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)
  + 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 change of language over the years? (e.g. past 100 years)

**Research Questions**

1. What is the solution? What is our end result?
2. 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?
3. 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?
4. 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
5. 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?
6. Why is a data-driven approach necessary in linguistic analysis?
   1. Its big data
7. 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>

contacts

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

# Dr. Damien Williams

* UPDATE: Waiting for response, emailed on 2025/02/06

# 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. 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/)

# Preparation

1. Read her research
   1. Connect her research to our research; What is our motivation for coming to her for help?
   2. What has she been researching recently? (Most recent work was in 2011)
   3. Does she know french? (<https://www.unm.edu/~jbybee/downloads/Bybee2005LaLiaison.pdf>)

* However, because most linguistic change appears to be quite gradual, with slowly changing meanings and distributions and overlapping stages, a problem arises for a theory with discrete constituent structure. Evidence from the gradualness of change has led some researchers to doubt discrete categories and structures ([Haspelmath, 1998](https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9922.2009.00534.x#b19); [Hoffmann, 2005](https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9922.2009.00534.x#b21); [Quirk, Greenbaum, Leech, & Svartvik, 1985](https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9922.2009.00534.x#b36))
* Continuing from [Bybee and Scheibman (1999)](https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9922.2009.00534.x#b11), we join these researchers in proposing that constituent structure can change gradually. We take the view that it is altogether common even for an individual speaker to have nondiscrete syntactic representations for the same word sequence. Taking a complex systems-based perspective, we hold that syntactic structure is in fact much richer than the discrete constituency view would indicate. There are multiple overlapping and, at times, competing influences on the shape of units in the grammar, and these multiple factors have an ongoing effect on each speaker's synchronic representations of syntactic structure. Specifically, syntactic constituents are subject to ongoing influence from general, abstract patterns in language, in addition to more localized, item-specific usage patterns. The foregoing perspective makes it possible that the same word sequence may be characterized by multiple constituent structures and that these structures have gradient strengths rather than discrete boundaries. Our position in this article is thus that constituency may change in a gradual fashion via usage, rather than via acquisition, and that structural reanalysis need not be abrupt..

# Agenda

1. Situation report

NOTES

Pagel

Loan words are not cognates

Identify a swadesh list using data science

Estimate a person’s working lexicon

Consider inflexionally invariant words

Count and group? Lemmas

Social good: education, language didactic, nlp

Languages at crossroads are more innovative than languages far from crossroads (English vs. Icelandic)

Spoken language

coca corpora

Interview subtitle

Soap corpora

potential-sources

*Documents all of the potential literature sources — categorized by their relevant subject matter. Claim the source you’re reading by changing the font color to your associated color. 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.*

*Dibs for literature review — Samy Bakikerali, Gloria Chen, Bao Dinh, Dakota Ellis, Uyen Le*

archive

*Documents archived literature 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.*

# 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)*

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.

Gloria Chen

[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.

Bao Dinh

Dakota Ellis

[Advantages and Disadvantages of Globalization of the English Language for Ukrainian People - University of North Carolina at Charlotte](https://charlotte.primo.exlibrisgroup.com/discovery/fulldisplay?docid=cdi_crossref_primary_10_24025_2707_0573_9_2024_293258&context=PC&vid=01UNCC_INST:01UNCC_INST&lang=en&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,language%20globalization%20theory,AND&sortby=rank&mode=advanced&offset=0)

Uyen Le

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

# Abstract

# Introduction

# Background

# Data Description