. Therefore, a balanced approach that combines statistical rigor with contextual understanding is essential.

V. Conclusion

Using three standard deviations as a threshold for identifying and removing large and small outliers, as opposed to a fixed threshold like 150 milliseconds, underscores the importance of conducting a thorough and context-specific analysis during data preprocessing. By recognizing the nuances of data distribution and the potential impact of outliers on analytical outcomes, researchers and analysts can develop more effective and appropriate strategies for ensuring the quality and reliability of their findings. This approach not only enhances the scientific rigor of the analysis but also promotes the generation of insights that are more robust, accurate, and applicable to the real-world problems they aim to address.

References

[1] P. Nation, “Learning Vocabulary in Another Language”, 2001. Cambridge: Cambridge University Press.

[2] S.Y. Wu, “The development of lexical and conceptual representation in Chinese EFL learners: Evidence from mixed-effects linear modelling”. 2017.Foreign Language Teaching and Research, Sept. 2017, Vol. 49, NO. 5, pp 767-779.

[3] JF. Kroll, and E. Stewart, “Category Interference in Translation and Picture Naming: Evidence for asymmetric connections between bilingual memory representations”. In Journal of Memory and Language. 33: 149-174. 1994.

[4] S.YWu, “Second Language Processing and R in Application”. Foreign Language Teaching and Research Press. 2019

[5] S.Y. Wu, “Using R in Linguistic Research”. Science Press, 2021.

[6] S.Y. Wu, “Multivariate Analysis in Linguistic Research With R”. Shanghai Jiao Ting University Press, 2024.

[7] VA. Brown, "An Introduction to Linear Mixed-Effects Modeling in R. Advances in Methods and Practices in Psychological Science". 2021, 4(1). doi:10.1177/2515245920960351

[8] B.T. West, K.B. Welch, and Galecki, A.T, "Linear Mixed Models: A Practical Guide Using Statistical Software" (3rd ed.). 2022, Chapman and Hall/CRC. https://doi.org/10.1201/9781003181064

[9] LM. DeBruine, Barr, DJ, "Understanding Mixed-Effects Models Through Data Simulation. Advances in Methods and Practices in Psychological Science". 2021, 4(1). doi:10.1177/2515245920965119

[10] L. Meteyard, R.A.I. Davies, Best practice guidance for linear mixed-effects models in psychological science,

Journal of Memory and Language, Volume 112, 2020, pp. 104092

[11] L. Kumle, Levi, "Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R" Behavior Research Methods, 2021, 2528-2543, 53. 6.