# HOW TEXTS CONFUSE ONLINE BUYERS: QUANTIFYING TEXT QUALITY OF ONLINE PRODUCT DESCRIPTIONS[1]

Yang Sun
College of Management
Park University
8700 NW River Park Dr, Parkville, MO 64152, United States
yang.sun@park.edu

Shaonan Tian[2]
Department of Marketing and Business Analytics
Lucas College of Graduate School of Business
San Jose State University
One Washington Square, San Jose, CA 95192, United States
shaonan.tian@sjsu.edu

Ming Zhou
School of Global Innovation & Leadership
Lucas College of Graduate School of Business
San Jose State University
One Washington Square, San Jose, CA 95192, United States
ming.zhou@sjsu.edu

**ABSTRACT**

Online marketplaces are growing internationally with sellers coming from around the world. Product description texts often stand as the first piece of information soliciting buyers' attention, where the seller's culture and linguistic backgrounds drive great variations in description style and appeal. Product descriptions, on top of being a self-selected disclosure of product information, offer first-impression signals to infer a seller's credibility, competency, experience, and more for a buyer's decision-making. Sellers therefore make strategic decisions in composing product descriptions to woo potential buyers. In this research, we studied composition quality of product descriptions, a type of marketer-generated content, in an online marketplace. We collected online product description data and quantified their properties using data mining methods to define perceived description quality and characterize quality differentiating dimensions. Our analysis identified ten defining characteristics for differentiating native versus non-native speaker drafted descriptions. By combining business text analytics methods with modern linguistic theories and tools, this study is the first research to systematically quantify and characterize the concept of composition quality within the context of consumer-perceived nativeness. Our findings further complete and advance the literature on marketer-generated content (MGC), consumer decision-making, and seller operations decisions.

Keywords: Description text quality; Online consumer decision; Online seller operations; Data mining; Marketer-generated content

## 1. Introduction

"I buy pretty much every gadget from Amazon.com." a California consumer said, "but it would be so much easier if some sellers could better describe their products." Another consumer chimed in: "Yeah, some sellers got great prices and solid stuff, just sometimes their descriptions threw me off." That was a real conversation one of the authors participated in 2020. The two having the conversation founded an Amazon vendor consulting company in 2021. They surveyed more than three hundred consumers in the Silicon Valley area. About 87% of the three hundred respondents picked product descriptions on Amazon.com as the most challenging issue in their shopping experience. A similar

---

[1] **Cite**: Sun, Y., Tian, S., & Zhou, M., (2024, May) How Texts Confuse Online Buyers: Quantifying Text Quality of Online Product Descriptions, *Journal of Electronic Commerce Research, 25*(2).
[2] Corresponding author. Authors are in alphabetic order.

percentage of respondents would be more inclined to buy and/or to pay more for better-described products. These respondents preferred the descriptions to be of or close to native US English in composition quality. Consumers grapple with every piece of product information and form their reservation prices (Myerson, 1981) of a product. Descriptions, along with other information, can indispensably impact consumer purchase decisions.

Now, let us switch to the other side of a transaction, the selling side. One of the authors helped the two founders of the consulting company connect with Amazon.com sellers in Shenzhen, China. A conversation with a medium-sized company with an annual sale of about US $5-6 Million best brought up the dilemma many sellers face on Amazon.com. Their operations manager in charge of posting products online said: "We try to write in English. Google translate and sometimes translating professionals are accessible to us. We know we haven't been good enough. But really, we don't know what "good" looks like. This market (Amazon.com) is full of sellers from China and other countries. We copy each other's descriptions". Sellers make decisions on their descriptions to best attract consumers. In the absence of knowing what "good" means in online descriptions, a common practice has been to plagiarize others, which leads to their trash-in-trash-out predicament. On top of confusing consumers, low-quality descriptions put international sellers at a communication/promotion disadvantage, which leaves pricing as the only competitive apparatus in their toolbox. Fierce price competition drives down margins, takes away budgets for branding, and then stifles R&D or quality spending. They keep coming back to this vicious cycle only to get more engulfed. None of us in the conversation with that Shenzhen manufacturer was surprised when Amazon banned over 600 Chinese brands across 3,000 seller accounts (Hollister, 2021). These sellers should bear full responsibility for violating platform policies. Unfortunately, the reality had always been that their strategic spaces were so limited that it was just a matter of time before gray areas had to be tapped. This seemingly small problem, namely description quality, bears profound relevance to both consumers' purchasing decisions and seller competitive strategies.

In this research, we attempted to tackle the problem of description quality in an online marketplace. In particular, we focus on defining and identifying lexical characteristics that define the quality (nativeness) of product descriptions. The practical purpose is to assist online e-commerce sellers to improve their product descriptions and provide better seller-buyer online information communication. We adopted Coh-Metrix, an automated text parsing tool to assess product description texts. Different from the myriad of linguistic studies based on academic compositions, our research validated and selected lexical measurements for their applicability within the context of commercial product descriptions. We exposed real textual descriptions to well-established lexical metrics. Lexical measurements that best defined the perceived composition quality (nativeness) of product descriptions in an e-commerce environment were identified in this study. This, to the best of our knowledge, is the first business research targeting the composition quality of commercial online product descriptions in an English environment.

Our study focuses on a new aspect of e-commerce that has not been well-studied in the current literature. Traditional e-commerce literature is predicated on the trust assumption (Gefen et al., 2003; Zhou et al., 2007) where both sides of a transaction could securely conduct business by interpreting information relevant to a seller's posting. Online information became signals to infer many aspects of a product and its sellers (Spence, 1973). Chung and Sarnikar (2021) classified online information into user-generated content (UGC) and marketer-generated content (MGC). UGC mostly refers to user reviews or word of mouth (WOM) while MGC is perceived as seller self-selected information for buyers to infer sellers' intention and/or other subtleties. Many studies in marketing and information system research have examined the effects of UGC in online markets. For example, Yoo et al. (2013), explored the effect of e-WOM in e-commerce. Li et al. (2019) examined the effect of online reviews on product sales. Other studies have further extended the analysis of UGC to factors such as review valence (Rosario et al., 2016) or agglomeration (Liu et al., 2018). While research on studying the impact of UGC has been voluminous, the literature on MGC received little attention (Chung & Sarnikar, 2022). To this end, we attempted to fill the gap by studying the implications of MGC through analyzing product description quality in an online marketplace. Our research takes a text analytics perspective by directly dissecting seller product descriptions. Our findings contribute to extant theories of consumer decision-making and trust building through MGC since product descriptions can unequivocally impact various stages of consumer online decision-making (Liang & Lai, 2002). E-commerce sellers benefit from our findings by obtaining highly practical and implementable insights for drafting product descriptions to better earn buyer trust. The traditional amorphous and intuition-dependent process of composing descriptions lacked quantitative directions and our findings bring unprecedented transparency and operational practicalities for e-commerce operations. Our interdisciplinary findings based on well-established linguistic theories enrich the solutions to a business problem using modern text mining techniques and well-recognized open-source tools. This work even sheds light on the popular large language models by identifying characteristics of natively drafted compositions that can be used to improve language model results.

In the next section, we review the current literature. Our data collection is described in Section 3. Section 4 contains our methodology and empirical findings. We discuss results and contributions in Section 5 along with managerial implications and research limitations.

## 2. Literature Review

A sizable portion of our commerce transitioned to online since the early 2000s. Ebay.com and Amazon.com reshaped how e-commerce transactions have been conducted. Instead of physically inspecting goods and interacting with a seller (Gefen et al., 2003), a buyer saved the trip to a store at the expense of information completeness. Online information is reduced to presented cues varying from simple texts to multimedia content, such as photos and videos (Koh & Cui, 2022). Chung and Sarnikar (2021) coined the term "marketer-generated content (MGC)" to refer to seller-provided information. Online environments of reduced information demanded a new era of trust-building (Dimoka et al., 2012). Kim et al. (2008) found that trust and perceived risk strongly impacted purchase decisions. Gefen et al. (2003) found ease of use in combination with trust and safety mechanisms on a website influenced trust. Unsurprisingly, current literature confirmed the idea that trust has been an indispensable factor in e-commerce environments.

A key paradigm in online trust research was about trust building in online transactions. One factor that has been identified as critical for online trust-building is information quality (Mun et al., 2013). In Gao et al. (2021), they showed that trust impacted consumer decision-making under the moderation of some other factors in online purchases. Higher quality information assured a buyer that the selling side could be more trustworthy in qualification, experience, and/or intention. Information quality is a multi-faceted concept that encompasses any information a buyer can obtain online before purchase decisions. Gursoy (2019) identified "perceived cost of information search, cognitive processing required, level of consumer processing capabilities, and level of consumer involvement with a product" as influencing factors for a consumer to pick an information source. Lu and Chen (2020) postulated that content routineness and time routineness impacted the clicking behavior of online advertising content. Traditional research on online information content dealt with website design, user interface, ease of use, and so on (Gefen et al., 2003). Some directed their attention to eyeball tracking with the exposure to various image content (Chocarro et al., 2022). Within the studies of text contents, many customers often rely on the online review or feedback system (Biswas et al., 2022; Zhou et al., 2008; Lee et al., 2017; Kim & Chun, 2019). Chung and Sarnikar (2021) believed listing descriptions are likely to be the lowest-cost information source for a consumer. Seller-provided description information is often not fully repeated in other information sources, such as consumer reviews. Description texts' low-cost nature also renders it an economic and highly efficient beginning point to initiate an online trade. Research of product descriptions shall offer important insights into decision-making for both sellers and buyers.

Extant research suggested that product descriptions were relevant to e-commerce participants' decision-making. Liang et al. (2020) clustered Airbnb listing description information. They did not attempt to quantify description quality. Instead, they correlated text patterns with the number of reviews. Chung and Sarnikar (2021) also analyzed Airbnb datasets. Their analysis identified the most frequent lexical patterns appearing in Airbnb descriptions. Pryzant et al. (2017) implemented a neural network architecture on product description texts from Rakuten.com, a Japanese e-commerce website. They specifically emphasized how "textual content of product descriptions impacted sales". Their results showed that narrative elements of descriptions were predictive of sales. More interestingly, they showed that the text features were mostly independent of other confounding factors in determining sales. Current research well analyzed textual patterns of product descriptions, but a direct assessment of description quality from the perspective of composition efficacy was still missing. The efficacy of any description lies in its quality as a union of all components with structure, flow, and cohesion. Product description quality is essentially composition quality. To this end, this research attempted to contribute to the literature by performing macro-level analysis to study the relationship between composition quality and L2 learner's online product descriptions.

1.1 In Linguistics, the quality of a composition may vary with rater, genre, and context differences, but it encompasses a basic set of criteria for quality writing. Diederich (1974) developed a six-trait rubric composed of ideas and development, organization, voice, word choice, sentence fluency, and conventions. Text quality prediction was "the computational approach to detect and score these traits in writing." (Louis, 2013). Predictions could be impacted by factors such as range of traits and audience. The range of traits refers to presentation format and style. Audiences often varied by age, technical expertise, and cognitive capacities. Composition quality in Linguistics literature was a multi-dimensional concept. A review by Louis (2013) thoroughly discussed major dimensions of text quality. Text readability received the most attention from existing research. In an academic context, readability was defined as a one-to-many mapping from reader competency levels to a set of texts. Matching competency with corresponding texts led to a higher level of readability. Other facets of readability included syntax complexity and discourse properties (Scharm & Ostendorf, 2005). Latent Semantic Analysis (LSA) overlap score was a good measurement for discourse properties (Foltz, et al., 1998). Text convention was another dimension of text quality. Its research has mostly been about spelling, grammar, and punctuation (Louis, 2013). Text organization could be understood as coherence. How words and sentences have been tied together to present a flow demonstrates coherence. Entity repetition and pronoun use between adjacent sentences based on centering theory (Grosz, et al., 1995) very much measured the coherence of an article. Reader interest (Flesch, 1948) was theorized to correlate with text quality but later research paid less attention to reader interest, where only word and sentence length were retained in measuring text quality.

Second language (L2) learners (Bardel, et al., 2013) inspired another stream of linguistics research highly relevant to our study. Online listings in English were often drafted by non-native English speakers to communicate commercial information and persuade their buyers to purchase. Writing skills have been proven to be the most challenging learning for a second language learner (Yu, 2018). Corpus analysis (Wilson, 2001) explored lexical richness, lexical diversity, lexical errors, and so on. Laufer (2012) listed vocabulary, grammar, structure, and use of vocabulary as fundamental qualities of an L2 composition. Crossley and McNamara (2011) proposed to use of lexical proficiency to analyze L2 writings, which included lexical originality, sophistication, diversity, density, error, and coherence. Originality turned out to be a less reliable indicator (Laufer & Nation, 1995). Lexical sophistication advocated the usage of advanced words as measured by lexical frequency profiles (LFP). Lexical diversity is about different types of vocabulary, and it has been measured by the measure of textual and lexical diversity (MTLD). Lexical density was the percentage of lexical words compared to the total number of words. Lexical errors were a ratio between lexical errors and the total number of errors. Lexical coherence was defined as "consistency of the text from the perspective of the reader's mental process (Louwerse, 2004).

## 3. Data

This section focuses on our data collection. Data source, data collection, and data processing/structuring are elaborated here in detail.

### 3.1. Data Collection

Our research endeavored to identify metrics that separate varying qualities of description compositions. Previous research collected data from e-commerce websites like Airbnb (Liang et al., 2020; Chung & Sarnikar, 2021) and Rakuten (Pryzant et al., 2017). We followed the convention and collected our description texts from US e-commerce giant Amazon. Thanks to the consulting firm's kind support, we were given access to a professional Amazon software, called JungleScout. JungleScout specializes in product and vendor information search.

The team went through multiple rounds of deliberation and came to the following criteria to better analyze descriptions. First, we searched the JungleScout vendor database to ensure geographical diversity. A product category fully dominated by US sellers or by international sellers was not interesting to this research. Second, we sought inputs from the consulting company. Their professional experience suggested products priced from US $20-50 per unit. Anything lower than US $20 often rendered product descriptions trivial in purchase decisions while markets of products over US $50 tended to contain dominating brands that reduce description diversity. Third, we intentionally identified products that were complex enough to entail description reading. For example, a laptop camera cover normally does not need much description reading before purchase. Fourth, we selected Amazon Prime, four-star and above ratings, and stock available options under JungleScout to better minimize impacts from system design or logistical factors.

Four product categories were selected for our data collection: oximeters, water flasks, lighted vanity mirrors, and ring lights. These four products offered good product variations ranging from medical needs during COVID-19 and daily accessories to cosmetic care and social media supplies. They were mostly priced within our target price range and vendor locations displayed balanced geographical diversity. Product descriptions were manually collected from the website. Our final sample, after reviewing and cleaning, contained three hundred and seventy (370) listings across four product categories.

### 3.2. Composition Quality Rating

Product listing descriptions from Amazon were further coded for quantitative analysis. Coding Step 1 was to develop a rating scale for description texts. Coding Step 2 parsed the texts into lexical metrics.

Rating scales were developed in our research using human raters. Human rater or human intervention/direction has been a widely accepted practice for empirical theory building in the field of Linguistics. More recently, language models and text mining research also adopted this method for using human inputs to gauge and train data mining models. In Linguistic literature of second language learning research, human raters have been prevalent for evaluating composition quality. Engber (1995) recruited and trained teachers to grade compositions based on pre-developed rubrics. Polio (1997) built a team of one professor and one graduate student to rate essay quality. Louis (2013) used annotators to manually evaluate the sentimental and emotional aspects of a composition. Li et al. (2022) similarly asked a group of teachers to rate one hundred and eighty-five student writings with established rubrics. Language model and text mining research referred to human inputs under the general theory of reinforcement learning from human feedback (RLHF). Casper et al. (2023) best described RLHF in three stages: feedback collection, reward modeling, and policy optimization. Feedback collection utilizes human evaluation of outputs or outcome variables to instruct a model and better results. Reward modeling is the stage where human evaluations are replicated or trained through supervised learning methods. The theoretical foundation came from the revealed preference theory in economics (Chambers & Echenique, 2016; Wirth et al., 2017). Lin et al. (2020) believe human trainers can provide demonstrations, instructions, and feedback to a machine system. Ouyang et al. (2022) used human testers to optimize their language model. Bai et al. (2022) crowdsourced human raters to interact with their models.

Our team of three faculty members and three volunteer raters worked on description ratings. We recruited three senior students born and educated in the US. Each student worked independently in one semester. All raters were clearly instructed on research goals and expected rating outcomes. The three faculty members held meetings with raters to discuss their concerns and questions during the rating process. Volunteer raters had adequate initial training periods for them to familiarize themselves with listing styles. Each rater first tested ratings on a small sample and presented their rationales to the faculty team. This process of training and iteration reduced bias and enhanced rating consistency.

The team agreed upon a rating of 1-3 scale, where 3 was the highest rating and 2 was a semi-native or undecided category. The highest rated descriptions were within a close range of native speaker quality whereas lowest rated descriptions were perceived as non-native quality. It should be noted that this research was designed to understand how consumer decisions might be influenced by perceived description quality or impression of a description. The quality of a description, native quality or not, was mostly a perception. Our raters were instructed to code the descriptions based on their perception/impression.

The first rater used a semester to complete rating all listings based on consented evaluation rubrics. The second rater was born in the US but spent a sizable amount of her early years out of the US. Her ratings better captured the nonnative nature of descriptions. The third rater was middle-aged. Different from the first two raters who were about 20 to 22 years old, this rater brought in more maturity and was an expert in English grammar. The third rater was encouraged to refine our rubrics while staying conscious of common grammatical/usage mistakes made by native speakers in the US. The purpose of this research was not to police grammar on Amazon. Very interestingly, more than 60% of the third raters' ratings agreed with previous raters, which attested to the overall reliability of our data. For the portion of listings where the third rater disagreed with previous raters, the faculty team and the third rater revisited every listing to come to a consensus.

The second step of coding demanded a systematic tool to parse and quantify the textual properties of a description. We needed to transform text corpuses into numerical units for mathematical model analysis. We resorted to a prevailing linguistic computational tool, Coh-Metrix. The Coh-Metrix system was originally funded by the Institute for Educational Research (IES) in 2002. It was designed to measure the cohesion of compositions from any source. Coh-Metrix has expanded its measuring scope to 11 lexical categories totaling 106 metrics since its inception: descriptive indices, text easability assessment, referential cohesion measures, Latent Semantic Analysis (LSA) indices, lexical diversity measures, connectives indices, situation model output, syntactic complexity measures, syntactic pattern density measures, word information, and its readability scores. Detailed explanations can be found on http://cohmetrix.com/ and in the Coh-Metrix book (McNamara et al., 2014). Each listing's Coh-Metrix metrics were saved under its ASIN number. ASIN is Amazon Standard Identification Number. No other analytical tools we tested offered results close to Coh-Metrix in comprehensiveness and granularity.

## 4. Model and Empirical Results

Dissecting texts and analyzing their characteristic properties have been prevalent in modern data mining research (Roberts et al., 2016). We resorted to classic data mining methods employed in extant business and linguistic studies in our analysis. In this section, we present our data analysis, methodological considerations, and our model results.

### 4.1. Models

Coh-Metrix offered a great set of quantitative metrics for the text input. In this work, we focused on tackling perceived composition quality in product descriptions. Product descriptions were textual data. Coh-Metrix gave each description 106 quantified lexical indices. To cope with the "curse of dimensionality" problem and choose effective feature selection, we first adopted principal component-based factor analysis (FA) to lower the number of dimensions among the lexical indices.

The FA was performed within each of the Coh-Metric categories as listed in Table 1 since lexical indices in the same category tended to have relatively strong correlations and are likely represented by a smaller number of underlying latent variables (factors). We used the principal component analysis approach in our FA to identify the appropriate number of factors that accounted for the maximum variance among the lexical indices in the same category.

We used the eigenvalue criterion to determine the factors, where factors with an eigenvalue greater than one were kept in the FA result as such a factor would explain more variance than any single variable.

The 106 Coh-Metrix index variables were pre-categorized into 11 categories, each corresponding to a major theme in lexical analysis. The independent variables within the same category could naturally carry multicollinearity. The principal component-based FA helped us create a set of orthogonal components within each category. A few principal components were extracted as latent factors, while components with low eigenvalues were neglected as they were likely noise components. While multicollinearity could still exist among factors that were across categories, this FA approach largely reduced the chance of having multicollinearity in our analysis. At the same time, sufficient data variance was maintained when the dimensionality of the data was reduced by aggregating the variables into fewer comprehensive latent factors within each category. In this dimension reduction effort, we did not discard any input variables and retained sufficient information from all variables. Because we obtained these latent factors from each of the themed categories, they remained interpretable. By looking into the factor loading matrix, we were able to identify the input variables corresponding to a specific latent factor and understand the relationship between the latent factor and the input variables based on the strength of the relationship between each variable and the factor.

To test the significance of the remaining latent factors affecting the rated quality, we used multiple ordinal logistic regression (MOLR), a supervised machine learning technique for classification using an interactive-reweighted least-squares algorithm to obtain maximum likelihood estimates of the model parameters. In ordinal logistic regression, the response was a categorical variable with ordinal classes and its relationship with a number of explanatory variables (Coh-Metrix components) was modeled.

Our logistic response function was $P\left(y \le y_j\right) = \frac{e^{\alpha_j + BX}}{1+e^{\alpha_j + BX}}$, where y was labeled with three ordinal values, and $y_j$ was the $j^{th}$ possible value of y, j = 1,2. Note that in this model, $P(y \le y_3)$ always equaled 1, as $y_3$ was the highest possible value of y. Therefore, the j = 3 case was not needed in the model. $\alpha_j + BX$ was the linear component in the logistic model, where X was a vector of Coh-Metrix components (variables) as a result of the dimension reduction, and B was a vector of parameters (coefficients) in correspondence with these components. $\alpha_j$ was a constant for each j, j = 1, 2.

Thus, $P\left(y \le y_j\right)$ was simply the cumulative probability that the readability quality of the text to be of category $y_j$, $\forall j$, or lower. The logit of logistic cumulative probability was a linear model: $ln\left(\frac{P\left(y \le y_j\right)}{1-P\left(y \le y_j\right)}\right) = \alpha_j + BX$. A significant positive coefficient in this model indicated that as the value of the Coh-Metrix component increases, the likelihood of a higher level of y decreases. We started the model building by including all Coh-Metrix components as a result of the FA-based dimension reduction, and insignificant components were removed one by one in a step-wise procedure until all remaining components (variables) in the model were significant. Once a model was established, we then estimated the perceived readability quality $P\left(y = y_j\right) = P\left(y \le y_j\right), j = 1$ and $P\left(y = y_j\right) = P\left(y \le y_j\right) - P\left(y \le y_{j-1}\right), j > 1$ based on the Coh-Metrix component values.

For verification purposes, we chose a modern intelligent learning algorithm: neural networks. Different from many traditional statistical models, neural network models (NNM) make no assumption about the data by allowing non-linear decision boundaries. While a neural network is often considered a “black box” in the sense that it is difficult to abstract insights from the model, it allows better model flexibility and often offers improved classification performance. To examine our model’s predictive accuracy, we reported the neural network model’s confusion matrix. Confusion matrix is a popular measure for tabulating classification results in machine learning. With two dimensions – actual and predicted classes, the confusion matrix reports the total counts of correct predictions in the diagonal cells of the table. Incorrectly classified cases are positioned in off-diagonal cells. Higher cumulative diagonal counts often suggest better prediction accuracy.

### 4.2. Empirical Results

The MOLR took in dimension-reduced factors (resulting from FA using a principal component algorithm) from each major category of Coh-Metrix. Table 1 contains a summary of dimension reduction results. FA reduced dimensionalities from one hundred and six (106) metrics to thirty (30) dimensions. Some categories received sizable reductions, such as DES, EAS, COHE, and WRD, etc. We used JMP Pro developed by the SAS Institute to estimate our model. Significant factors with 5% or lower p-values were kept in our final model.

Table 1: Dimension Reduction with Principal Component-based Factor Analysis

| **Category** | **Original Dimension** | **Latent Variables** |
|---|---|---|
| 1. Descriptive (DES) | 11 | DES 1, DES 2, DES 3, DES 4 |
| 2. Text Easability (EAS) | 16 | EAS 1, EAS 2, EAS 3, EAS 4, EAS 5 |
| 3. Referential Cohesion | 10 | COHE 1, COHE 2 |
| 4. LSA | 8 | LSA 1, LSA 2 |
| 5. Lexical Diversity (LD) | 4 | LD 1, LD 2 |
| 6. Connectives (CON) | 9 | CON 1, CON 2 |
| 7. Situation Model (SM) | 8 | SM 1, SM 2, SM 3 |
| 8.Syntactic Complexity (SC) | 7 | SC 1, SC 2, SC 3 |
| 9.Syntactic Pattern Density (DEN) | 8 | DEN 1, DEN 2, DEN 3 |
| 10. Word Information (WRD) | 22 | WRD 1, WRD 2, WRD 3 |
| 11. Readability (READ) | 3 | READ 1 |
| *Sum:* | *106* | *30* |

Table 1: Dimension Reduction with Principal Component-based Factor Analysis (Cont.)

Table 2 reports the parameter estimation results of MOLR fitting on our dataset. Standard errors are reported in the parenthesis.

Table 2: Parameter Estimation Results

| | Coefficient Estimate | Chi-Sq Test Statistics | P-Value |
|---|---|---|---|
| Intercept[1] | 0.59 (0.13) | 21.71 | <.0001 |
| Intercept[2] | 1.64 (0.15) | 111.74 | <.0001 |
| DES3 | 0.96 (0.20) | 24.18 | <.0001 |
| EAS2 | 0.90 (0.19) | 22.62 | <.0001 |
| EAS5 | -0.63 (0.16) | 15.13 | 0.0001 |
| COHE1 | 0.47 (0.18) | 6.96 | 0.0083 |
| LSA1 | 0.40 (0.14) | 7.94 | 0.0048 |
| LD2 | 0.35 (0.14) | 6.51 | 0.01 |
| SM3 | 0.60 (0.15) | 16.67 | <.0001 |
| SC1 | 0.50 (0.13) | 14.27 | 0.0002 |
| DEN3 | 0.41 (0.14) | 7.91 | 0.0049 |
| WRD2 | 0.34 (0.12) | 7.85 | 0.0051 |

The confusion matrix of our MOLR is presented in Table 3.

Table 3: MOLR Confusion Matrix

| | Predicted Rating | | |
|---|---|---|---|
| Actual Rating | 1 | 2 | 3 |
| 1 | 0.943 | 0.000 | 0.057 |
| 2 | 0.759 | 0.000 | 0.241 |
| 3 | 0.530 | 0.000 | 0.470 |

Our confusion matrix displayed strong classification accuracy and precision in classifying descriptions rated 1 by our team. Because the data we obtained naturally carried an imbalance class distribution, it was expected that the prediction was skewed towards supporting the majority class (Armah et al. 2014). The model correctly classified 215 out of the 228 descriptions with a rating value of 1, a 94.3% accuracy rate. The accuracy obtained in the confusion matrix was reasonable for the minority classes, as well. The classification performance of our model for descriptions with rating value 3 was 39 out of 83, with a 46.99% accuracy rate. This class of rating was a minority class as descriptions with rating value 3 accounted for only 22.43% of our sample. We expected our model to perform poorly for the group of descriptions with a rating value of 2 as human raters were not sure about the writing being “native” or “nonnative”. Interestingly, our model swept 44 out of the 59 descriptions with a rating of 2 to the lowest rating group, where the remaining 15 were grouped with the highest rating descriptions. It could be plausible that an objective system found the second group of descriptions sharing more commonalities with the other two groups, where human raters were incapable of identifying the nuances in them. From a practical point of view, helping practitioners detect low-quality or “nonnative” descriptions was of the most importance to provide directional insights for improvement.

Our model retained the following factors, DES3, EAS2, EAS5, COHE1, LSA1, LD2, SM3, SC1, DEN3, and WRD2. We further inspected each factor to understand their overarching themes to give each of them a comprehensible meaning. Fortunately, these factors displayed strong themes for us to gain insights and explain our results. The probabilistic results of our model are graphed in Figure 1.

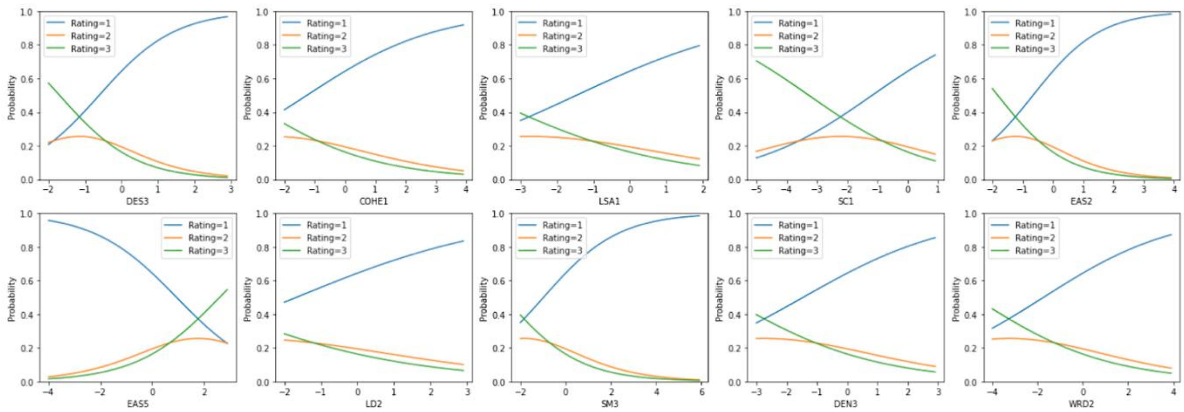


Figure 1: Cumulative Probability of Multinomial Logit Model Results

DES3 was the 3rd factor of the descriptive category. This category mostly contained basic descriptive statistics of a description. The factor DES3 was loaded with three original metrics, DESWC, DESSL, and DESSLd. DESWC is the total number of words in a description text. DESSL is the average number of words in a sentence of a text. DESSLd refers to the standard deviation of sentence length (DESSL). Please see McNamara et al. (2014) for a detailed explanation of each acronym. Descriptively, DES3 aggregated information related to description length and sentence length. The loadings of all three variables with DES3 were positive while parameter estimation of DES3 in our MOLR was positive. A higher value of DES3 value increased the probability of rating 1 as shown in Figure 1, which meant it led to lower perceived quality. Longer texts, longer sentences, and larger variations of sentence lengths turned out to suggest a worse impression of a description. The consulting company’s research seemed to confirm this pattern. Clumsier usage of English and lack of succinctness in lower-quality descriptions tended to push product descriptions longer with higher variations of sentence length.

EAS2 was the 2nd factor of text easability. The category captured readability and text comprehension. This factor contained only two metrics, PCSYNp and PCSYNz. PCSYNp and PCSYNz both capture sentences characterized by fewer words and simpler/more familiar structures either as a percentile or a z-score. Positive loadings and positive parameter values of EAS2 suggested that higher values of the metrics increased the probability of being a lower-quality composition. Simple structure sentences were straightforward in structure without many components. The presence of simply structured sentences in a product description might suggest that the writer lacked the necessary linguistic sophistication to eloquently describe a product to its readers.

EAS5 as a factor category showed divergence in its Coh-Metrix metrics’ impacts. The factor contained PCVERBp, PCVERBz, PCNARp, and PCNARz. The first two metrics, PCVERBp and PCVERBz, measure repeating and overlapping verbs in a text. These two metrics touch on the concept of referential cohesion (Sabatini, et al., 2012). Referential cohesion measures “the degree to which there is overlap or repetition of word concepts across sentences, paragraphs, or the entire text.” Both metrics’ loadings were negative and EAS5’s parameter estimation value was negative. Their overall impact was to increase the probability of low-quality writing. Repetition of verbs, maybe due to limitation in vocabulary or lacking sophistication of verb selection, suggested a description was more likely drafted by a non-native writer. The last two metrics, PCNARp and PCNARz, described a text that “tells a story, with characters, events, places, and things that are familiar to the reader”. The narrative is closely affiliated with every day, oral conversation” (Coh-Metrix). These two metrics reflected the context or cultural familiarity of everyday conversations. Their positive loadings within the factor, jointly with a negative parameter value, pointed to a higher probability of high-quality descriptions. We found this result intuitive to understand. Non-native English speakers lacked multiple linguistic capacities, which might mean insufficient context/cultural familiarity. A description deploying colloquial and familiar cultural usages was less likely to be drafted by someone who was not immersed in an English culture or daily environment.

COHE1 was the first factor of referential cohesion. CRFNOa, CRFNO1, CRFSOa, CRSFO1, CRFAOa, CRFAO1, CRFCWOa, and CRFCWO1 were Coh-Metrix metrics of this factor. The eight sub-factors brought up four dimensions. Overlap of nouns with every other sentence of a text (global) and with the immediate previous sentence (local) was captured by CRFNOa and CRFNO1. Stem overlaps, globally and locally, were measured by CRFSOa and CRSFO1. Stem was the root commonality between the two words. For example, the words "pricy" and "priced" shared the common stem word "price". CRFAOa and CRFAO1 were about global and local overlaps of nouns and pronouns. CRFCWOa and CRFCWO1 measured the proportion of explicit content words that overlapped between sentences. These two were overlapping measures normalized by sentence length. These metrics all scored positive loadings in the factor and COHE1's parameter value was positive. Higher overlapping of nouns and pronouns as well as higher usage of words sharing stems could be indicative of low quality, or less native, English writings. This factor seemed to point to language ability inadequacies as the culprit, such as writers' deficiency in vocabulary of nouns and pronouns, usage of nouns and pronouns, and selection of words.

LSA1 was the 1st factor of the LSA category. The category was all about Latent Semantic Analysis measurements. Per Coh-Metrix, "LSA provides measures of semantic overlap between sentences or between paragraphs." Please refer to the Coh-Metrix documentation for examples. This factor included LSASSp, LSASSpd, LSASS1, LSASS1d, LSAGN, and LSASSNd. LSASSp measured sentence conceptual similarity between any pairs of sentences (global) in a text, where LSASSpd was LSASSp's standard deviation. LSASS1 measured sentence conceptual similarity between any adjacent pairs (local) of sentences in a text, where LSASS1d was the standard deviation. LSAGN measured givenness, which means existing or provided information in current content, where LSASSNd was the standard deviation. All loading values were positive, and the multinomial logit model parameter value was also positive. A higher value of factor LSA1 increased the probability of rating 1, which meant it hinted lower quality writing or more non-native writing. The overarching theme in this factor was conceptual overlaps and givenness in a description. Our results suggested that a higher level of global and local conceptual overlaps, a higher level of givenness, and higher variations of overlaps and givenness were more likely a composition habit of non-native speakers. High-quality commercial writers, whose English was more native to our context, tended to avoid conceptual overlaps or repeating given information to optimize their composition within their word limits.

LD2 was a single metric factor that contained only LDVOCDa from Coh-Metrix. LDVOCDa was a VOC lexical diversity measure for all words. The concept of lexical diversity pertained to two key concepts, type and token. The type was the variety of unique words, and the token was the total number of words in a text. Quoting Coh-Metrix again, if the word "Dog" appeared in a text 7 times, its type value was 1 but the token value was 7. A text with a lexical diversity measure of 1, meaning each word showed up only once, could be hard to read or the text was very short. Our result suggested that a higher value in lexical diversity tended to present a description of lower quality, more likely composed by a non-native speaker. In other words, a description with many unique yet unrepeated words or an abnormally short description was less native a writing.

SM3 was another single metric factor that contained only SMCAUSr from Coh-Metrix. SMCAUSr was within the situation model category, which was the ratio of causal particles to causal verbs. Causal cohesion is one source of cohesion in a composition. Causal particles or connectives, such as because, so, and therefore, serve as linguistic cues for relationships. Causal verbs often include make, cause, allow, help, enable, and so on. They indicate that people or things cause something to happen (Nordquist, 2020). Higher values can be either too many causal particles or too few causal verbs. Positive loading and positive parameter value suggested that a higher value in this metric pertains to a lower-quality description. Crossley et al. (2007) noticed similar patterns in materials provided to second language learners (L2). Authentic and simplified texts were compared. Simplified texts were simpler or abstract versions of authentic texts. Authentic texts enjoyed a higher causal verb to causal particle ratio (the inverse of SMCAUSr). It was noted that simplified texts helped L2 learners as L2 learners "benefit from texts that are lexically, syntactically, and rhetorically less dense than authentic texts." As non-native speakers learn English often through reading simplified texts, their skewed exposure to simplified texts might train them to write with a higher SMCAUSr, a lower causal verb to causal particle ratio.

SC1 grouped three metrics from the syntactic complexity category of Coh-Metrix. Coh-Metrix documentation explained syntactic complexity as "some sentences are short and have a simple syntax that follows an actor-action-object syntactic pattern, have few if any embedded clauses, and have an active rather than passive voice. Some sentences have a complex, embedded syntax that potentially places heavier demands on working memory." Three metrics were the main members of the factor, SYNMEDlem, SYNMEDpos, and SYNMEDwrd. All three were related to the concept of minimum editorial distance. Minimum editorial distance essentially deals with the question of how similar two text strings are. It was the minimum amount of editing operations one would need to transform one string to another (Bringmann et al, 2019), such as converting the word “grail” to “giraffe”. All three metrics’ loadings were positive and SC1’s parameter value was also positive, where higher values in the three metrics suggested lower-quality writings. The implication was that descriptions demanding less working memory to read would sound more native in an online listing context.

DEN3 referred to syntactic pattern density. It was the third factor of the category. The category captured the density of particular syntactic patterns, word types, and phrase types, such as the incidence of noun phrases, adverbial phrases, and propositions. The density of these syntactic patterns could impact the processing difficulty of a text. Two metrics were part of DEN3, DRINF and DRVP. DRINF was the density of infinitives, such as “to do”. DRVP was the density of verb phrases. Our results suggested a lower-quality description tended to have higher densities of infinitives and verb phrases given both metrics loadings were positive and the DEN3 parameter value was also positive.

WRD2 was the second factor in the word information category. Words were assigned into syntactic categories, such as content words (nouns, verbs, adjectives, etc.) and function words (prepositions, determiners, pronouns, etc.). Coh-Metrix assigns only one category to each word within the context of the text. For example, the word bank in the phrase “river bank” is a noun, whereas the word bank in “don’t bank on it” is a verb. The factor housed four metrics, WRDIMGc, WRDCNCc, WRDMEAc, and WRDAOAc. Interestingly, these metrics differed in their impacts on description quality. WRDIMGc measured if a word was conducive to the mental image. For example, high mental image words could be bracelet and hammer whereas low mental image words could be reason and dogma. WRDCNCc measured if a word was concrete or non-abstract, things that one could hear, taste, or touch. Box and ball were more concrete than protocol and the difference was. WRDMEAc rated words by their meaningfulness, where a high-scored word tended to be a word highly associated with other words, such as people vs abbess. All three metrics’ loadings were positive and WRD2’s parameter value was also positive. The three metrics pointed to lower-quality writing. A common theme of the three metrics was the usage of basic and simple words. Gardner (2013) believed that usage patterns of vocabulary could be a differentiator between non-native and native English speakers. The last metric, WRDAOAc, directly measured the age of acquisition of a word. Words such as milk and pony surely were more likely acquired at early ages than words such as dogma and cortex were. Its negative loading combined with WRD2’s positive parameter value pointed to a higher probability of high-quality writing. This was not so hard to see as the usage of English vocabulary acquired later in one’s life is an indication of education, sophistication, and experience in an English environment. Overall, all four metrics conformed to one conclusion: the usage of more sophisticated words was an indication of higher quality or more native English descriptions.

We constructed a neural network model (NNM) to validate the classification performance of our MOLR. NNM is a machine learning method inspired by the networks of neurons in the brain and is particularly adept at classification tasks where the goal is to predict the correct label for given input data. First, there is an input layer in NNM that receives the raw feature data. In the context of our classification purpose, all retained factors after the dimension reduction were used as input nodes in the input layer in our NNM. The output layer contained one node, the rating classification.

Between the input and output, we constructed one hidden layer. Given our sample size and 30 input nodes, we chose to have 4 neuron nodes in the hidden layer to prevent overfitting. Each neuron node in the hidden layer transforms the values from the input layer with a nonlinear transformation of the weighted linear summation. Similarly, the output node transforms the values from the hidden layer as a logistic transformation of the linear combination of the transformed values. That is, connections between neuron nodes of different layers have associated weights, which represent the biases neurons have. These weights are the parameters in the NNM. The weights are initially set randomly and then optimized through training. In the optimization mechanism behind NNM, the objective function is a loss function, where the model’s predictions are compared to the true labels in the raw data. This measures how well the model’s predictions match the manually labeled classes.

We constructed the NNM in JMP Pro. 67% of our dataset was used as the training set and 33% were the testing set in our NNM effort. Both in-sample training and out-of-sample testing results were close to our MOLR results. For descriptions with rating 1, NNM in-sample training and out-of-sample testing results were 96.7% and 93.4% accurate. Both percentages were very close to our 94.3% accuracy rate from MOLR. Training and testing results from NNM for descriptions with rating 3 were 58.2% and 28.6% respectively, where MOLR scored 46.99%. Descriptions with rating 2 stayed consistent and it performed equally poorly between NNM and MOLR results. Overall, the NNM validated that factors retained from our MOLR model could provide good predictability for the perceived quality rating.

Table 4: In-Sample Confusion Matrix -- NNM

| | Predicted Rating | | |
|---|---|---|---|
| Actual Rating | 1 | 2 | 3 |
| 1 | 0.967 | 0.000 | 0.033 |
| 2 | 0.667 | 0.077 | 0.256 |
| 3 | 0.400 | 0.018 | 0.582 |

Table 5: Out-of-Sample Confusion Matrix -- NNM

| | Predicted Rating | | |
|---|---|---|---|
| Actual Rating | 1 | 2 | 3 |
| 1 | 0.934 | 0.026 | 0.039 |
| 2 | 0.800 | 0.050 | 0.150 |
| 3 | 0.679 | 0.036 | 0.286 |

## 5. Discussions and Conclusions:

5.1. Discussion of findings

Product descriptions, as a key category of MGC (Chung & Sarnikar, 2021), are highly conducive to a better understanding of consumer decision-making, which in turn feeds into businesses' decisions in system, product, and promotion designs. This research is the first study that directly tackles the problem of defining and quantifying the composition quality of product descriptions based on linguistic theories. We hope to build the foundation of a new stream of cross-disciplinary research and inspire more explorations to further our knowledge of textual characteristics/properties of MGC for business decision-making.

Table 6: Impact Summary

| **Factor** | **Subfactor Impact** | **Description Impact** | | **Theme** |
|---|---|---|---|---|
| | | **Low Quality** | **High Quality** | |
| DES3 | Same for all | + | - | Longer text, longer sentences, and higher sentence length variations |
| EAS2 | Same for all | + | - | Simply structured sentences |
| EAS5 | PCVERBp | + | - | Repetition and overlapping of verbs |
| | PCVERBz | + | - | |
| | PCNARp | - | + | Cultural familiarity and daily conversation |
| | PCNARz | - | + | |
| COHE1 | Same for all | + | - | Overlapping nouns and pronouns + word stem sharing. |
| LSA1 | Same for all | + | - | Concept overlaps and higher givenness |
| LD2 | Same for all | + | - | Many unique yet unrepeated words |
| SM3 | Same for all | + | - | Causal particle to causal verb ratio |
| SC1 | Same for all | + | - | Minimum editorial distance |
| DEN3 | Same for all | + | - | Density of infinitives and verb phrases |
| WRD2 | WRDIMGc | + | - | Concrete yet simple words |
| | WRDCNCc | + | - | |
| | WRDMEAc | + | - | |
| | WRDAOAc | - | + | Age of acquisition |

We summarize our results in Table 6. The remaining significant factors in the first column were originally obtained as a consolidation of a number of input variables through an FA dimension reduction process. By looking into the factor loading matrix in the FA output, we were able to identify which input variables were strongly connected with a specific latent factor and therefore interpret its conceptual meaning that led to our findings.

In our study, we analyzed commercial description texts. Commercial descriptions demand clarity, relevance, succinctness, informativeness, and more to facilitate a transaction. The ten factors reported in Table 6 revealed insights for a seller to make better decisions on describing their products to connect with US market consumers. Our data showed that a low-quality description, less native or non-native, displayed one or more of the identified characteristics. First, a low-quality description tends to display long texts and long sentences with higher sentence length variations. The lengthiness sometimes is a compromise for a less skilled writer to compensate for his/her inability to get to the point using the right structure or the most relevant selection of words. Research results by Yoshida et al. (2022) attested to clumsiness as a typical style of second language learners' writing. Succinctness in writing develops over time with repeated exposure to advanced compositions. A foreign seller may not obtain the needed frequency of exposure as their main focus would not be on language learning. Second, sentences tend to have simple structures as a non-native seller's writing skill is more likely to be at a primitive level. A long and simply structured sentence can be insufficient in linguistic sophistication and promotional efficacy. Communicating at a rudimentary level may suffice for basic daily usage. Unfortunately, sales persuasion demands a caliber of language sophistication that far exceeds basic daily chats. Simple structured sentences signal lower nativeness and worse composition quality. Third, repetition and overlapping of verbs, nouns, and pronouns can be prevalent in low-quality descriptions. We believe vocabulary deficiency may be the culprit of word reuse. Similarly, more frequent usage of words sharing the same stem can be another manifestation of vocabulary deficiency. Neither of these traits would be observed if the seller had a larger reservoir of vocabulary. The description would then benefit from a richer selection of words to reduce boredom and mechanicalness. Fourth, concept repetition and higher givenness may suggest weakness in expression precision where contents are repeated or re-explained to make up for known vagueness. A usual challenge for a non-native speaker is to precisely and succinctly explain a concept or convey an idea. The explanation is often unclear at the beginning and evolves to clarify after repetition and compensatory additions. Fifth, excessive word uniqueness, the extreme case of lexical diversity, makes a description very difficult to process. Achieving a balance between the repetition of words and polarizing word uniqueness is a must. This phenomenon can be understood as a result of over-compensation. A non-native speaker, due to his/her unfamiliarity or unskillfulness with the language, may resort to dictionaries to replace their simple words with more advanced words. Conflating quantity with quality is not an absolute sin, but the description would not read well. Sixth, a high ratio between causal particle to causal verb, a high value of minimum editorial distance, and a high density of infinitives and verb phrases are observed patterns from our datasets to watch out for in drafting product descriptions. We do not have an intuitive explanation for these data-driven patterns. These present themselves as the artifacts of analytical methods. Seventh, word selection sophistication is a direct indicator of description quality. Selecting concrete and simple words is convenient yet it may not meet linguistic bars set for marketing and promotion purposes. A promotional message that inspires interest, affinity, and sometimes impulsiveness demands skillfulness and better mastery of a language.

High-quality descriptions, meaning more native quality composition, displayed some interesting characteristics. Our dataset was not abundant in high-quality descriptions, which limited our model's ability to decipher lexical patterns embedded in high-quality descriptions. The first set of high-quality indicators includes cultural familiarity and content affiliation embedded in oral and daily conversations. Both metrics essentially emphasize immersion in a cultural environment which leads to familiarity with idiomatic words, phrases, and usages. The third metric of high-quality descriptions captures word maturity. Later age of acquisition surely relates to a higher level of education and sophistication. Staying away from words used often by kids and young students is expectedly correlated with composition styles of high-quality descriptions.

### 5.2. Research Contributions

Gao et al. (2012) concluded that higher quality information provided buyers with better assurance of seller qualifications, experience, and intention. Information quality in a competitive market is a self-selection process (Stiglitz, 1983) that signals seller credibility to consumers. User/consumer-generated content (UGC), such as online reviews, has been long studied by existing information systems, marketing, and decision science literature (Zhou et al, 2008; Biswas et al., 2022). UGCs could be good third-party signals to verify the seller's credibility (Akerlof, 1970). Marketer-generated content (MGC) so far has been less often studied. Conventional wisdom would question a seller's honesty in information disclosure, such as MGC. Subtleties in disclosed information, regardless of seller honesty, can be further analyzed to reveal hidden cues related to seller qualification, experience, and more. Hidden cues in MGC, maybe less ubiquitous as UGSs are, consciously or subconsciously solicit trust from consumers. Going beyond traditionally researched topics of MGCs, such as website design, user interfaces, and more, this research took a novel perspective to analyze the text quality of a description. It is in this sense that this research enriched our existing understanding of online trust-building processes by bringing in unrecognized patterns/factors hidden in description texts. To the best of our knowledge, this is the first research that endeavored to study description quality, defined as perceived composition nativeness, in online trust research.

This research sheds new light on consumer and business decision-making. Buying is an aggregated outcome of interactions among many decisions influencing factors. Understanding these decision-influencing factors and how they may lead to end outcomes is of paramount importance to both business entities and consumers. Liang and Lai (2002) mapped consumer purchase experience into a six-stage process, namely, problem recognition, information search, evaluation, choice, transaction, and post-sales services. Product descriptions directly and indirectly interact with five of the six stages. In the problem recognition stage, a consumer reads product descriptions to match solutions for his/her product needs. A big chunk of the task of information search involves analyzing and digesting product descriptions. Evaluation and choices cannot be successfully conducted without product descriptions. Even post-sales services can be impacted by product descriptions as low-quality descriptions can either damper usage impressions or even trigger a return. Product descriptions' relevance to the entire consumer decision process cannot be overstated. Given the description's low cost and easy accessibility nature (Gursoy, 2019; Chung & Sarnikar, 2021), our analysis of description quality adds new assets to existing theories of consumer purchasing decisions.

On the selling side, a business entity draws practical yet easily implementable insights from the ten factors in our findings. Our findings went beyond frequency-based textual patterns and presented higher-level lexical regularities inherent to description quality. The ten quality defining lexical characteristics discovered in this research have not been discussed in extant literature. A business equipped with our findings can systematically improve its decision efficacy from dependency on experience and intuition to the construction of assessment and composition systems for quality evaluation, problem identification, solution recommendation, and concept execution.

This research is cross-disciplinary in nature. Most linguistic research developed theories using academic text corpuses, such as student term papers and/or research publications. Their findings have been informative references for similar considerations in a business environment. Nevertheless, theories developed in academic settings are not directly transferable to business applications. Even small variations in linguistic research's experimental design require a thorough reexamination of result validity (Diederich, 1974), not to mention drastic context changes from academic to business settings. This research applies existing linguistic metrics to business compositions. Our findings are categorically more applicable and unique to business environments and their applications. Our contributions lie in our subject of analysis and our context-specific findings. Research like Biswas et al. (2022) and Lee et al. (2017) used text mining methods to analyze online review texts, not product descriptions. Their subject of analysis pertained to UGC studies. Liang et al. (2020) and Chung and Sarnikar (2021) analyzed product descriptions, yet their analysis was purely data-driven with little relevance to linguistic theories. Pryzant et al. (2017) aimed to build a link between product description and financial outcomes where a very limited set of linguistic metrics were included. The comprehensiveness of their theoretical relevance was weaker in relativity to our work that assessed the entire one hundred and six linguistic metrics in Coh-Metrix. We painted a much more holistic picture for a strategic understanding of metric applicability in defining and assessing description quality. Our findings built theories specific to business applications, where existing linguistic conclusions were confirmed, altered, or rejected for their usability in business decision-making.

### 5.3. Managerial implications and limitations

Picturing yourself in the shoes of the Shenzhen manufacturer mentioned at the beginning, our findings offer executable solutions squarely to address their very problem of "But really, we don't know what "good" looks like." Sellers from non-English speaking countries are almost one million strong on Amazon.com (Bryant, 2023). The rise of Walmart's online platform and the recent progress of traditional online playgrounds like eBay.com all amass sellers from around the world. A seller who strives to better its information quality now has clear instructions to work on ten dimensions of descriptions. Our results potentially level the playground by shrinking the quality discrepancy between native English sellers and non-English sellers. A non-native seller is endowed with more strategic options to fulfill their desires for brand building and better presenting their products. Maybe our results are not adequate for sellers to break out of the vicious cycle of endless price competition. However, this research endeavored to break open a refreshing crack for sellers regardless of their nationalities or language capacities. Consumers like the two partners running the consulting company can better comprehend product descriptions for purchasing decisions. A larger set of sellers meeting consumer expectations on information quality shall enhance market structure competitiveness and improve consumer surplus. Simply put, consumers will be less confused by product descriptions and start to enjoy the availability of more seller choices. Platforms like Amazon.com may use our results to construct a description quality score that automates quality assessment and even moves towards an expert system for quality improvement suggestions.

Our research is not without limitations. As the first research to analyze description quality using linguistic theories, our research is more exploratory in nature. Our sample size can be expanded to include more product varieties and more description data points. Description samples from other ecommerce platforms may provide extra sentiments on our results. Our linguistic analysis was conducted using Coh-Metrix, which was a prevailing tool for text analysis. While Coh-Metrix is linguistically comprehensive, other text analysis tools do exist and offer their own merits. More text analysis metrics from other tools shall help enrich our linguistic relevance and provide further dimensions of linguistic properties. Another issue that shall be better discussed is the ultimate question of whether description quality matters. The work by Pryzant et al. (2017) was a good start. More outcome variables can be linked with a larger set of description quality influencing factors to further validate the links between description quality and business performance.

**Acknowledgment**

We are grateful to Mr. Woody Hunt for his invaluable support of our research. His generosity in granting our team access to professional Amazon.com market research through his company's licenses has been instrumental to our progress. This access has provided us with crucial insights that would not have been possible otherwise, and we deeply appreciate his commitment to advancing research and knowledge in this field.

We also would like to thank our research assistants, Amaranta A. Quintero, Diana Sosa, Gosia Hoot, Jonathan Xiang, and Austin Meng. Their support and hard work were truly invaluable. We couldn't have asked for a better team!

## REFERENCES

Akerlof, G. A. (1970). The market for "Lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, *84*(3), 488-500.

Armah, G. K., Luo, G., & Qin, K. (2014). A deep analysis of the precision formula for imbalanced class distribution. *International Journal of Machine Learning and Computing*, *4*(5), 417-422.

Babić Rosario, A., Sotgiu, F., De Valck, K., & Bijmolt, T. H. (2016). The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *Journal of marketing research*, *53*(3), 297-318.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*. Cornell University.

Bardel, C., Lindqvist, C., & Laufer, B. (2013). *L2 vocabulary acquisition, knowledge and use*. Lulu. com.

Biswas, B., Sengupta, P., Kumar, A., Delen, D., & Gupta, S. (2022). A critical assessment of consumer reviews: A hybrid NLP-based methodology. *Decision support systems*, *159*, 1-13.

Bringmann, K., Grandoni, F., Saha, B., & Williams, V. V. (2019). Truly subcubic algorithms for language edit distance and RNA folding via fast bounded-difference min-plus product. *SIAM Journal on Computing*, *48*(2), 481-512.

Bryant, D. (2023, August 2). *How Chinese Sellers are Manipulating Amazon in 2024*. Ecomcrew. Retrieved from https://www.ecomcrew.com/chinese-sellers-manipulating-amazon/

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., ... & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*. Cornell University.

Chambers, C. P., & Echenique, F. (2016). *Revealed preference theory* (Vol. 56). Cambridge University Press.

Chocarro, R., Cortiñas, M., & Villanueva, A. (2022). Attention to product images in an online retailing store: An eye-tracking study considering consumer goals and type of product. *Journal of Electronic Commerce Research*, *23*(4), 257-281.

Chung, Y., & Sarnikar, S. (2022). Understanding host marketing strategies on Airbnb and their impact on listing performance: A text analytics approach. *Information Technology & People*, *35*(7), 2075-2097.

Crossley, S. A., & McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, *20*(4), 271-285.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15-30.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 37-54.

Diederich, P. B. (1974). *Measuring growth in English*. National Council of Teachers English.

Dimoka, A., Hong, Y., & Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence. *MIS quarterly*, 395-426.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of second language writing*, *4*(2), 139-155.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, *25*(2-3), 285-307.

Gao, J., Zhang, C., Wang, K., & Ba, S. (2012). Understanding online purchase decision making: The effects of unconscious thought, information quality, and information quantity. *Decision Support Systems*, *53*(4), 772-781.

Gardner, D. (2013). *Exploring vocabulary: Language in action*. Routledge.
Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS quarterly*, 51-90.
Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, *21*(2), 203-226.
Gursoy, D. (2019). A critical review of determinants of information search behavior and utilization of online reviews in decision making process (invited paper for 'luminaries' special issue of International Journal of Hospitality Management). *International Journal of Hospitality Management*, *76*, 53-60.
Hollister, S. (2021, September 18). *Amazon says it's permanently banned 600 Chinese brands for review fraud*. The Verge. Retrieved from https://www.theverge.com/2021/9/17/22680269/amazon-ban-chinese-brands-review-abuse-fraud-policy
Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision support systems*, *44*(2), 544-564.
Kim, E. G., & Chun, S. H. (2019). Analyzing online car reviews using text mining. *Sustainability*, *11*(6), 1611.
Kim, Y. S. G., Wolters, A., Mercado, J., & Quinn, J. (2022). Crosslinguistic transfer of higher order cognitive skills and their roles in writing for English-Spanish dual language learners. *Journal of educational psychology*, *114*(1), 1.
Koh, B., & Cui, F. (2022). An exploration of the relation between the visual attributes of thumbnails and the view-through of videos: The case of branded video content. *Decision Support Systems*, *160*, 1-13.
Kohro, Y. (2009). A contrastive study between L1 and L2 compositions: Focusing on global text structure, composition quality, and variables in L2 writing. *Dialogue*, *8*(2009), 1-19.
Laufer, B. (2012). Vocabulary and writing. *The encyclopedia of applied linguistics*, 1-5.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307-322.
Lee, M., Jeong, M., & Lee, J. (2017). Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach. *International Journal of Contemporary Hospitality Management*, *29*(2), 762-783.
Li, M., & Pham, Q. N. (2022). Three heads are better than one? Digital multimodal composition completed collaboratively versus individually. *Language Teaching Research*, 13621688221102536.
Li, X., Wu, C., & Mai, F. (2019). The effect of online reviews on product sales: A joint sentiment-topic analysis. *Information & Management*, *56*(2), 172-184.
Liang, S., Schuckert, M., Law, R., & Chen, C. C. (2020). The importance of marketer-generated content to peer-to-peer property rental platforms: evidence from Airbnb. *International Journal of Hospitality Management*, *84*, 102329.
Liang, T. P., & Lai, H. J. (2002). Effect of store design on consumer purchases: an empirical study of on-line bookstores. *Information & management*, *39*(6), 431-444.
Lin, J., Ma, Z., Gomez, R., Nakamura, K., He, B., & Li, G. (2020). A review on interactive reinforcement learning from human social feedback. *IEEE Access*, *8*, 120757-120765.
Liu, A. X., Steenkamp, J. B. E., & Zhang, J. (2018). Agglomeration as a driver of the volume of electronic word of mouth in the restaurant industry. *Journal of Marketing Research*, *55*(4), 507-523.
Louis, A. P. (2013). *Predicting text quality: Metrics for content, organization and reader interest*. University of Pennsylvania.
Louis, A. P. (2013). *Predicting text quality: metrics for content, organization and reader interest*. University of Pennsylvania.
Louwerse, M. M. (2004). Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities*, *38*, 207-221.
Lu, X., & Chen, Y. (2020). Situations matter: Understanding how individual browsing situation routineness impacts online users' advertisement clicks behavior. *Journal of Electronic Commerce Research*, *21*(2), 113-129.
McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
Mun, Y. Y., Yoon, J. J., Davis, J. M., & Lee, T. (2013). Untangling the antecedents of initial trust in Web-based health information: The roles of argument quality, source expertise, and user perceptions of information quality and risk. *Decision support systems*, 55(1), 284-295.
Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research*, *6*(1), 58-73.
Nordquist, R. (2020, February 5). *What Are Causative Verbs?* ThoughtCo. Retrieved from https://www.thoughtco.com/what-is-causative-verb-1689833

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, *35*, 27730-27744.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language learning*, *47*(1), 101-143.

Pryzant, R., Chung, Y., Jurafsky, D., & Britz, D. (2017). JESC: Japanese-English subtitle corpus. *arXiv preprint arXiv:1710.10639*. Cornell University.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, *111*(515), 988-1003.

Sabatini, J., Albro, E., & O'Reilly, T. (2012). *Measuring up: Advances in how we assess reading ability* (Vol. 83). R&L Education.

Schwarm, S. E., & Ostendorf, M. (2005, June). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics* (pp. 523-530).

Spence, M. (1978). Job market signaling. In *Uncertainty in economics* (pp. 281-306). Academic Press.

Stiglitz, J. E. (1983). Risk, incentives and insurance: The pure theory of moral hazard. *The Geneva papers on risk and insurance-issues and practice*, 8, 4-33.

McEnery, T., & Wilson, A. (2001). Corpus Linguistics: An Introduction. Edinburgh University Press. http://www.jstor.org/stable/10.3366/j.ctvxcrjmp

Wirth, C., Akrour, R., Neumann, G., & Fürnkranz, J. (2017). A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136), 1-46.

Yoo, C. W., Sanders, G. L., & Moon, J. (2013). Exploring the effect of e-WOM participation on e-Loyalty in e-commerce. *Decision Support Systems*, *55*(3), 669-678.

Yoshida, A., Teranishi, M., Nishihara, T., & Nasu, M. (2022). The impact of L1 on L2: A qualitative stylistic analysis of EFL learners’ writings. In *Pedagogical Stylistics in the 21st Century* (pp. 343-369). Cham: Springer International Publishing.

Yu, X. (2018). Analyses and comparisons of three lexical features in native and nonnative academic English writing. *Electronic Theses and Dissertations*, 6061.

Zhou, L., Dai, L., & Zhang, D. (2007). Online shopping acceptance model-A critical survey of consumer factors in online shopping. *Journal of Electronic commerce research*, *8*(1).

Zhou, M., Dresner, M., & Windle, R. J. (2008). Online reputation systems: Design and strategic practices. *Decision support systems*, *44*(4), 785-797.