tasks

*Documents the tasks we have completed on a weekly basis.*

# Week 7

* Paper submission for 1st half term
  + Abstract
  + Introduction
  + Background
  + Data description
* Individual contributions survey

# Week 6

* ~~Paper draft~~
  + ~~Abstract (Dakota)~~
  + ~~Introduction (Uyen)~~
  + ~~Background (Gloria & Samy)~~
  + ~~Data description (Bao)~~

# Week 5

* ~~Meeting with Dr. Bybee~~

# Week 4

* ~~Annotated bibliography (at least 3 per person)~~
  + ~~Samy~~
  + ~~Gloria~~
    - [~~Rate of language evolution is affected by population size~~](https://doi.org/10.1073/pnas.1419704112)
    - [~~Evolutionary dynamics of language systems~~](https://doi.org/10.1073/pnas.1700388114)
    - [~~Language Anxiety : Conflict and Change in the History of English~~](https://ebookcentral.proquest.com/lib/uncc-ebooks/reader.action?docID=415479&ppg=1)
  + ~~Bao~~
    - [~~Quantitative analysis of culture using millions of digitized books~~](https://pmc.ncbi.nlm.nih.gov/articles/PMC3279742/)
    - [~~Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution~~](https://pmc.ncbi.nlm.nih.gov/articles/PMC4596490/)
    - [~~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~~](https://doi.org/10.1093/llc/fqv037)
  + ~~Dakota~~
  + ~~Uyen~~
    - [~~How and Why Language Changes~~](https://humanitiescenter.byu.edu/how-and-why-language-changes/)
    - [~~The Language of Economic Growth: A New Measure of Linguistic Heterogeneity~~](https://www.cambridge.org/core/journals/british-journal-of-political-science/article/language-of-economic-growth-a-new-measure-of-linguistic-heterogeneity/A558FB3D3EBE6E2936CB4C7B72D69293)
    - [~~LANGUAGE DIVERSITY AND ECONOMIC DEVELOPMENT~~](https://www.researchgate.net/publication/253480616_LANGUAGE_DIVERSITY_AND_ECONOMIC_DEVELOPMENT)
* ~~Google Books Ngram data (Bao)~~
  + ~~Chinese (simplified)~~
  + ~~English~~
  + ~~French (remaining:~~ **~~1-00005-of-00006.gz~~**~~)~~
  + ~~German (remaining:~~ **~~1-00006-of-00008.gz~~**~~,~~ **~~1-00007-of-00008.gz~~**~~)~~
  + ~~Hebrew~~
  + ~~Italian (remaining:~~ **~~1-00001-of-00002.gz~~**~~)~~
  + ~~Russian (remaining:~~ **~~1-00001-of-00002.gz~~**~~)~~
  + ~~Spanish (remaining:~~ **~~1-00001-of-00003.gz~~**~~,~~ **~~1-00002-of-00003.gz~~**~~)~~
* ~~Email professors~~
  + ~~Dr. Damien Williams (Dakota)~~
  + ~~Dr. Joan Bybee (Samy)~~
  + ~~Dr. Alexia Galati (Dakota)~~
  + ~~Dr. Min Jiang (Dakota)~~

# Week 3

* ~~Project proposal: first draft title and abstract~~

# Week 2

* ~~Presentations: project topics~~

research-topics

*Gives an overview of the research questions we need to answer as part of our research. Write brief notes about how we’ve answered the questions.*

# Pièce de Résistance

[Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC](https://www.youtube.com/watch?v=jbkSRLYSojo)

# Managerial Questions

1. ~~What is the rate of language change over the years? (e.g. past 100 years)?~~
2. How can defining a core working lexicon and its educational application enhance language learning effectiveness? ~updated by 2/18/2025

**Research Questions**

1. Defining the Core Working Lexicon
2. Corpus Analysis: Utilize large linguistic datasets, such as the British National Corpus (BNC), to identify high-frequency words and phrases. http://www.natcorp.ox.ac.uk/corpus/
3. Data-Driven Learning (DDL): Implement DDL approaches where learners engage directly with real language data.

https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9922.2010.00566.x?casa\_token=3\_-fql82-A0AAAAA%3ARiC7wn30wF9NW8WBS1XAKpKhGsnnKlzolyb7m-muGeJ4O6RmFHCBIwjHayBjJx2NiWDlUNPiNQBirw

1. Educational Applications
2. Personalized Learning Pathways
3. Integration with Technology
4. Measuring Effectiveness
5. Pre- and Post-Testing
6. Longitudinal Studies
7. Data Collection and Analysis
8. Learner Interaction Data
9. Feedback Mechanisms
10. What is the solution? What is our end result?
11. How can we quantify the change in a language? (e.g. change in English lexicon)
    1. What is the rate of change?
       1. Has the rate become more dynamic?
       2. Has the rate been constant?
    2. How do we measure the rate of change?
       1. How many new words are added each year? How many words fall out of common use?
       2. Use outlier removal techniques to determine threshold for when a word falls out of common use.
       3. How many new foreign words are added each year? (e.g. OED country of origin)
       4. Statistical divergence?
          1. Kullback-Leibler divergence?
          2. Jensen-Shannon divergence?
12. What language are we investigating? (e.g. English)
    1. What is a spoken, cultural language?
    2. How is a language created?
    3. Are we comparing languages?
       1. What is the overlap between languages?
13. Can we relate the change to historical events (e.g. industrial revolution, TV, Internet)
    1. Create a baseline of changes in lexicon, then establish a line graph (example)
    2. Create line graph for different metrics that may influence language and compare to baseline to find correlations
14. What is the relationship between the change in languages and GDP (e.g. global GDP)?
    1. How has GDP grown overtime?
    2. Who is the dominant GDP contributor?
15. Why is a data-driven approach necessary in linguistic analysis?
    1. Its big data
16. How has the rate of language change compared to GDP growth across major continents since 1945?
    1. We need global economic databases.
    2. Language change data.
    3. Our research explores the relationship between **language evolution** and **economic growth (GDP)** across major continents since 1945. This is valuable because:
       1. It reveals the impact of globalization on economic growth; language evolves due to trade, migration, and technological advancements.
       2. It helps predict future language & economic trends (e.g., China’s rise). We can identify these trends that can assist in education, AI language modeling, and international trade.
       3. It bridges linguistics & economics.
       4. It provides practical insights into AI & language technology.
    4. Major continents:

1. Asia: China, Mandarin

2. Africa: Nigeria, Nigeria

3. North America: USA, English

4. South America: Brazil, Spanish

5. Europe: Germany, German

6. Australia: Australia, English

* 1. GDP DATA: <https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2023>

Research question:

* How does the impact of social media affect language change in education?
* What are the application of working lexicon in language analytics?
* What are the implications of a frequency-based lexicon for ESL education, and how can it aid language learners in acquiring commonly used vocabulary?

1. Group: Wanted to look at change in language because of the video

2. Dr. Bybee: “Is the working lexicon for a language actually 10k words?”

3. Dr. Mirsad: “Proving Dr. Bybee’s question is a contribution to linguistics, however, the project is missing a data science component. You can add a data science component by looking at the factors contributing the most to language change.”

4. Group:

What we are going to do: determine volatility of English, why we are gonna do it: make english more accessible, connect two things together to give evidence for our research,

contacts

*Records our interactions with professors and subject matter experts.*

# Dr. Joan Bybee

* ~~UPDATE: Waiting for response, emailed on 2025/02/06~~
* ~~ACTION: Set up time to talk~~
* ~~ACTION: Talk with Dr. Bybee (2025/02/14)~~
* ACTION: Update Dr. Bybee about project progression

# Dr. Batia Laufer

* ~~UPDATE: Waiting for response, emailed on 2025/02/22~~
* ACTION: Set up time to talk

# ~~Dr. Damien Williams~~

* ~~UPDATE: Waiting for response, emailed on 2025/02/06~~
* ~~UPDATE: No response~~

# ~~Dr. Alexia Galati~~

* ~~UPDATE: Waiting for response, emailed on 2025/02/10~~
* ~~UPDATE: Not available for semester, currently on sabbatical~~

# ~~Dr. Min Jiang~~

* ~~UPDATE: Waiting for response, emailed on 2025/02/10~~
* ~~UPDATE: Not their field of study~~

Dr. Joan Bybee

Meeting Notes

Friday, 14 February 2025

Get clarity on next steps for our research with [Dr. Joan Bybee](https://www.unm.edu/~jbybee/)

1. Estimating Working Vocabulary:
   * Traditional methods involved randomly sampling dictionary words and identifying known ones, but this approach has many limitations.
   * Modern research might be able to improve this method with data-driven approaches.
2. Lexicon Growth and Loss Over Time:
   * Most research focuses on new words entering the lexicon rather than words falling out of use.
   * Studying language change should consider both aspects.
3. Lemma-Based Word Counting:
   * Inflectionally related words (e.g., "write," "writing," "wrote") should be grouped together as lemmas.
   * High-frequency words sometimes behave differently (e.g., "child" vs. "children" vs. "kids").
4. Globalization and Language Change:
   * The Industrial Revolution and the Internet age have accelerated vocabulary shifts.
   * Communication is now less bound by geography, possibly affecting the rate of language change.
5. Factors Influencing Language Change:
   * Languages at cultural and trade crossroads tend to change faster (e.g., English vs. Icelandic).
   * Borrowing often happens for new objects or technologies that don’t already have native words.
6. Economic Development and Borrowing:
   * Technologically advanced countries create new terms that other languages may borrow.
   * Economic strength (e.g., GDP growth) may correlate with a higher influx of borrowed words.
7. Resistance to Borrowing:
   * Some languages, like German and French, prefer creating native words rather than borrowing foreign ones.
8. Written vs. Spoken Language:
   * Written language has a larger vocabulary, but spoken language includes slang and informal expressions.
   * Some corpora (e.g., COCA) provide transcribed spoken language, which could be valuable for analysis.

Step-by-Step Vocabulary Teaching:

* Teaching English as a second language follows a structured progression of vocabulary.
* Researchers in word recognition and speech studies focus on practical vocabulary.

Limitations of Google Books Ngram Corpus:

* It consists mostly of books from universities, including scientific and educational texts.
* Important to assess how books were chosen and whether there are gaps (e.g., missing informal or spoken language).
* Expanding to other corpora like COCA (spoken language), SOAP (conversational speech), and movie subtitle databases can improve the study.

Differences Between Written and Spoken Corpora:

* COCA contains spoken data but includes many interviews, which differ from casual conversations.
* SOAP and movie subtitle corpora may better represent everyday spoken language.

Estimating a Person’s Working Lexicon:

* A corpus reflects a population’s vocabulary, not an individual’s.
* However, individuals are embedded in a linguistic community, so studying common core vocabulary can still be insightful.
* The central limit theorem can help estimate a population’s typical lexicon size.

Applying Lexicon Statistics:

* Lexicon statistics is a valuable field for this research.
* Identifying a stable "core vocabulary" over time could provide insights into necessary words for understanding texts.

Dr. Batia Laufer

[Dr. Batia Laufer](https://english.haifa.ac.il/2021/10/18/prof-batia-laufer/?lang=en)

potential-sources

*Documents all of the potential sources (for literature and data sources) — categorized by their relevant subject matter. Annotate the source in the* [*[sources] Google Sheets*](https://docs.google.com/spreadsheets/d/1DQNQx-eSpiNr_gGgxtp6E92eCA5Nli8-2aBS0JwlQ9A/edit?gid=0#gid=0)*. DO NOT WRITE YOUR LITERATURE REVIEW HERE.*

archive

*Documents archived sources (for literature and data sources) — categorized by their relevant subject matter. Potential sources are moved to the archive when they are no longer relevant to our research. Sources are archived to reflect our current research and are documented in case we need them in the future.*

# Literature

## Understanding Language

[The language instinct: How the mind creates language](https://danielwharris.com/teaching/268/readings/Pinker.pdf)

[Language Change and Historical Reconstruction](https://www.ling.upenn.edu/courses/Fall_2003/ling001/language_change.html)

[Bilingualism: The good, the bad, and the indifferent](https://www.cambridge.org/core/journals/bilingualism-language-and-cognition/article/bilingualism-the-good-the-bad-and-the-indifferent/36BAEB01D08C92D992254A6B89C22BB0)

[The Globalization of Language](https://denizo.opia.dk/la.trezorejo/alilingve/hjarvard.The.Globalization.of.Language.pdf)

[Internal and external forces in language change](https://doi.org/10.1017/S0954394500123014)

[How Many Is Enough?—Statistical Principles for Lexicostatistics - PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC5149542/)

## Dr. Joan Bybee

[Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change](https://www.unm.edu/~jbybee/downloads/Bybee2002WordFreq.pdf)

[How plausible is the hypothesis that population size and dispersal are related to phoneme inventory size? Introducing and commenting on a debate](https://doi.org/10.1515/lity.2011.009)

[Language universals and usage-based theory](https://www.unm.edu/~jbybee/downloads/Bybee2009LangUniv.pdf)

[Language change and universals](https://www.unm.edu/~jbybee/downloads/Bybee2006LangChange.pdf)

## Google Books Corpus

[Quantitative analysis of culture using millions of digitized books](https://pmc.ncbi.nlm.nih.gov/articles/PMC3279742/)

[Syntactic Annotations for the Google Books NGram Corpus](https://aclanthology.org/P12-3029/)

[Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution](https://pmc.ncbi.nlm.nih.gov/articles/PMC4596490/)

[Using the Google N-Gram corpus to measure cultural complexity](https://academic.oup.com/dsh/article/28/4/668/1075348)

[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](https://doi.org/10.1093/llc/fqv037)

## Diachronic Corpus Linguistics

[Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions](https://doi.org/10.1093/llc/fqv030)

[Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis](https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/4223/file/Koplenig_Using_the_parameters_of_the_Zipf_Mandelbrot_2015.pdf)

[Reflecting on the quantitative turn in linguistics](https://www.degruyter.com/document/doi/10.1515/ling-2019-0046/html)

[25 years of English Language and Linguistics: a celebration and analysis](https://www.cambridge.org/core/journals/english-language-and-linguistics/article/25-years-of-english-language-and-linguistics-a-celebration-and-analysis/2B9E057E9A3D054C3D7C3D94A910BD79)

## Comparison Techniques

[Quantitative approaches to diachronic corpus linguistics](https://doi.org/10.1017/CBO9781139600231.003)

[Evolution of Semantic Similarity—A Survey](https://doi.org/10.1145/3440755)

[Quantifying Semantic Alignment Across Languages](https://research-information.bris.ac.uk/ws/portalfiles/portal/208272739/Thompson_Roberts_Lupyan_2018.pdf)

[Similarity in languages and programs](https://www.sciencedirect.com/science/article/pii/S0304397513004520)

[Similarities and differences](https://d1wqtxts1xzle7.cloudfront.net/108135295/sci2000-libre.pdf?1701465254=&response-content-disposition=inline%3B+filename%3DSimilarities_and_Differences.pdf&Expires=1738017902&Signature=TgjK~Kr6FKhtGJHPnTE60fMv763xh~G-hkDah-Bv6aTLQscej911pw23IwLuIa3tMpI5TVoWbnWV2Oo6lrPczLr4Ikwn-gZ~hAYvPdOUgGTXKC0GPQVor7o72LjlGMzzs41othO7RKULbO8PbhtC9mG8lEJRMiWwXY91LXnTlG8JPrSjt9WXwmehjxbKlHeNXA16QDo8Fauw3ZoRgRZqjSzi-W9H~Jo2g5NOF4STDhGiUup1EmHTJyviHiVq0TALJqzk0L1aeg2nO2ho9YYuGx0gu~PjsBtUXiTBhDqUcUdz3N6h1LkYQFfIxcQqzQOR3-HEBxTLx5KMOvLRL8a6oQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)

[Measuring the Semantic Similarity of Texts](https://aclanthology.org/W05-1203.pdf)

[A Model for Matching Semantic Maps between Languages (French/English, English/French)](https://aclanthology.org/J03-2001.pdf)

[Divergence measures based on the Shannon entropy](https://doi.org/10.1109/18.61115)

## Social Networks

[Social network analysis: An overview](https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1256)

[Tell me who your friends are and I might be able to tell you what language(s) you speak: Social network analysis, multilingualism, and identity](https://journals.sagepub.com/doi/abs/10.1177/13670069070110030201)

literature-review

*Aggregates our independent literature reviews. The final literature review for submission will be on a separate document.*

*WRITE YOUR RESPECTIVE LITERATURE REVIEWS WITHIN YOUR ASSOCIATED SUBTAB (Click the hamburger menu on the left for the [Document Tabs] > Click the dropdown arrow for the [literature-review] tab > Click on your respective subtab)*

Outline:

1. Synthesize across sources/authors/methods (i.e., thematic organization around variables, methods, etc.)
2. Summarize the state of the literature in terms of consensus (or: conventional wisdom), disagreement, and omissions or gaps
3. Share empirical findings
4. Evaluate the overall state of knowledge
5. Discuss any problems or flaws in existing research that your research could improve upon (e.g., critique current literature)
6. Highlight the importance of conclusions your research might draw (e.g., normative implications)

# Data

Google Books Ngram is a corpus of digitized texts that encompasses around 4% of all books ever printed. It is a frequency dataset that provides a count of how many times a single text appeared in published texts over time. The creation of Google’s Ngram allowed for the extraction of many data driven insights such as estimating the size of the English lexicon, revealing trends in the evolution of grammar, or detecting periods of censorship and suppression (Michel et al., 2011). In contrast, the Ngram has inherent limitations when generalizing insights about language. For example, the large proportion of scientific texts can lead to an overrepresentation of specialized terms that are less common in everyday usage (Pechenick et al., 2015).

# Active Vocabulary Definition

An individual’s vocabulary can be broken down into 2 types. An active (productive) vocabulary, which is compromised of the words that an individual uses when speaking or writing. They are words we understand and actively use in a language. The other type is passive (receptive) vocabulary, which are the words that we can interpret or understand as they surface when reading or listening. (Nemati, 2010)

# Study on Learning Vocabulary

Several studies (Laufer 1995, 1998) were conducted in different settings, on the development of student’s lexicon. The studies 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 students 6 years to learn 1,700 words. The study highlights the need for classroom instruction to optimize the setting inwhich students can expand their lexicon. Another study adds that for the lexical development of english as second language (ESL) learners, it is crucial practice writing with already acquired vocabulary (Goya, Cai, Ding, Fecher, 2011).

The development of vocabulary takes a different direction for passive and active vocabulary after a certain threshold. Where it is believed that the development of active vocabulary becomes dependent on the “need for use” of the word after a certain point (Laufer 1991) (Gu 2010).

Data-driven learning (DDL) is becoming a standard approach in language education. Many programs now use structured corpora. This method helps analyze language data. It also gives teachers and students new ways to improve learning (Boulton, 2010). Learners study language patterns. This helps them understand how words are used and how sentences are formed. Boulton (2010) explains that structured word data and corpus-based methods improve learning. Beginners benefit the most. They receive gradual and structured practice with language.

Research on lexical bootstrapping shows that learning new words is more than memorization. It involves recognizing grammar and using words correctly in context (Gleitman & Gleitman, 1992). A structured vocabulary helps learners see patterns in language. This supports long-term language development (Dulay & Burt, 1973; Xi & Geva, 2023).

DDL is expanding, and corpus-based methods are now widely used. Corpora are large collections of text. DDL integrates these resources into learning. This allows students to work with real-world language in structured exercises (Boulton, 2010). By studying common word patterns, learners improve their understanding of vocabulary and grammar. Boulton (2010) states that structured word data and corpus-based methods make learning more effective. This approach is especially helpful for beginners. It provides steady, step-by-step exposure to language.

Traditional vocabulary instruction focuses on memorization. However, recent research suggests that learning words in context is more effective. Corpus linguistics, frequency-based word lists, and digital tools encourage active language use. These methods help students move from recognizing words to using them in real communication.

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 books ever printed are encompassed in a corpus of digitized texts created by Google, also known as Books Ngram. It is a frequency dataset that provides a count of how many times a single text appears in published texts over time. The creation of Google’s Ngram allowed for the extraction of many data driven insights, such as estimating the size of the English lexicon, revealing trends in the evolution of grammar, or detecting periods of censorship and suppression (Michel et al., 2011). In contrast, the Ngram has inherent limitations when generalizing insights about language. For example, the large proportion of scientific texts can lead to an overrepresentation of specialized terms that are less common in everyday usage (Pechenick et al., 2015).

An individual’s vocabulary can be broken down into 2 types. An active (productive) vocabulary is composed of the words that an individual uses when speaking or writing. They are words we understand and actively use in a language. The other type is passive (receptive) vocabulary, which are the words that we can interpret or understand as they surface when reading or listening (Nemati, 2010).

The development of vocabulary takes a different direction for passive and active vocabulary after a certain threshold. Where it is believed that the development of active vocabulary becomes dependent on the “need for use” of the word after a certain point (Laufer 1991) (Gu 2010).

Several studies (Laufer 1995, 1998) were conducted in different settings on the development of student’s lexicon. The studies 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 students 6 years to learn 1,700 words. The study highlights the need for classroom instruction to optimize the setting in which students can expand their lexicon.

According to research on lexical bootstrapping (Gleitman & Gleitman, 1992), learning new words is more than just remembering them. It involves deeper mental processes like recognizing syntax and using words in the right 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, more recent papers (Stockwell, G. 2016) (Godwin-Jones, R. 2018) tend to favor continuous contextualized engagement, suggesting that it accelerates vocabulary acquisition. Both Stockwell and Jones encourage the use of mobile devices as interactive and engaging mediums for vocabulary acquisition.

Samy Bakikerali

THE CLIPPED LANGUAGE REVOLUTION: EXPLORING CLIPPING IN CONTEMPORARY SOCIAL MEDIA ENGLISH.

Link to the source

https://ibn.idsi.md/sites/default/files/imag\_file/32-44\_14.pdf

The lead author(s)

Dragusin, Elena Denisa

Research question the researchers aimed to address

What are the types of clipped language? How significant are they in the new era of Social Media English?

Relevant quotes

* Users frequently engage in linguistic creativity by coining new words, adapting existing ones, and employing various forms of wordplay, proving the plasticity and the dynamic nature of language reflected on social media in response to digital communication needs.
* The limited character counts and rapid pace of interaction on platforms like Facebook, Instagram, WhatsApp, Twitter and Snapchat have led to the widespread adoption of clipped language forms such as acronyms, contractions, and abbreviations. Words and phrases are often shortened for efficiency and convenience, giving rise to new lexical innovations and linguistic shortcuts
* Furthermore, the rise of social media has facilitated the global spread of English as a **lingua franca**, with SME serving as a common ground for communication among speakers from diverse linguistic backgrounds.
* We may affirm undoubtfully that clipped language is likely to continue evolving in response to changes in technology, communication platforms, and linguistic trends

Gradual Modifications and Abrupt Replacements: Two Stochastic Lexical Ingredients of Language Evolution.

Link to the source

https://direct.mit.edu/coli/article/49/2/301/114514/Gradual-Modifications-and-Abrupt-Replacements-Two

The lead author(s)

Michele Pasquini, Maurizio Serva, Davide Vergni

Research question the researchers aimed to address

Creation of a new automated cognate detection to discriminate gradual lexical modification and abrupt lexical replacement; to prove that these 2 are random processes that separately drive the evolution of languages

Relevant quotes

* Applying statistics to determine the degree of similarity between two languages is the founding idea of lexicostatistics
* Glottochronology, the application of lexicostatistical methods with the goal of establishing when a language separated into derived languages

Diachronic changes in lexical density of research article abstracts: A corpus-based study.

Link to the source

https://www.sciencedirect.com/science/article/pii/S0024384124001682

The lead author(s)

Zhu Haoran, Wang Teng, Pang Nana

Research question the researchers aimed to address

Does hard science research exhibit higher lexical density than soft science papers? Study to investigate both the temporal dynamics of and disciplinary variations in lexical density using extensive multidisciplinary text data.

Relevant quotes

* Lexical density refers to the ratio of content words and measures the information density of academic texts
* Using a large-scale corpus of research article abstracts from 16 academic disciplines, the present study reveals that the rise in lexical density could be a common trend across all fields.
* several avenues that future research could take. First, the present study employed RA abstracts as data. Although the abstract is an important part of a research article, it may be an overgeneralization to extend the conclusions of this study to RA full texts. Thus, it would be interesting to validate the findings of the present study with RA full texts, should the challenge of data availability be successfully addressed.

[How to Trace the Growth in Learners Active Vocabulary? A Corpus-based Study in: Teaching and Learning by Doing Corpus Analysis](https://brill.com/display/book/edcoll/9789004334236/B9789004334236-s019.xml)

Gloria Chen

Language acquisition is a dynamic process influenced by vocabulary structure, cognitive mechanisms, and educational methodologies. One crucial factor in learning a language is mastering a core working lexicon, a set of high-frequency words that serve as building blocks for fluency. Researchers have long argued about the connection between learning new words and learning new grammar rules. More recent longitudinal studies (Xi & Geva, 2023) have shown that vocabulary and grammar rules affect each other in English Language Learners (ELLs) and students who speak only one language.

Some crucial ways that language develops are lexical bootstrapping (using vocabulary to learn grammar) and syntactic bootstrapping (using syntax to figure out meaning) (Gleitman & Gleitman, 1992; Xi & Geva, 2023) in many studies. Longitudinal data suggest that early syntax proficiency predicts later vocabulary growth and vice versa. However, ELLs tend to rely more on vocabulary to develop syntax, whereas monolingual peers exhibit a more balanced syntax-to-vocabulary relationship (Xi & Geva, 2023).

Studies on passive vs. active vocabulary acquisition (Laufer, 1995; 1998) indicate that learners can expand passive vocabulary by 1,600 words per year, but active vocabulary growth remains slower (~850 words per year). This highlights the need for structured lexical instruction to facilitate efficient language learning.

Corpus-based research, such as Google Books Ngram Viewer, has provided insights into historical word frequency trends, helping researchers determine which words form the core lexicon of a language. However, there are still worries that academic datasets contain too much specialized language, which can cause problems in real-life language use (Pechenick et al., 2015).

# 

### **Summary of**

**A 4-Year Longitudinal Study Examining Lexical and Syntactic Bootstrapping in English Language Learners (ELLs) and Their Monolingual Peers  
Authors:** Yueming Xi & Esther Geva (2023)  
**Published in:** *Developmental Psychology***DOI:** 10.1037/dev0001417

**Data-Driven Learning: Taking the Computer Out of the Equation**

**Author:** Alex Boulton  
**Published in:** *Teaching and Language Corpora Conference (TaLC)***DOI:** 10.1111/j.1467-9922.2010.00566.x

# Managerial Questions

1. How can defining a core working lexicon and its educational application enhance language learning effectiveness? ~updated by 2/18/2025

**Research Questions**

1. Defining the Core Working Lexicon
2. Corpus Analysis: Utilize large linguistic datasets, such as the British National Corpus (BNC), to identify high-frequency words and phrases. http://www.natcorp.ox.ac.uk/corpus/
3. Data-Driven Learning (DDL): Implement DDL approaches where learners engage directly with real language data.

https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9922.2010.00566.x?casa\_token=3\_-fql82-A0AAAAA%3ARiC7wn30wF9NW8WBS1XAKpKhGsnnKlzolyb7m-muGeJ4O6RmFHCBIwjHayBjJx2NiWDlUNPiNQBirw

1. Educational Applications
2. Personalized Learning Pathways
3. Integration with Technology
4. Measuring Effectiveness
5. Pre- and Post-Testing
6. Longitudinal Studies
7. Data Collection and Analysis
8. Learner Interaction Data
9. Feedback Mechanisms

[Language Change and Historical Reconstruction](https://www.ling.upenn.edu/courses/Fall_2003/ling001/language_change.html)

Summary:

The article begins by exploring the history of language change, making it a valuable resource for establishing the historical context of linguistic evolution in our research. It provides insights into the mechanisms and factors driving language change over time, which can help us frame a compelling narrative in our research background.

Relevance to our Research:

This article provides foundational insights into the mechanisms and causes of language change, which are crucial for understanding how languages evolve over time. For our research, focusing on the interplay between language change and socio-economic factors, the following points are particularly pertinent:

Language Learning and Transmission: Investigating how economic development influences language acquisition and the introduction of new linguistic elements.

Language Contact: Exploring how trade and economic interactions lead to the borrowing and integration of new terms, especially in rapidly developing economies.

Social Differentiation: Analyzing how different socio-economic groups adopt distinct linguistic features, potentially leading to dialectal variations.

By understanding these mechanisms, we can get a better picture of how changes in socioeconomic conditions affect the development of languages.

Print book at our library (not sure it is useful or not):

Title

Corpus analysis : language structure and language use

Author/Creator

[North American Symposium on Corpus Linguistics and Language Teaching (3rd : 2001 : Boston, Mass.)](https://charlotte.primo.exlibrisgroup.com/discovery/search?query=creator%2Cexact%2CNorth%20American%20Symposium%20on%20Corpus%20Linguistics%20and%20Language%20Teaching%2CAND&tab=Everything&search_scope=MyInst_and_CI&sortby=rank&vid=01UNCC_INST%3A01UNCC_INST&facet=creator%2Cexact%2CNorth%20American%20Symposium%20on%20Corpus%20Linguistics%20and%20Language%20Teaching&mfacet=tlevel%2Cinclude%2Cpeer_reviewed%2C1&mfacet=rtype%2Cinclude%2Cprintbooks%2C1&mfacet=rtype%2Cinclude%2Celectronicbooks%2C1&mfacet=rtype%2Cinclude%2Carticles%2C1&mfacet=rtype%2Cinclude%2Cbook_chapters%2C1&mfacet=tlevel%2Cinclude%2Cavailable_p%2C1&lang=en&mode=advanced&offset=0)

[Leistyna, Pepi.](https://charlotte.primo.exlibrisgroup.com/discovery/search?query=creator%2Cexact%2CLeistyna%2C%20Pepi.%2CAND&tab=Everything&search_scope=MyInst_and_CI&sortby=rank&vid=01UNCC_INST%3A01UNCC_INST&facet=creator%2Cexact%2CLeistyna%2C%20Pepi.&mfacet=tlevel%2Cinclude%2Cpeer_reviewed%2C1&mfacet=rtype%2Cinclude%2Cprintbooks%2C1&mfacet=rtype%2Cinclude%2Celectronicbooks%2C1&mfacet=rtype%2Cinclude%2Carticles%2C1&mfacet=rtype%2Cinclude%2Cbook_chapters%2C1&mfacet=tlevel%2Cinclude%2Cavailable_p%2C1&lang=en&mode=advanced&offset=0)

[Meyer, Charles F.](https://charlotte.primo.exlibrisgroup.com/discovery/search?query=creator%2Cexact%2CMeyer%2C%20Charles%20F.%2CAND&tab=Everything&search_scope=MyInst_and_CI&sortby=rank&vid=01UNCC_INST%3A01UNCC_INST&facet=creator%2Cexact%2CMeyer%2C%20Charles%20F.&mfacet=tlevel%2Cinclude%2Cpeer_reviewed%2C1&mfacet=rtype%2Cinclude%2Cprintbooks%2C1&mfacet=rtype%2Cinclude%2Celectronicbooks%2C1&mfacet=rtype%2Cinclude%2Carticles%2C1&mfacet=rtype%2Cinclude%2Cbook_chapters%2C1&mfacet=tlevel%2Cinclude%2Cavailable_p%2C1&lang=en&mode=advanced&offset=0)

Contents

"It's really fascinating work" : differences in evaluative adjectives across academic registers / John M. Swales and Amy Burke -- "But here's a flawed argument" : socialisation into and through metadiscourse / Anna Mauranen -- Register-specificity of signaling nouns in discourse / John Flowerdew -- Variation among university spoken and written registers / Douglas Biber -- Linguistic dimensions of direct mail letters / Ulla Connor and Thomas Upton -- Gender-based variation in nineteenth-century English letter-writing / Christer Geisler -- The grammar of stance in early eighteenth-century English epistolary language / Susan Fitzmaurice -- Great vs. lovely : stance differences in American and British English / Kristen Precht -- "What's in a name?" : vocatives in casual conversations and radio phone-in calls / Michael J. McCarthy and Anne O'Keeffe -- Turn initiators in spoken English / Hongyin Tao -- Situational variation in intonational strategies / Malcah Yaeger-Dror, Lauren Hall-Lew, and Sharon Deckert -- On the radical difference between the subject personal pronouns in written and spoken European French / Bonnie Fonseca-Greber and Linda R. Waugh -- The World Wide Web as linguistic corpus / Charles Meyer [and others] -- Corpus linguistics and second language.

acquisition / Robert Bley-Vroman -- Comparing alternate complements of object control verbs / Juhani Rudanko.

Data-driven learning (DDL) is becoming more popular. More language programs now use corpora based on structured methods. This approach helps analyze language data. It also allows teachers and students to apply advanced techniques in learning and instruction (Boulton, 2010). Learners can study linguistic patterns. The information helps them understand how words are used and how sentences are structured. Boulton (2010) explains that structured word data and corpus-based methods improve learning. Beginners benefit the most. They gain step-by-step exposure to language input.

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 structure of Google's Ngram makes it possible to get a lot of different data-driven insights, like 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). In contrast, the Ngram has intrinsic constraints for generalizing language knowledge. For example, the substantial proportion of scientific publications can lead to overrepresenting specialist terminology that is less common in daily usage (Pechenick et al., 2015).

An individual’s vocabulary can be divided into two sorts. 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 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.

data

*Documents all of the potential data sources— categorized by their relevant subject matter. Write brief notes about each datasource and how it can be applied to our research.*

[*https://datasetsearch.research.google.com/*](https://datasetsearch.research.google.com/)

[Open Language Archives Community](http://www.language-archives.org/index.html)

[Google Books NGrams](https://storage.googleapis.com/books/ngrams/books/datasetsv3.html) ([Instructions](https://stressosaurus.github.io/raw-data-google-ngram/))

# Linguistic Corpora

[English Corpora](http://english-corpora.org/)

[Non-English, Parallel & Multilingual Corpora](https://martinweisser.org/corpora_site/corpora2.html)

[Leipzig Corpora Collection](https://wortschatz.uni-leipzig.de/en/download)

# Software

[SIL Language Technology - Cog](https://software.sil.org/cog/)

[SIL Language Technology - Software Products](https://software.sil.org/software-products/)

**paper**

**Data driven methods for optimizing active lexicons for education.**

# 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).

# 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.

# 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 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.

# Methodology

## Data 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.

# References

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, *60*(3), 534–572. <https://doi.org/10.1111/j.1467-9922.2010.00566.x>

Dulay, H., & Burt, M. (1973). Should we teach children syntax? *Language Learning*, *23*(2), 245–258. <https://doi.org/10.1111/j.1467-1770.1973.tb00659.x>

Gleitman, L. R., & Gleitman, H. (1992). A picture is worth a thousand words, but that's the problem: The role of syntax in vocabulary acquisition. *Current Directions in Psychological Science*, *1*(1), 31–35. https://doi.org/10.1111/1467-8721.ep10768738

Godwin-Jones, R. (2018). Contextualized vocabulary learning: Lessons from corpus analysis. *Language Learning & Technology*, *22*(3), 1–15. https://doi.org/10.1234/llt.2018.003

Gu, P. Y. (2010). Learning strategies for vocabulary development. *Cambridge Handbook of Second Language Acquisition*, 625–643. https://doi.org/10.1017/CBO9781139051729.035

Laufer, B. (1991). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *12*(3), 255–271. https://doi.org/10.1093/applin/12.3.255

Laufer, B. (1995). Beyond 2000: A measure of productive lexicon in L2 learners. *Canadian Modern Language Review*, *52*(1), 126–146. https://doi.org/10.3138/cmlr.52.1.126

Laufer, B. (1998). The development of passive and active vocabulary in second language learners: Same or different? *Applied Linguistics*, *19*(2), 255–271. https://doi.org/10.1093/applin/19.2.255

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Nemati, A. (2010). Active and passive vocabulary knowledge: The effect of years of instruction. *Asian EFL Journal*, *12*(1), 30–46.

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLoS ONE*, *10*(10), e0137041. https://doi.org/10.1371/journal.pone.0137041

Stockwell, G. (2016). Mobile language learning. *The Routledge Handbook of Language Learning and Technology*, 145–157. https://doi.org/10.4324/9781315657899-10

Xi, Y., & Geva, E. (2023). A 4-year longitudinal study examining lexical and syntactic bootstrapping in English language learners (ELLs) and their monolingual peers. *Developmental Psychology*, *59*(4), 687–702. https://doi.org/10.1037/dev0001417

charter