Differential Contributions of Machine Learning and Statistical Analysis to Language and Cognitive Sciences

Kun Sun*[1, 3] and Rong Wang[2]

[1]The Department of Linguistics, University of Tübingen, Germany
[2]The Institute of Natural Language Processing, Stuttgart University, Stuttgart, Germany
[3]The School of Foreign Languages, Fudan University, Shanghai, China

Abstract

Data-driven approaches have revolutionized scientific research. Machine learning and statistical analysis are commonly utilized in this type of research. Despite their widespread use, these methodologies differ significantly in their techniques and objectives. Few studies have utilized a consistent dataset to demonstrate these differences within the social sciences, particularly in language and cognitive sciences. This study leverages the Buckeye Speech Corpus to illustrate how both machine learning and statistical analysis are applied in data-driven research to obtain distinct insights. This study significantly enhances our understanding of the diverse approaches employed in data-driven strategies.

Keywords: data-driven approaches, features, objectives, mixture,

*email: kun.sun@uni-tuebingen.de

1 Introduction

Advancements in technology have revolutionized the ability of scientists to gather and analyze data on an unprecedented scale. This surge in data availability has led to the necessity for more sophisticated analytical techniques and computational tools, fundamentally changing the way research is conducted and facilitating new discoveries across various domains (Langley, 1981; Han et al., 1993; Kitchin, 2014; Montáns et al., 2019; Jack et al., 2018; Núñez et al., 2019).

Data technologies serve as powerful illuminators, revealing hidden patterns and insights within vast datasets, much like a flashlight cutting through darkness. For example, machine learning has become indispensable in scientific inquiry, allowing researchers to decipher complex patterns, enhance their understanding of intricate phenomena, and make precise predictions. This technology is applied across diverse fields including medicine, astronomy, genomics, social sciences, and environmental studies, enabling researchers to swiftly extract significant insights from large volumes of data, much faster than traditional methods would permit (Mjolsness and DeCoste, 2001; Jordan and Mitchell, 2015; LeCun et al., 2015; Webb et al., 2018). Such acceleration not only deepens the exploration of research questions but also unveils knowledge that was once concealed or unreachable. Furthermore, statistical methods play a critical role in the modern landscape of scientific research. They are integral in every phase of a study, from planning and design to data collection, analysis, and the interpretation and reporting of results (Carleo et al., 2019; Núñez et al., 2019; Butler et al., 2018; Grimmer et al., 2021). While machine learning helps in identifying patterns that may elude human detection in massive datasets, statistical techniques continue to enrich scientific research by providing robust frameworks for making inferences and validating findings (Fisher, 1955; Box, 1976; Nelder, 1986; Nosek et al., 2012). Together, these tools are transforming scientific paradigms, propelling forward our capacity to understand and manipulate the natural world.

Data-driven research has made significant strides across various fields, bringing to the forefront disciplines such as data science, statistics, machine learning, and deep learning (Solomatine and Ostfeld, 2008; Wolf, 2010; Miller and Goodchild, 2015; Zhang et al., 2011). Among these, machine learning and statistics form the foundational core. While often conflated due to their shared capability to process data, machine learning and statistics frequently overlap, creating ambiguity around their distinct roles and unique contributions. Moreover, both machine learning and statistics take advantage of the shared knowledge of probability theory, linear algebra etc. Many researchers have found it challenging to distinguish clearly between these two areas. Although both fields are fundamentally concerned with extracting knowledge from data, they differ significantly in their approaches and objectives.

The rapid advancement of statistics in the early 20th century significantly enhanced scientists' ability to quantitatively evaluate hypotheses such as "does

treatment X affect outcome Y?". These inquiries often required definitive, binary answers, leading to the development of statistical tools focused primarily on validating binary propositions through hypothesis testing and generating confidence intervals to estimate the magnitude and uncertainty of observed effects. Contrastingly, machine learning seemingly emerged from the field of engineering, driven by the ambition to enhance machine functionality. Initially, machine learning aimed to augment machine capabilities, with a secondary focus on comprehending intelligence itself. In this discipline, verifying real-world truths ranked lower in priority, considered relevant mostly as components of intelligent behaviors, a point still open to debate.

While both statistics and machine learning are interconnected, they hold distinct identities, as evidenced by various studies. Gaining a clear understanding of these differences is crucial for mastering both fields and effectively applying data-driven strategies. Contemporary research methodologies frequently incorporate advanced techniques such as machine learning and statistical analysis, often leveraging both to harness the strengths of each approach. Despite the prevalent use of these methods, there remains a notable gap in studies that systematically apply both techniques to identical datasets to highlight potential differences in outcomes. This comparative approach is rarely employed. However, it could provide valuable insights into how each method processes and interprets the same dataset differently. Moreover, the specific contributions of these techniques to various academic fields have not been extensively explored. In particular, the fields of language sciences, cognitive research, and social sciences could greatly benefit from targeted studies examining how machine learning and statistical analysis can uniquely advance our understanding of complex phenomena within these disciplines. Such investigations could elucidate the distinct advantages or limitations of each method, potentially leading to more refined and effective research methodologies in these areas.

The current study aims to bridge these gaps by employing both statistical analysis and machine learning on the same dataset to achieve different objectives. This study is intended to provide fresh perspectives to researchers in language sciences and cognitive studies, illustrating distinct applications of each method to extract unique insights.

2 Relation and differences between statistics and machine learning

Before starting our demonstration, we should systematically have a detailed account of how statistics and machine learning are related and differ from each other in theoretical perspectives. After having an understanding of these differences on theory, we believe that the practical implementation abilities will be enhanced, and this is our purpose of this paper.

2.1 How are statistics and machine learning related?

Breiman (2001b) highlighted the divergent approaches to data analysis, emphasizing the ongoing evolution of these methodologies over the decades. Many machine learning techniques have their roots in traditional statistical methods, such as linear and logistic regression, while also drawing from disciplines like calculus, linear algebra, and computer science. This intersection has led some to mistakenly merge the concepts of machine learning and statistics. Moreover, the introduction of user-friendly machine learning packages, such as Python's `scikit-learn` (Pedregosa et al., 2011), has further abstracted machine learning from its statistical foundation. This has propagated a belief among some newcomers to the field that a deep understanding of statistics isn't necessary for machine learning applications. While basic tasks might not require intensive statistical knowledge, advanced modeling and the development of new algorithms heavily rely on a solid grounding in statistics and probability theory.

Statistical learning theory, formulated in the 1960s, lays the theoretical foundation for machine learning. It introduces concepts such as the hypothesis space and loss functions, providing a framework for supervised learning. The theory defines a dataset: Let S be the set of all pairs (x_i, y_i), where x_i and y_i are elements related by some function or condition. We then define S as $S = \{(x_i, y_i)\}$, comprising n data points, where each data point consists of features x and a corresponding output y. The objective in statistical learning is to discover the function that maps the input features to the output, navigating through a hypothesis space of potential functions, guided by the minimization of a loss function which evaluates the expected risk over the dataset.

2.2 How are two fields different?

Despite the original intentions to remain separate, the core components of machine learning algorithms often rely on statistical principles that have been examined by statisticians for over a century. These mathematical principles apply universally, indifferent to whether the objective is to achieve artificial intelligence, publish research, or develop unbiased estimators. Consequently, many pressing questions in machine learning are, at their heart, statistical challenges previously unexplored by mainstream statistics. In recent decades, the field of statistics has found itself both challenged and invigorated by the successes of machine learning, particularly in areas traditionally dominated by statistics, such as predictive modeling (L'heureux et al., 2017; Saidulu and Sasikala, 2017; Ratner, 2017; Rudin et al., 2022; Ratner, 2017; Ley et al., 2022). This has sparked vigorous efforts to merge the theories and tools of both fields. However, for these efforts to be successful, they must first recognize and address the underlying reasons for their differences (Bzdok et al., 2018; Makridakis et al., 2018; Boulesteix and Schmid, 2014). Understanding these differences is essential for leveraging the strengths of both fields effectively, whether the goal is to uncover deep insights from data or to develop

robust predictive algorithms.

One of the fundamental distinctions between statistics and machine learning is their purpose. Statistics aims to infer properties about a population through samples, focusing on understanding and describing data relationships. In contrast, machine learning seeks to predict outcomes based on patterns identified in data, often using large and complex datasets to train predictive models. This training process typically involves dividing data into subsets for training, validation, and testing, which helps refine the models for better accuracy.

Another significant difference lies in how data is approached. In statistics, the focus is on the quality of data and the validity of the conclusions drawn from it through significance testing, which considers the presence of noise and potential confounding variables. Machine learning, however, emphasizes the quantity of data, often requiring large datasets to achieve the accuracy needed for effective predictions.

Interpretability also varies greatly between the two. Statistical models, often simpler and based on fewer variables, tend to be more interpretable. This clarity comes from the use of statistical significance tests that validate the relationships within the data. Machine learning models, in contrast, can become highly complex, especially with the inclusion of many variables, making them accurate yet sometimes difficult to decipher. This complexity can render machine learning models as "black boxes" (Adadi and Berrada, 2018; Gilpin et al., 2018; Linardatos et al., 2020), where it is challenging to trace how inputs are transformed into outputs.

2.3 An example

Linear regression is fundamentally a statistical method designed to minimize the squared error between data points. However, it is also widely used in machine learning for predictive tasks. An example of employing regression models is taken to illustrate different intentions and outcomes in language sciences.

In language sciences, regression models can be particularly illuminative. For example, researchers might use regression analysis to explore how the complexity of syntactic structures in children's language development correlates with cognitive development indicators. Statistically, this would involve collecting data on children's language use and cognitive tests, then applying linear regression to understand and quantify how changes in one variable relate to changes in another, typically across a complete dataset. The statistical approach focuses on inference—determining whether there is a statistically significant relationship between syntactic complexity and cognitive development. This method does not necessitate training and testing subsets of data; instead, it aims to characterize the relationship across the entire data set, assessing the significance and reliability of the observed relationships.

Conversely, in a machine learning context, the same regression model could be employed to predict cognitive developmental outcomes based on the existing

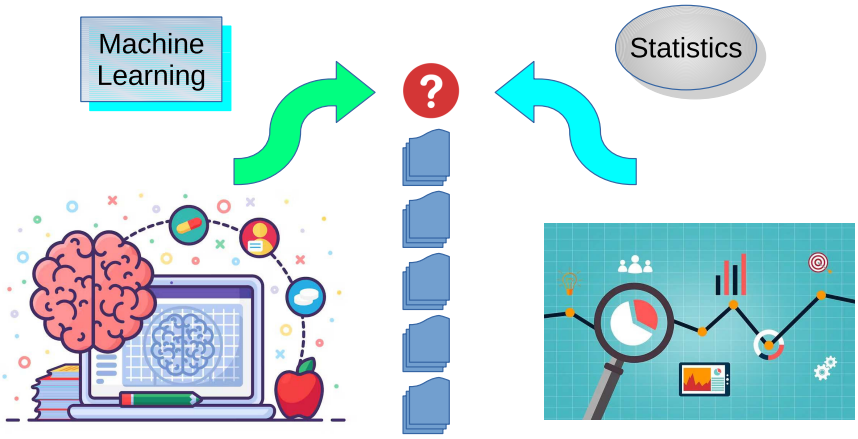

Figure 1: Differences between machine learning and statistics

linguistic data. The model would be trained on a subset of data, with the model's parameters adjusted to optimize performance as measured on a separate test set. The objective is less about understanding the underlying dynamics of the relationship and more about achieving high predictive accuracy.

Suppose a language scientist wants to predict future cognitive development based on early linguistic behaviors using machine learning. In that case, they might employ a regression model on a divided dataset (training and testing), focusing on how well the model predicts developmental outcomes in new, unseen data. This approach would likely prioritize predictive power over interpretability, with model adjustments driven by performance metrics on the test set. Thus, while both fields use regression models, the context and objectives dictate their implementation: statistical modeling seeks to uncover and explain relationships within the data, providing comprehensive insights into language development patterns. In contrast, machine learning leverages these models primarily for their predictive capabilities, often in applications that require rapid assessments of developmental trajectories based on linguistic inputs.

These differences underscore the importance of distinguishing between the purposes and methodologies of statistics and machine learning in language sciences, ensuring that the chosen approach aligns with the specific goals of the research or application.

2.4 Key differences

In short, the contrast between machine learning and statistics is rooted in their different approaches and priorities. These key differences are summarized as shown in Table 1 and Fig. 1. The following provides some details on these key differences.

Machine learning is primarily focused on achieving high predictive accuracy and is comfortable with models that are effective but not easily interpretable, often referred to as 'black box' models. On the other hand, statistics places a greater emphasis on relation among variables in data. Statisticians value models that are transparent and explainable, reflecting the field's deep roots in mathematics and science, where theoretical foundations and provable properties are paramount. This includes a strong focus on the behavior of estimates as sample sizes increase and ensuring that models are robust even with small datasets.

Moreover, statistical methods rigorously address the bias and variance of estimates, seeking models that not only perform well but also provide insights into the certainty and reliability of predictions. In contrast, while machine learning does consider these factors, the emphasis is more on how the model performs in practical scenarios, often prioritizing larger datasets that feed the algorithmic complexity necessary for modern applications.

Table 1: Key Differences between Machine Learning and Statistics

Characteristic	**Machine Learning**	**Statistics**
predictive accuracy	High	Lower
black box models	High	Low
relation among variables	Low	High
various asymptotic properties	Low	High
provable characteristics and bounds	Low	High
bias and variance of estimates	Less Concerned	Highly Concerned

3 Experiments

After having a systematic understanding the differences between statistics and machine learning, we carried out the experiments to demonstrate how machine learning and statistics differently contribute to linguistic and cognitive research.

3.1 Materials & methods

The Buckeye Corpus of conversational speech contains high-quality recordings from 40 speakers in Columbus OH conversing freely with an interviewer (Pitt et al., 2005). The speech has been orthographically transcribed and phonetically labeled. The sessions were conducted as sociolinguistics interviews, and are essentially monologues. The speech has been orthographically transcribed and phonetically labeled. The corpus includes 357908 words.

The original dataset from the Buckeye provided diverse factors: speaker's gender (female, male), age (old, young), word duration, etc. However, we can add more factors such as PoS tag for each word, word length, word frequency, the

phrase rate, deletions and semantic relevance. Most of these factors have been investigate to show that they are closely related with word duration (Cowan et al., 1997; Baker and Bradlow, 2009; Terroir and Lavandier, 2014; Cohen Priva, 2015; Pierrehumbert, 2016; Bürki, 2018). However, these relevant research mostly employed statistical correlation and simply linear regression to explore how these factors affect word duration in speech. We adopt machine learning methods and some advanced mixed-effect regression models to examine them in the current study. The specific information on these factors is detailed in the Table 2:

Table 2: Factors and Their Descriptions

Factor	**Description**
Word Duration	The time to articulate a target word in spontaneous speech.
Word Length	The number of alphabets in a target word.
Word Frequency (log)	(Logarithm of) the normalized frequency of the word in the subtitle corpus [1].
CiteLength	The number of syllables in transcript phonetic form of this word.
PhraseRate	[Word number in this phrase] / [duration from the beginning of the phrase to the end of the phrase in a target word].
Deletion	The number of segments in a target word deleted or reduced.
Semantic Relevance	The semantic relatedness degree with the context.
Speaker/Sex/Age	Speakers in corpus; Female vs. male; young vs. old.

Additionally, semantic relevance is a novel measures, representing how a target word is semantically related with the context. It measures the semantic degree of how the contextual information influence the target word (Sun et al., 2023; Sun, 2023).

3.2 Machine learning methods

Machine learning can be categorized mainly into three types: supervised learning, where models predict an outcome based on input data; unsupervised learning, where models identify patterns and relationships in data without any specific outcome to predict; and reinforcement learning, where an agent learns to make decisions by receiving rewards for actions (Bishop and Nasrabadi, 2006; Hastie et al., 2009; Murphy, 2012; Harrington, 2012; Raschka and Mirjalili, 2019).

Due to the features of our data, we introduced supervised machine learning models to recognize patterns in the new data. The main purpose of ML is to predict word speech duration. To reduce the training load, we classify the data of word duration into eight ranges. Originally there are 357908 words, and their durations differ from each other, in other words, there are 357908 different values. However, we classified 357908 values into eight ranges. Put it simply, our trained LM models were required to predict which one belongs to a specific range among the eight ones. The data used for training the ML models include age, gender, word length, word frequency. The dataset is divided by 75% as the training one, and the remaining 25% as testing dataset.

3.2.1 Random forest

Random Forest is an ensemble machine learning algorithm that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forests correct for decision trees' habit of overfitting to their training set (Breiman, 2001a). During the training phase, random forest creates multiple decision trees. Each tree is trained on a random subset of the data samples, and at each node, a subset of features is randomly chosen to determine the split. For predictions, each tree in the forest votes, and the final class assigned to a sample is based on the majority vote for classification tasks, or an average in case of regression. This aggregation helps to mitigate errors from individual trees and exploit the strength of multiple learners. Random Forest's performance can typically be assessed using standard metrics such as accuracy for classification tasks. Since the model uses multiple trees, it is usually more robust to overfitting compared to a single decision tree. The final accuracy result is about 51%. The following code is used to implement the task of the classification of different classes on durations.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
from sklearn import tree

df = pd.read_csv("buckeye1.csv", delimiter='\t')

# Define the breaks for the ranges of 'Duration'
breaks = np.quantile(df['Duration'], np.arange(0, 1,
   0.2))
```

```
labels = [‘1’, ‘2’, ‘3’, ‘4’, ‘5’]  # Number of labels
    matches number of quantile edges minus one

# Cut the ‘Duration’ variable into ranges and assign
    labels
df[‘range_label’] = pd.cut(df[‘Duration’], bins=breaks,
    labels=labels, include_lowest=True)

df.dropna(subset=[‘range_label’], inplace=True)

# Feature selection
feature_names = [‘CiteLength’, ‘PhraseLength’,  ‘
    Deletions’, ‘wordlen’, ‘logfreq’]
X = df[feature_names]
y = df[‘range_label’]

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y
    , test_size=0.25, random_state=42)

# Create and train the Random Forest Classifier
clf = RandomForestClassifier(n_estimators=100,
    random_state=42)
clf.fit(X_train, y_train)

# Make predictions
preds_rf = clf.predict(X_test)
accuracy = accuracy_score(y_test, preds_rf)
print(‘‘Accuracy:", accuracy)
#accuracy 51.02712%
```

We selected five estimators from 100 ones to visualize them, as shown in Fig. 2. RandomForest model using Python's `matplotlib.pyplot` and `sklearn.tree` libraries. This kind of visualization can be particularly useful for understanding how individual trees in the ensemble make decisions, which can provide insights into the model's operation.

However, we can streamline the factors in our training dataset. For instance, by selecting only 'WordLength' and 'WordFrequency', the prediction accuracy achieved was 50.75%. Conversely, using 'CiteLength' and 'PhraseRate' as factors, the prediction accuracy dropped to 37.13%. Furthermore, combining 'Deletions' with 'WordLength' yielded a prediction accuracy of 42.36%. By experimenting with various factor combinations, we can deduce the impact of different factors on machine learning accuracy. Evidently, 'WordLength' and 'WordFrequency' appear to play more significant roles in enhancing machine learning performance.

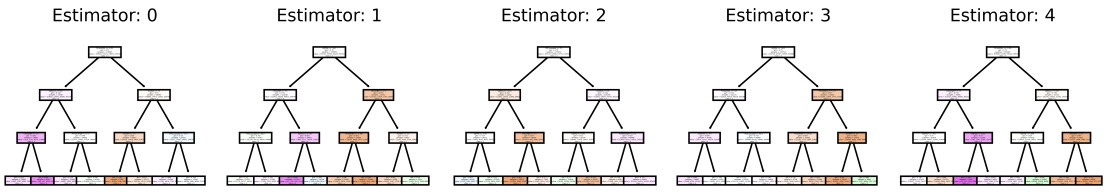

Figure 2: The five decision trees in the random forest

3.2.2 Support vector machine

Support vector machine (SVM): SVM is a powerful and versatile supervised machine learning model, particularly well-suited for classification tasks (Hearst et al., 1998). It works by finding the hyperplane that best divides a dataset into classes with the maximum margin, i.e., the maximum distance between data points of both classes. SVMs are effective in high-dimensional spaces and relatively immune to overfitting, especially in cases where the number of dimensions exceeds the number of samples.

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

# Feature selection
X = df[['Age', 'Score', 'CiteLength', 'PhraseLength', '
   PhraseRate', 'Deletions']]
y = df['range_label']
X = X.dropna()
y = y.loc[X.index]

X_train, X_test, y_train, y_test = train_test_split(X, y
   , test_size=0.25, random_state=42)

# Create and train the Support Vector Machine Classifier
clf = SVC(kernel='linear', C=1.0, random_state=42)
clf.fit(X_train, y_train)

preds_svm = clf.predict(X_test)
accuracy = accuracy_score(y_test, preds_svm)
print(''Accuracy:", accuracy)
```

```
#accuracy 51.2334%
```

We used the SVC from the package "scikit-learn" to create a Support Vector Machine classifier. kernel='linear' specifies that we are using a linear kernel for the SVM. "C=1.0" is the regularization parameter, which controls the trade-off between maximizing the margin and minimizing the classification error. We fit the classifier to the training data using the fit() method. Then, we make predictions on the test data using the predict() method. Finally, we calculate the accuracy of the model using "accuracy_score()" from scikit-learn. We can adjust the kernel type (linear, rbf, poly, etc.) and other hyperparameters of the SVM as needed based on the specific requirements and the characteristics of the dataset. The final accuracy reaches about 51%, which is close to the one predicted by the random forest. The consistent accuracy results show that these factors play stable roles in machine learning. The classifier function from SVM is shown in Fig. 3. We selected four features to visualize how SVM helps classify them.

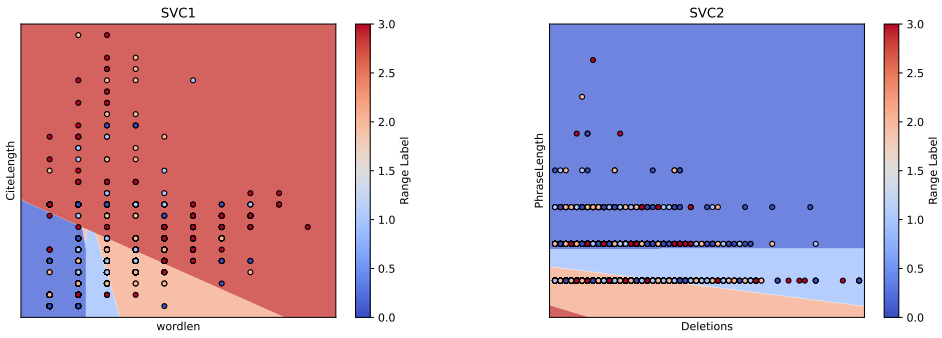

Figure 3: The two classifiers from the SVM. Each plot selected two factors as *x-axis* and *y-axis*.

Following the random forest strategies, we streamlined the factors in our training dataset. For example, using only 'WordLength' and 'WordFrequency,' we achieved a prediction accuracy of 50.92%. In contrast, CiteLength' and PhraseRate' resulted in a lower accuracy of 38.36%, while combining Deletions' with WordLength' reached 42.72%. Such results are quite close to those run in the random forest. This approach helped us understand that 'WordLength' and 'WordFrequency' are more critical for improving performance. In the following statistical analysis, we should treat both factors as control predictors.

3.3 Statistical analysis

Statistical analysis is broadly divided into two main types: descriptive statistics and inferential statistics. Descriptive statistics summarize and describe the fea-

tures of a dataset through metrics like mean, median, mode, and standard deviation, helping to visualize and understand data distributions. Inferential statistics, on the other hand, use samples of data to make generalizations or predictions about a larger population, employing techniques such as hypothesis testing, regression analysis, and confidence intervals. This distinction allows statisticians and researchers to both understand the data they have and to infer properties about data they do not have, supporting decision-making across fields like economics, medicine, engineering, and social sciences.

We first employed descriptive correlation to explore in the data, and then applied some advanced regression models to make further explorations. Machine learning needs to use training data to train models and do tests in the test data. In contrast, statistical analysis does not require to do this.

3.3.1 Correlation

The first statistical analysis is to explore the Pearson's correlationship among these factors we have applied. Using in-built R function, we could obtain the correlation matrix for these variables, and it highlights significant relationships among various features, as shown in Fig. 4. Specifically, a very strong positive correlation exists between "WordLength" and "CiteLength" (ρ = 0.926), indicating that as words get longer, they generally contain more syllables. This is complemented by strong negative correlations, such as between "WordLength" and "LogFrequency" (ρ = -0.662), where longer words appear less frequently. Similarly, "WordDuration" and "LogFrequency" show a moderate negative correlation (ρ = -0.589), suggesting that less frequent words tend to have longer durations. These patterns suggest a close link between the physical characteristics of words and their usage frequencies.

Further analysis shows relationships affecting phrase dynamics and overall speech patterns. Here we included more factors, such as PhraseLength (the alphabet number of phrase where the target word is located), SpeakerRate (the average speaking speed for this speaker) to show more divergent correlations. For instance, there is a moderate positive correlation between "PhraseRate" and "PhraseLength" (ρ = 0.249), indicating that longer phrases typically have a faster speech rate. Conversely, a moderate negative correlation between "SpeakerRate" and "PhraseRate" (ρ = -0.349) suggests that faster phrase rates correlate with slower overall speaking rates, potentially reflecting variations in speaking dynamics based on phrase complexity. These findings highlight the specific ways in which speech elements interact, providing valuable insights for studies in linguistic patterns and speech processing. Importantly, while these correlations reveal trends, they do not imply causation, and correlations above approximately 0.5 are particularly noteworthy in social science contexts.

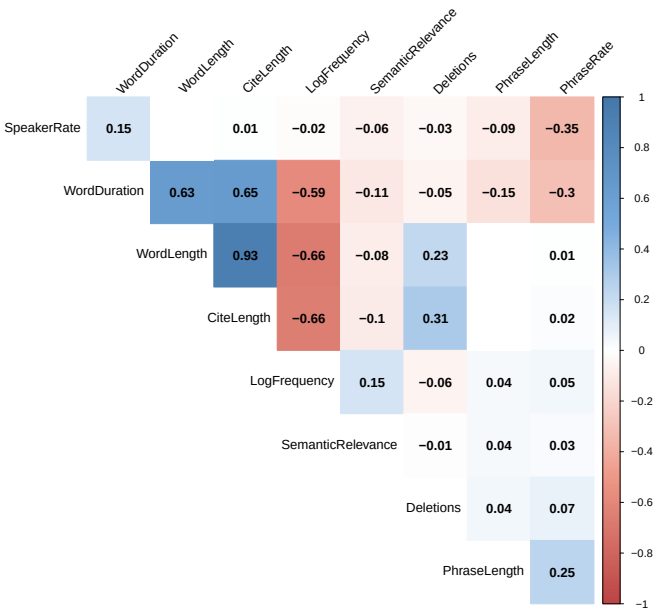

Figure 4: Correlations among various factors

3.3.2 Mixed-effect regression analysis

Data collection often involves variables such as the characteristics of speakers, age groups, and gender. Furthermore, the sample sizes available for these studies may be limited, posing challenges when attempting to apply complex models that include a multitude of parameters. Additionally, the assumption of data independence can be problematic. For instance, data may be gathered using quadrats within specific sites, creating a hierarchical data structure with quadrats nested inside the sites. Mixed models are specially designed to manage these types of intricate, structured datasets—even when faced with smaller sample sizes and a large number of variables. Notably, these models also provide the benefit of conserving degrees of freedom, an improvement over traditional linear models in handling such data scenarios. We employed two popular mixed-effect regression models to explore how different factors take effect on durations of words.

The first is Linear Mixed-Effects Regression (**LMER**) (Bates, 2010; Kuznetsova et al., 2017). LMER is a statistical model used to analyze data with both fixed effects, which represent the main variables of interest expected to have a consistent impact across the dataset, and random effects, which account for inherent variations from grouped or nested data sources. This model is particularly useful for dealing with hierarchical structures, such as students within classrooms, or data from repeated measures on the same subjects, allowing for more accurate estimates by accounting for both within-group and between-group variability. LMER helps in understanding complex datasets by modeling the dependencies and structure within the data, making it a powerful tool for robust statistical analysis. The present study employed LMER to explore how independent variables and random variable take effects on the dependent variables. The dependent variable is "word duration", and other independent variables include word length, word frequency, and the random variables (i.e. Speaker, Age and Sex). The data number (n) in the regression fittings is 262342.

The LMER fittings were summarized as follows, and we made comparison via AIC (Akaike Information Criterion). Control predictors such as word length and word frequency were included, and random variables such as age, sex and speaker could be included. almUnderstanding mixed-effect models is essential for appreciating the significance of these indicators, highlighting the Akaike Information Criterion (AIC) as the preferred tool for model comparison. A smaller AIC indicates better performance, as shown in Table 3.

- lm1=lmer(WordDuration∼WordLength+LogWordFreq+CiteLength+PhraseRate+(1|Sex)+(1|Speaker))
- lm2=lmer(WordDuration∼WordLength+LogWordFreq+CiteLength+SemanticRelevance+PhraseRate+Deletions+(1|Age)+(1|Sex)+(1|Speaker))
- lm3=lmer(WordDuration∼WordLength+LogWordFreq+CiteLength+SemanticRelevance+PhraseRate+Deletions+(1|Age)+(1|Speaker))
- lm4=lmer(WordDuration∼WordLength+LogWordFreq+CiteLength+SemanticRelevance+PhraseRate+(1|Age)+(1|Speaker))

- lm5=lmer(WordDuration∼WordLength+LogWordFreq+CiteLength+SemanticRelevance+PhraseRate+Deletions+(1|Sex)+(1|Speaker))

Table 3: LMER Model Comparison with AIC Values (n=262342)

Model	Degrees of Freedom (df)	AIC
lm1	8	-433759.4
lm2	11	-459522.0
lm3	10	-459523.8
lm4	9	-433952.1
lm5	10	-459523.8

First, most of factors listed were significant in these fittings. In other words, these factors significantly influenced word duration. Despite this, clearly, the smaller AIC indicates better performance. "lm3" / "lm5""has the smallest AIC, and this suggests that these factors and the random variable have significant effects on word duration. Compared with "lm3" and "lm4", we find that the factor "Deletion" did significantly contribute to AIC in the model. However, when the random effect "age" or "sex" or both appears, they have no great difference. Through comparison, we found that the random factor "sex" had no significant effect on word duration.

Next, we applied Generalized Additive Mixed Models (**GAMM**) in analyzing these factors in a similar way. GAMMs are an extension of Generalized Linear Mixed Models (GLMM), incorporating non-linear relationships between the dependent and independent variables through smooth functions (Wood, 2017). GAMMs allow for both fixed and random effects, accommodating complex variations within hierarchical data structures. The "additive" part of GAMM means that the model expresses the dependent variable as a sum of smooth functions of predictors, along with any random effects and an error term. This flexibility makes GAMMs particularly useful for modeling non-linear trends in data, where the effect of variables is not strictly linear and may vary by group or over time.

We still listed a number of GAMM fittings and made comparison by referring to AIC. The biggest differences between GAMM and LMER is that GAMM could leverage the function `s()` and interaction smooth `te()`. The smooth function better gets model fittings for some factors, and the interaction smooth could find the interaction among some given factors. The independent and dependent variables were set similarly as the ones in LMER. The AIC results are shown in Table 4.

- t1=bam(WordDuration∼s(WordLength)+s(LogWordFreq)+s(CiteLength)+s(SemanticRelevance)+s(PhraseRate)+s(Deletions)+s(Age, bs="re")+s(Speaker, bs="re"))
- t2=bam(WordDuration∼s(WordLength)+s(LogWordFreq)+s(CiteLength)+s(SemanticRelevance)+s(PhraseRate)+s(Deletions)+s(Sex, bs="re")+s(Speaker, bs="re"))

- t3=bam(WordDuration∼te(WordLength,LogWordFreq)+s(CiteLength)+s(SemanticRelevance)+ s(PhraseRate)+s(Deletions)+s(Speaker, bs="re"))
- t4=bam(WordDuration∼te(WordLength,LogWordFreq)+s(CiteLength)+s(PhraseRate)+s(Deletions)+ s(Sex, bs="re")+s(Speaker, bs="re"))
- t5=bam(WordDuration∼te(WordLength,LogWordFreq)+s(PhraseRate)+s(Deletions)+s(Sex, bs="re")+ s(Speaker, bs="re"))
- t6=bam(WordDuration∼s(WordLenght)+s(LogWordFreq)+s(CiteLength)+s(PhraseRate)+s(Deletions)+s(Sex, bs="re")+
s(Speaker, bs="re"))

Table 4: GAMM Model Comparison with AIC Values (n=262342)

Model	Degrees of Freedom (df)	AIC
t1	281.8724	-486185.2
t2	281.7350	-486185.1
t3	282.5285	-486017.5
t4	279.2962	-485994.0
t5	275.9720	-471297.8
t6	278.3133	-486159.7

Similarly, most of these factors are significant in these cases, which is consistent with the results in LMER. Random effect plays a crucial role in regression analysis (Baayen et al., 2017). Through the analysis, we found that the random variable "age" is not significant in "t1". T2 is the best fitting compared with other ones. Comparing "t2" and "t6", we could find that the factor "semantic relevance" substantially contributed to the model, namely, the factor "semantic relevance" had a significant impact on word duration.

To further understand the impacts of these factors, we visualized the partial effects of various predictors on word duration, as shown in Fig. 5. We found that word length and "CiteLength" are negatively correlated with word duration, indicating that as these factors increase, the word duration decreases. It is understandable that words with more syllables and letters could be spoken more slowly. Conversely, fast phrase rates and more deletions lead to shorter word durations. Word frequency and semantic relevance exhibit complex effects. When word frequency is low (less than 2), it is positively correlated with word duration; however, when word frequency exceeds 3, its impact turns negative. Similarly, when a word is less semantically related to the context (score less than 2), it negatively affects word duration. In contrast, high semantic relevance (score greater than 2) positively influences word durations.

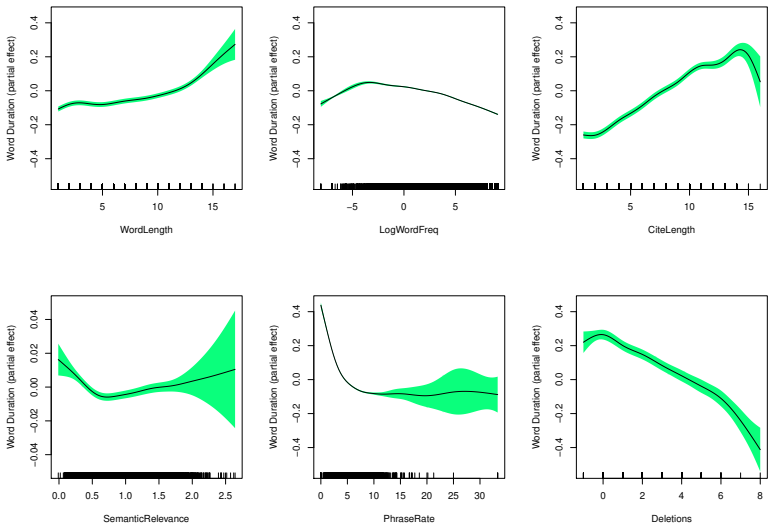

Figure 5: Partial effects of a given factor on word duration. (Note: The *x-axis* signifies the metrics, while the *y-axis* delineates word duration. Each curve visually articulates the relation between a predictor variable and the response variable. A steeper incline on these curves underscores a more robust impact between the predictor and reading speed, whereas gentler slopes imply a less pronounced effect. Moreover, when a curve fluctuates around zero, its effect vanishes. The p-values is smaller than 0.0001 in each plot.)

3.4 Discussion

In summary, we employed machine learning methods such as random forests and SVMs to train on the training dataset, allowing them to identify patterns that were then applied to predict various word durations in the test dataset. In contrast, statistical methods like LMER or GAMM were used primarily to analyze how specific factors affect word duration. These statistical methods do not require training on a testing dataset to determine necessary model parameters, and they are not typically involved in pattern recognition. Consequently, datasets used for statistical analysis do not need to be divided into training and test datasets.

Let us correlate our experiments with the key differences outlined in Table 1 to illustrate their specific applications in research on language and cognitive sciences. First, random forest and SVM were primarily employed to predict varying ranges of word durations in a new dataset. It was observed that both methods achieved a prediction accuracy of 51%. However, correlation analyses and mixed-effects regression demonstrated weaker predictive performance for word durations on new data. Second, machine learning methods often prioritize model accuracy over the interpretability of the extracted features. While the features used in these methods are interpretable, real-world machine learning algorithms focus on feature effectiveness in prediction. In contrast, features (i.e., variables) in statistical analysis should be interpretable from linguistic or cognitive perspectives. In fact, all the factors we selected are highly interpretable. Third, machine learning focuses on how features contribute to the prediction accuracy of word duration. Regression models, however, aim to determine how certain factors affect word duration and the strength of these effects, often within the framework of established statistical theories and constraints. For instance, regression models provide insights into how various factors influence model performance, thereby helping researchers understand the significant impact of these factors on the dependent variable (i.e., word durations). Fourth, statistical analysis addresses the bias and variance of estimates. This focus is crucial as researchers are interested in the distribution of data and their correlations; hence, bias and variance are key considerations.

As mentioned earlier, while machine learning and stastistics are closely related, the choice between employing a machine learning model or a statistical model often hinges on the intended use of the analysis. In some cases, a mixture of machine learning and statistical analysis could be probably used, especially in data-driven research. For instance, in computing the factor `semantic relevance`, we introduced word embeddings to represent word meanings. Word embedding is a well-known machine learning (specifically, deep learning) technique for deriving word meanings (Mikolov et al., 2013; Kusner et al., 2015; Ethayarajh, 2019), and the algorithm for semantic relevance also incorporates various machine learning methods. However, the computation of semantic relevance remains transparent and interpretable. We employed methods such as Generalized Additive Mixed Models (GAMM) or Linear Mixed Effects Regression (LMER) to investigate how semantic relevance influences word durations. The ultimate goal is to elucidate that

semantic relevance is a crucial aspect of human language comprehension; in other words, contextual information plays a vital role when humans produce language. In this sense, GAMM analysis provides interpretable results, further confirming the impact of semantic relevance on the complexity of language production.

In short, machine learning models are preferable for tasks requiring high accuracy in predictions. Conversely, when the goal is to ascertain relationships between variables or to draw inferences from data, statistical models are more suitable, offering the rigor and transparency needed for such analyses. In many instances, combining machine learning and statistical methods can enhance research outcomes. Despite this potential synergy, it is common for one approach to dominate within a specific study.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Baayen, H., Vasishth, S., Kliegl, R., and Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94:206–234.

Baker, R. E. and Bradlow, A. R. (2009). Variability in word duration as a function of probability, speech style, and prosody. *Language and Speech*, 52(4):391–413.

Bates, D. M. (2010). lme4: Mixed-effects modeling with r.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern Recognition and Machine Learning*, volume 4. Springer.

Boulesteix, A.-L. and Schmid, M. (2014). Machine learning versus statistical modeling. *Biometrical Journal*, 56(4):588–593.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Breiman, L. (2001a). Random forests. *Machine Learning*, 45:5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Bürki, A. (2018). Variation in the speech signal as a window into the cognitive architecture of language production. *Psychonomic Bulletin & Review*, 25(6):1973–2004.

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715):547–555.

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15(4):233–234.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002.

Cohen Priva, U. (2015). Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, 6(2):243–278.

Cowan, N., Wood, N. L., Nugent, L. D., and Treisman, M. (1997). There are two word-length effects in verbal short-term memory: Opposed effects of duration and complexity. *Psychological Science*, 8(4):290–295.

Ethayarajh, K. (2019). How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 17(1):69–78.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24:395–419.

Han, J., Cai, Y., and Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(1):29–40.

Harrington, P. (2012). *Machine learning in action*. Simon and Schuster.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4):18–28.

Jack, R. E., Crivelli, C., and Wheatley, T. (2018). Data-driven methods to diversify knowledge of human psychology. *Trends in Cognitive Sciences*, 22(1):1–5.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):2053951714528481.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13).

Langley, P. (1981). Data-driven discovery of physical laws. *Cognitive Science*, 5(1):31–54.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Ley, C., Martin, R. K., Pareek, A., Groll, A., Seil, R., and Tischer, T. (2022). Machine learning and conventional statistics: making sense of the differences. *Knee Surgery, Sports Traumatology, Arthroscopy*, 30(3):753–757.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18.

L'heureux, A., Grolinger, K., Elyamany, H. F., and Capretz, M. A. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access*, 5:7776–7797.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3):e0194889.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, H. J. and Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80:449–461.

Mjolsness, E. and DeCoste, D. (2001). Machine learning for science: state of the art and future prospects. *Science*, 293(5537):2051–2055.

Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., and Kutz, J. N. (2019). Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique*, 347(11):845–855.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT press.

Nelder, J. A. (1986). Statistics, science and technology. *Journal of the Royal Statistical Society: Series A (General)*, 149(2):109–121.

Nosek, B. A., Spies, J. R., and Motyl, M. (2012). Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., and Semenuks, A. (2019). What happened to cognitive science? *Nature Human Behaviour*, 3(8):782–791.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, 2:33–52.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (2005). The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Raschka, S. and Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd.

Ratner, B. (2017). *Statistical and machine-learning data mining:: Techniques for better predictive modeling and analysis of big data*. Chapman and Hall/CRC.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85.

Saidulu, D. and Sasikala, R. (2017). Machine learning and statistical approaches for big data: issues, challenges and research directions. *International Journal of Applied Engineering Research*, 12(21):11691–11699.

Solomatine, D. P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*, 10(1):3–22.

Sun, K. (2023). Optimizing predictive metrics for human reading behavior. *bioRxiv*, pages 2023–09.

Sun, K., Wang, Q., and Lu, X. (2023). An interpretable measure of semantic similarity for predicting eye movements in reading. *Psychonomic Bulletin & Review*, 30(4):1227–1242.

Terroir, J. and Lavandier, C. (2014). Perceptual impact of distance on high-speed train sound quality. *Acta Acustica United with Acustica*, 100(2):328–340.

Webb, S. et al. (2018). Deep learning for biology. *Nature*, 554(7693):555–557.

Wolf, G. (2010). The data-driven life. *The New York Times Magazine*, pages 38–L.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., and Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639.