Exploring Rate of Language Change

Samy Bakikerali
UNC Charlotte
Charlotte, NC
sbabike2@charlotte.edu

Gloria Chen
UNC Charlotte
Charlotte, NC
wchen38@charlotte.edu

Bao Dinh
UNC Charlotte
Charlotte, NC
bdinh2@charlotte.edu

Dakota Ellis
UNC Charlotte
Charlotte, NC
dellis23@charlotte.edu

Uyen Le
UNC Charlotte
Charlotte, NC
ple12@charlotte.edu

ABSTRACT

The vocabulary available for use within a language, referred to as a lexicon, does not possess a uniform distribution of implementation. This paper proposes a quantitative method that drops words of low frequency to establish a working lexicon, or a subset of the lexicon that contains only statistically significant words over a specific timeframe. The working lexicon for each year can be used in a variety of applications, including: determining rates of change of the language, training machine-learning models, and providing an optimized list of vocabulary words for English education. This paper will highlight the benefits of the latter, addressing the global demand for English skills as the language remains the dominant medium of global business (Makarenko, 2023).

KEYWORDS

historical linguistics, lexicon, language attrition, corpus analysis

ACM Reference Format:
Samy Bakikerali, Gloria Chen, Bao Dinh, Dakota Ellis, and Uyen Le. 2025. Exploring Rate of Language Change. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Language is constantly changing. New words appear. Some words change meaning. Others fade away. This phenomenon makes learning English challenging. Learners may struggle to know which words matter most. Learning every word is not realistic. Some words are rarely used. A working lexicon helps solve this problem. It includes the most common and useful words.

This study explores an important question. How can we identify the English working lexicon to support learners? Understanding this information can make language learning more effective. It helps learners focus on essential words. A vocabulary list based on actual usage is better than guessing. It speeds up fluency. It also allows researchers to track language changes. It could improve language-learning tools.

To answer this, this study uses a data-driven approach. It uses the Google Books Ngram corpus. It analyzes word frequency over time. The analysis reveals which words stay relevant. It also shows which words become less common. The goal is to create a practical word list for learners.

2 BACKGROUND

With the rise of data-driven learning (DDL), corpus based methodologies have gained popularity in language education. It enabled the use of language data to perform advanced analytics that support both language learning and teaching (Boulton, 2010). By analyzing linguistic patterns, learners can enhance their understanding of word usage and its syntactic structures. Boulton (2010) highlights that structured lexical data and corpus driven approaches significantly enhance learning efficiency, especially for beginners.

Around 4% of all printed books are in a corpus of digitized texts generated by Google, also known as Books Ngram. The frequency dataset counts how often a specific text appears in published texts. The Google Books corpus makes it possible to extract insights using quantitative methods, such as figuring out how big the English language is, finding patterns in how grammar has changed over time, or finding times when speech was restricted or censored (Michel et al., 2011). However, the corpus has intrinsic constraints for generalizing language knowledge due to its structure (Koplenig, 2015; Pechenick et al., 2015). Despite the constraints, the availability and size of the Google Books corpus enables meaningful insights into language that few corpora can.

An individual's vocabulary can be divided into two types. An active (productive) vocabulary comprises words spoken or written. These are words we understand and actively employ in a language. The other type is passive (receptive) vocabulary, which refers to words we can interpret or understand as they surface when reading or listening (Nemati, 2010). Vocabulary growth takes a different trajectory for passive and active vocabulary after a certain threshold. Where it is considered that the development of active vocabulary becomes contingent on the "need for use" of the word after a certain point (Laufer 1991) (Gu 2010).

Several studies (Laufer 1995, 1998) were conducted on developing a student's lexicon in different settings. The analyses revealed that passive vocabulary can improve by 1,600 words in 1 year; however, it took students 6 years to learn 1,900 words. Similarly, active vocabulary can improve by 850 words in 1 year; yet again, it took

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

students 6 years to learn 1,700 words. The study highlights the need for classroom instruction to optimize the setting where students can expand their lexicon.

According to research on lexical bootstrapping (Gleitman & Gleitman, 1992), learning new words is more than remembering them. It involves deeper mental processes like recognizing syntax and using words in the proper context. Exposure to a structured lexicon enables learners to acquire and apply new linguistic structures more effectively, fostering long-term language proficiency (Dulay & Burt, 1973; Xi & Geva, 2023).

The traditional method of instruction in language emphasizes rote memorization of vocabulary. However, newer studies (Stockwell, G. 2016) and (Godwin-Jones, R. 2018) support continuous contextualized engagement, which helps students learn new words faster. Stockwell and Jones encourage using mobile devices as interactive and engaging mediums for vocabulary acquisition.

3 METHODOLOGY

3.1 Dataset Description

The Google Books Ngram corpus is a collection of yearly usage frequencies for words and phrases extracted from millions of digitized books (Michel et al., 2011). Before its creation, researchers inferred trends in social sciences by reading carefully-chosen literature (Michel et al., 2011). The unprecedented availability and size of the Google Books corpus (Koplenig, 2015) enables scholars to extend the boundaries of quantitative methods in the study of culture (Michel et al., 2011).

However, researchers have since cautioned against the broad conclusions drawn from the Google Books corpus due to its inherent limitations (Koplenig, 2015; Pechenick et al., 2015). First, prolific authors can noticeably influence the corpus lexicon because each text is given a single entry (Pechenick et al., 2015). Second, the corpus consists largely of scientific texts; therefore, its lexicon is not representative of pop cultural vocabulary (Pechenick et al., 2015). Third, only texts with quality scans and metadata were included (Michel et al., 2011); while no study has characterized the omitted texts, there are potentially systemic differences between the omitted and included texts. Lastly, the omission of texts' metadata makes the broad conclusions hard to verify (Koplenig, 2015). With these limitations, claims drawn from the Google Books corpus must address its shortcomings (Pechenick et al., 2015) and restrict the claim's scope to the lexicon represented in the corpus (Koplenig, 2015). The Google Books corpus will be used in this study because of its availability and size. Although its lexicon is not representative of pop cultural vocabulary, it enables the trial of the quantitative method to estimate a working lexicon proposed in this study.