Data-Driven Approach to Identify the English Working Lexicon

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 used in a language, or its lexicon, is not used uniformly as some words are used more than others. This paper proposes a corpus-based, quantitative method to establish a working lexicon, or the subset of a language's vocabulary that is used frequently. Traditionally, linguists estimate the working lexicon in a manual process based on their subjective intuition. However, language is constantly changing and the ad-hoc approach used by linguists can be streamlined using a data-driven method. Identifying the working lexicon for each year has important applications in determining rates of change of language, in the creation of large language models, and in language education. With the dominance of English in global business and education (Makarenko, 2023), we will identify the statistically significant vocabulary English learners can focus their efforts on by learning only the words that are relevant.

KEYWORDS

lexicon, historical linguistics, language attrition, language acquisition, corpus analysis

ACM Reference Format:
Samy Bakikerali, Gloria Chen, Bao Dinh, Dakota Ellis, and Uyen Le. 2025. Data-Driven Approach to Identify the English Working Lexicon. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Language is a dynamic system of communication where new words surface, some change meaning, and others fade away. This dynamic nature of languages makes the process of learning relevant vocabulary difficult. Learners often struggle with finding the appropriate meaning and use of words [13] but learning every word's use case is impractical, as some use cases are rarely used. This study explores a data-driven solution for identifying a working lexicon for the optimization of English language learning.

This study will start by reviewing past studies and their suggestions for future research. Followed by an examination of the data used and how it is processed. It will then provide a thorough description of the methodology used to identify a working lexicon. Finally, an in depth data analysis will be conducted, followed by a discussion of the implications and conclusions. Ultimately, the goal of this study is to develop a practical and functional working lexicon to assist instruction, language learning, and vocabulary acquisition.

2 BACKGROUND

An individual's vocabulary can be divided into their active and passive vocabulary [10]. Their active, or productive, vocabulary is the words they actively employ in speaking and writing. Their passive, or receptive, vocabulary is the words they can interpret. An individual's active and passive vocabulary will continue to grow as they come into contact with new words [6, 4]; however, the growth of their vocabulary is contingent on their need to use the words they encounter [6]. The standard methods for growing individuals' vocabulary don't emphasize this need and, thus, have proven to be inefficient [8, 7]. For instance, students' vocabulary acquisition in the classroom over six years is comparable to what's possible in just one year [8]. The inefficiency stems from the rote memorization used by traditional language education, which is missing the context words are used in [3]. Exposure to the ways a word is used engages mental processes involved in recognizing patterns, leading to better vocabulary acquisition and retention [2, 14].

Data-driven learning (DDL) techniques have been proven to add the needed context to vocabulary acquisition [1]. DDL adds the context by enabling learners to analyze corpora, thereby fostering their understanding of word usage and syntactic structures [11]. However, the novelty and complexity of DDL techniques have limited their use to advanced learners in university environments [1]. A shortcoming of DDL is that it requires learners to navigate a corpora, which requires linguistic training [1]. While the use of paper-based corpus materials can alleviate this difficulty, the preparation of the materials is resource-intensive and requires linguistic expertise [1]. Given the limitations of past research and the need for more efficient vocabulary acquisition, this study will use a corpus-based, quantitative method to identify a working lexicon which will help in the preparation of language learning materials.

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

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

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 [9]. Before its creation, researchers inferred trends in social sciences by reading carefully chosen literature [9]. The unprecedented availability and size of the Google Books corpus [5] enables scholars to extend the boundaries of quantitative methods in the study of culture [9].

However, researchers have since cautioned against the broad conclusions drawn from the Google Books corpus due to its inherent limitations [5, 12]. First, prolific authors can noticeably influence the corpus lexicon because each text is given a single entry [12]. Second, the corpus consists largely of scientific texts; therefore, its lexicon is not representative of pop cultural vocabulary [12]. Third, only texts with quality scans and metadata were included [9]; 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 [5]. With these limitations, claims drawn from the Google Books corpus must address its shortcomings [12] and restrict the claim's scope to the lexicon represented in the corpus [5]. 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.

The corpus of Contemporary American-English (COCA) and the corpus of American Soap Operas (SOAP) will be used to supplement the Google Books corpus. The COCA contains over one million words used from 1990–2019 and is gathered from sources such as: TV, radio, movies, and newspapers. The SOAP contains dialogue from 20,000 transcripts of American soap operas, supplying significant instances of informal spoken language.

REFERENCES

[1] Alex Boulton. "Data-Driven Learning: Taking the Computer Out of the Equation". In: *Language Learning* 60.3 (2010), pp. 534–572. doi: https://doi.org/10.1111/j.1467-9922.2010.00566.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9922.2010.00566.x. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9922.2010.00566.x.

[2] Heidi C. Dulay and Marina K. Burt. "SHOULD WE TEACH CHILDREN SYNTAX?" In: *Language Learning* 23.2 (1973), pp. 245–258. doi: https://doi.org/10.1111/j.1467-1770.1973.tb00659.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-1770.1973.tb00659.x. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-1770.1973.tb00659.x.

[3] Lila R. Gleitman and Henry Gleitman. "A Picture Is Worth a Thousand Words, but That's the Problem: The Role of Syntax in Vocabulary Acquisition". In: *Current Directions in Psychological Science* 1.1 (1992), pp. 31–35. doi: 10.1111/1467-8721.ep10767853. eprint: https://doi.org/10.1111/1467-8721.ep10767853. url: https://doi.org/10.1111/1467-8721.ep10767853.

[4] Peter Gu. "Learning strategies for vocabulary development". In: *Reflections on English Language Teaching* 9 (Jan. 2010), pp. 105–118.

[5] Alexander Koplenig. "The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII". In: *Digital Scholarship in the Humanities* 32.1 (Sept. 2015), pp. 169–188. issn: 2055-7671. doi: 10.1093/llc/fqv037. eprint: https://academic.oup.com/dsh/article-pdf/32/1/169/17506188/fqv037.pdf. url: https://doi.org/10.1093/llc/fqv037.

[6] BATIA LAUFER. "The Development of L2 Lexis in the Expression of the Advanced Learner". In: *The Modern Language Journal* 75.4 (1991), pp. 440–448. doi: https://doi.org/10.1111/j.1540-4781.1991.tb05380.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-4781.1991.tb05380.x. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-4781.1991.tb05380.x.

[7] BATIA LAUFER. "The Development of Passive and Active Vocabulary in a Second Language: Same or Different?" In: *Applied Linguistics* 19.2 (June 1998), pp. 255–271. issn: 0142-6001. doi: 10.1093/applin/19.2.255. eprint: https://academic.oup.com/applij/article-pdf/19/2/255/9741631/255.pdf. url: https://doi.org/10.1093/applin/19.2.255.

[8] BATIA LAUFER and PAUL NATION. "Vocabulary Size and Use: Lexical Richness in L2 Written Production". In: *Applied Linguistics* 16.3 (Sept. 1995), pp. 307–322. issn: 0142-6001. doi: 10.1093/applin/16.3.307. eprint: https://academic.oup.com/applij/article-pdf/16/3/307/9740387/307.pdf. url: https://doi.org/10.1093/applin/16.3.307.

[9] Jean-Baptiste Michel et al. "Quantitative Analysis of Culture Using Millions of Digitized Books". In: *Science* 331.6014 (2011), pp. 176–182. doi: 10.1126/science.1199644. eprint: https://www.science.org/doi/pdf/10.1126/science.1199644. url: https://www.science.org/doi/abs/10.1126/science.1199644.

[10] Azadeh Nemati. "Active and passive vocabulary knowledge: The effect of years of instruction". In: *Journal of Applied Sciences* 12 (Jan. 2010), pp. 3746–3751.

[11] Íde O'Sullivan. "Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy". In: *ReCALL* 19.3 (2007), pp. 269–286. doi: 10.1017/S095834400700033X.

[12] Eitan Adam Pechenick, Christopher M. Danforth, and Peter Sheridan Dodds. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution". In: *PLOS ONE* 10.10 (Oct. 2015), pp. 1–24. doi: 10.1371/journal.pone.0137041. url: https://doi.org/10.1371/journal.pone.0137041.

[13] Sardor Surmanov and Maftuna Azimova. "ANALYSIS OF DIFFICULTIES IN VOCABULARY ACQUISITION". In: *The Journal of Legal Studies* 6 (Sept. 2020), pp. 144–153.

[14] Yueming Xi and Esther Geva. "A 4-year longitudinal study examining lexical and syntactic bootstrapping in English Language Learners (ELLs) and their monolingual peers". en. In: *Developmental Psychology* 59.1 (Nov. 2022), pp. 161–172.