Diachronic Language Models

Lecture 1 - Kick off

Natural Language Processing Group

Dr. Wei Zhao 15/10/2023

Outline

About this course

- Motivation
- Organization
  - Resources and Communication
  - Presentation
  - Term Paper

Motivation

Why are we learning this

The distinction between past, present and future is only a stubbornly persistent illusion.

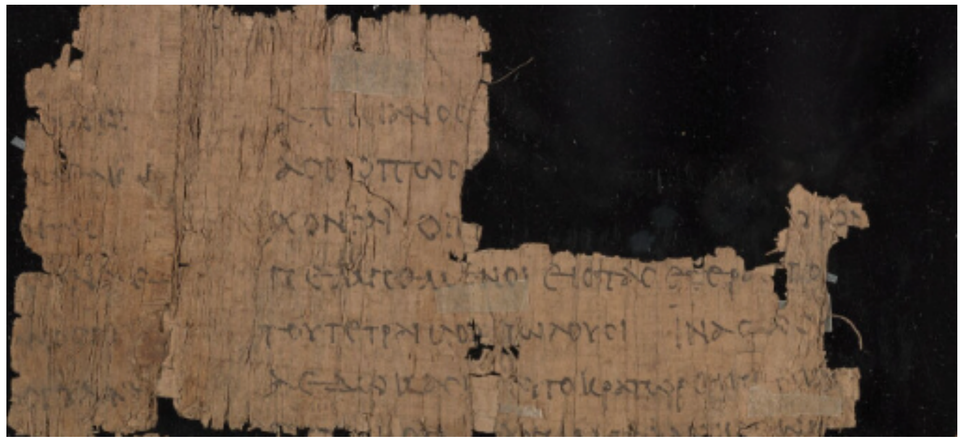
— Albert Einstein

Travel into the future

Motivation

Why are we learning this

We cannot physically travel to the past to know what was recorded in historical manuscripts. However, by using statistics-empowered diachronic language models, we can make a good guess about the contents of these manuscripts after the LMs study a large amount of textual data.

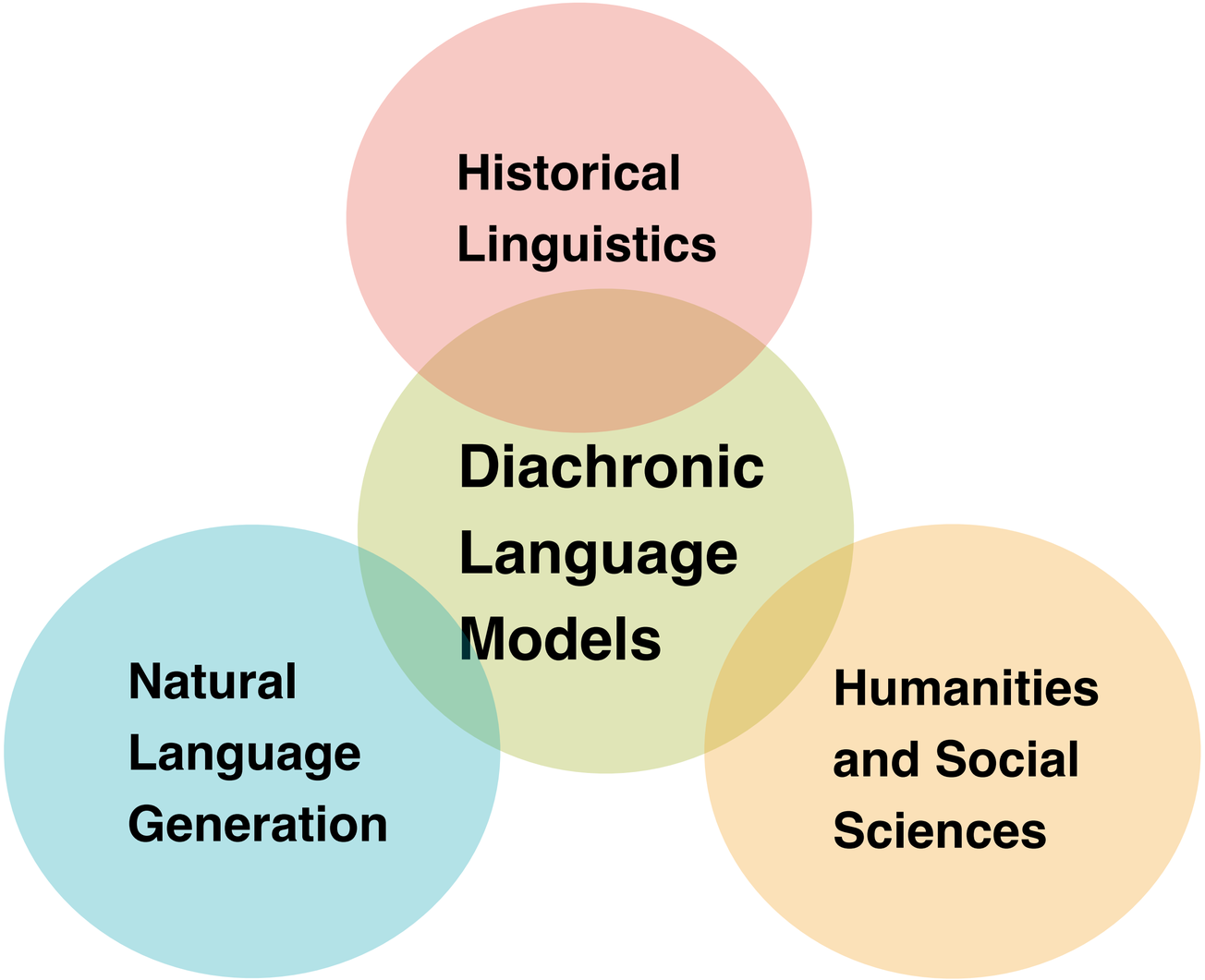


The cat is chasing the _____

- **dog** 5%
- **mouse** 70%
- **squirrel** 20%
- **boy** 5%
- **house** 0%

Seminar goals

Get to know DLMs and their implications



Lecturers

Who are we

- Dr. Wei Zhao
  - PhD in NLP at TU Darmstadt
  - Postdoc at HITS and Institute for Mathematics (by courtesy)
- Helen Chai
  - PhD student in CL at HITS and Heidelberg University

Feedback

It is your course

- Your feedback is important
- Constructive criticisms are welcomed
- We will post anonymous feedback forms twice at mid-term and the end of the course

Outline

About this course

- Motivation
- Organization
  - Resources and Communication
  - Presentation
  - Term Paper

Organization

About resources and communication

- Github for lecture slides and literature
  - https://github.com/andyweizhao/diaclms
- Slack for discussions and announcements
  - https://dlm-eau8707.slack.com/
- Email addresses
  - wei.zhao@h-its.org
  - haixia.chai@h-its.org
- Talk to us: in-person, slack, email anytime
- Office hours: Monday 13.00-14.00; Friday 13.00-14.00

Organization

To pass the course, you need to

- Give a presentation
- Active participation
  - Attendance (please sign up your name on the paper)
  - Write a paper review and prepare two questions
- Write a term paper

Your grade = 60% x term_paper + 30% x presentation + 10% x active participation

Organization

Presentation

- Choose a topic from the list and select 1-2 papers or a textbook chapter from our references.
- You will give a 30 minutes presentation + 15 minutes QA.
- Each lecture will have two presentations.
- Please fill out time lots. First come, first served.
  - Let us know if slots are sold out. We will release additional slots or organize three talks in a lecture.
  - https://docs.google.com/spreadsheets/d/1M6V_kFJpUR17jQJWV1xBR7zJKU9zdKW2tRFZitV_f_c/edit?usp=sharing
  - Deadline for this: 20, October (at least for the first presenter)

Organization

Presentation

- All non-presenters should prepare for the upcoming presentation
  - Read the materials selected by the presenters
  - Write a constructive review for the materials
  - Prepare at least two questions
  - Send them in email by Tuesday 12.00 before the next lecture
- For presenters on that day,
  - No need for a review and questions, just send us your slides

Organization

Presentation - a list of topics

#	Date	Topics	Disciplines	Presenters
1	17/10	Language change & Language models	CL	Wei Zhao
2	24/10	Speed of linguistic changes	HL	
3	31/10	Grammaticalization	HL	
4	7/11	Syntactic change	HL	
5	14/11	No seminar (GaML 2023)	-	
6	21/11	Large language models	NLP	
7	28/11	Diachronic language models	NLP	
8	5/12	No seminar (EMNLP 2023)	-	
9	12/12	Guest lecture on semantic change	CL	Dominik Schlechtweg
10	19/12	Semantic change detection	CL	
11	26/12	No seminar (Christmas)	-	
12	2/1	No seminar (Christmas)	-	
13	9/1	Syntactic change detection	CL	
14	16/1	Temporal machine translation	NLP	
15	23/1	Temporal text summarization	NLP	
16	30/1	Reconstruction of historical texts	NLP	
17	6/2	Temporal misinformation detection	CSS	

After Christmas I will be not in Heidelberg, so the sessions from Jan to Fed will take place in the hybrid format.

Organization

Presentation - a list of topics

#	Date	Topics	Discipline
1	17/10	Language change & Language models	CL
2	24/10	Speed of linguistic changes	HL
3	31/10	Grammaticalization	HL
4	7/11	Syntactic change	HL
5	14/11	No seminar (GaML 2023)	-
6	21/11	Large language models	NLP
7	28/11	Diachronic language models	NLP
8	5/12	No seminar (EMNLP 2023)	-
9	12/12	Guest lecture on semantic change	CL
10	19/12	Semantic change detection	CL
11	26/12	No seminar (Christmas)	-
12	2/1	No seminar (Christmas)	-
13	9/1	Syntactic change detection	CL
14	16/1	Temporal machine translation	NLP
15	23/1	Temporal text summarization	NLP
16	30/1	Reconstruction of historical texts	NLP
17	6/2	Temporal misinformation detection	CSS

README.md

References

- Speed of linguistic changes
  - Dialect contact and the speed of Jespersen's cycle in Middle Low German
  - The Determinants of Diachronic Stability
- Syntactic change
  - Complexity as L2-difficulty: Implications for syntactic change
  - Syntactic change: A Minimalist Approach to Grammaticalization
- Grammaticalization
  - Introduction to Grammaticalization
  - What is it then, this Grammaticalization?
  - The Formal Semantics of Grammaticalization
  - What counts as (an instance of) grammaticalization?
  - Information-based Modeling of Diachronic Linguistic Change: from Typicality to Productivity
- Diachronic language models
  - Dynamic Word Embeddings
  - Contextualized diachronic word representations
  - Dynamic Contextualized Word Embeddings
- Large language models
  - A survey on evaluation of large language models
  - A Survey of Large Language Models
  - Chatgpt: A meta-analysis after 2.5 months
- Semantic change detection
  - Diachronic word embeddings and semantic shifts: a survey
  - Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models
  - "Cultural shift or linguistic drift? comparing two computational models of semantic change
  - Diachronic word embeddings reveal statistical laws of semantic change
  - Grammar and meaning: Analysing the topology of diachronic word embeddings
  - What about Grammar? Using BERT Embeddings to Explore Functional-Semantic Shifts of Semi-Lexical and Grammatical Constructions
  - Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach

Organization

How to structure a presentation

For a research paper presentation, the focus is to present research findings but they are not always true.

Thus, consider this structure:

- Background
- Research Questions
- Methodologies
- Datasets
- Results and Analyses
- Conclusions

Organization

How to structure a presentation

For a textbook chapter presentation, the focus is to present basic facts that are true.

Thus, consider this structure:

- Background
- Fact 1
- Fact 2
- Fact 3
- Conclusions

Organization

How we evaluate a presentation

- Clarity: Use clear language that is easy to understand
- Entertainment with Pictures: Use visual aids such as images, charts and diagrams to illustrate key points.
- Answering Questions
  - Prepare some questions in advance
  - Think about the concise and to-the-point answers before responding

Organization

About writing a term paper

- Work in groups of 2 people on a topic from our recommendation list.
- Each group writes a paper of 8 pages of content + unlimited pages of references.
- We accept three paper types:
  - Survey paper (read a lot)
  - Analysis paper (empirical study)
  - Position paper (be critical and creative)
- We will provide the template for term papers on Slack

Organization

How we evaluate a survey paper (read a lot)

Provide a structured overview of the literature to date on a topic.

- Readability: Is the paper easy to follow and well-structured?
- Substance: Does the paper include sufficient amount of content that presents your own opinions and ideas?
- **Thoroughness**: Does the paper include most of the relevant literature from the past to date?
- Structure: Background, Related Work, Methods, Datasets, Outlook, Conclusions
- Example 1: https://aclanthology.org/C18-1326.pdf
- Example 2: https://aclanthology.org/C18-1117.pdf

Organization

How we evaluate an analysis paper (empirical study)

Provide new insights by conducting empirical study on a topic.

- Readability: Is the paper easy to follow and well-structured?
- Substance: Does the paper include sufficient amount of content that presents your own opinions, ideas and results?
- **Soundness**: Can one trust the claims of the paper - are they justified by empirical results?
- Structure: Background, Research Questions, Experimental Analyses, Conclusions
- Example 1: https://aclanthology.org/C18-1325.pdf
- Example 2: https://aclanthology.org/W19-4722.pdf

Organization

How we evaluate a position paper (be critical and creative)

Articulate open issues on a topic, present your stance and a futuristic vision.

- Readability: Is the paper easy to follow?
- Substance:Does the paper include sufficient amount of content that presents your own opinions and ideas?
- **Soundness**: Is the position of the paper clear? Foundations of your position?
- Structure: Background, Open Issues, Your Stance and Vision
- Example 1: https://aclanthology.org/C18-1324.pdf
- Example 2: https://aclanthology.org/C18-1094.pdf

Organization

Deadlines for term papers

- 12, March: Preliminary version of your paper (not required)
- 19, March: Final version of your paper
- We will select the top-three term papers at the end of the course.

Topics for Term Papers

A list of topics we recommend

Use language models to detect

- Speed of linguistic change over time
- Grammaticalization in language change
- Structural and grammatical changes in dialect continua
- Syntactic change over time
- Semantic change over time
- Dynamics of misinformation on social media over time.

Topics for Term Papers

A list of topics we recommend

Use language models to perform

- Temporal text summarization
- Temporal machine translation
- Reconstruction of incomplete historical manuscripts

Topics for Term Papers

A list of topics we recommend

A diachronic analysis of ChatGPT:

- Does it understand the past, and the changes from the past to present?
- Does it understand the development of human language?

Topics for Term Papers

How we assign topics to you

- Cast your vote on each topic at https://forms.gle/PxJeRZGfftZei4qE9
  - Let us know if you want to add a new topic
  - Deadline for this: 24, October
- We will pair students to topics based on their votes
- It is fine if many of you work on the same topic, but each group will only have two people
- If you are alone for some reasons, we will try to pair you with another student. If not successful, you can write a paper of 4 pages. Otherwise, write 8 pages.
- Let us and your peer know timely if you decide to quit.
- Discuss your topic with us before starting to write

Plagiarism

This is Forbidden!

- With any plagiarism, you will fail the course.
- Behaviors considered plagiarism (https://www.plagiarism.org/article/what-is-plagiarism)
  - Turning in someone else's work as your own
  - Copying words or ideas from someone else without giving credit
- About ChatGPT, you can use it as a search engine to gain knowledge, but not to write a term paper.
  - We will use a powerful tool to detect such plagiarisms!
- Know why you are here.

To-do Items (Shopping Cart)

All your tasks and deadlines

- 20, October: Fill out time slots and reading materials for your presentations at https://docs.google.com/spreadsheets/d/1M6V_kFJpUR17jQJWV1xBR7zJKU9zdKW2tRFZitV_f_c/edit?usp=sharing
  - Rolling basis for the later presenters
- 24, October: Vote on topics for your term papers at https://forms.gle/PxJeRZGfftZei4qE9
- By Tuesday 12:00, two presenters should submit their slides
- By Tuesday 12:00, non-presenters should read the materials selected by presenters, submit a constructive review, and prepare two questions
- 12, March: All should submit the preliminary version of your paper (not required)
- 19, March: All should submit the final version of your paper

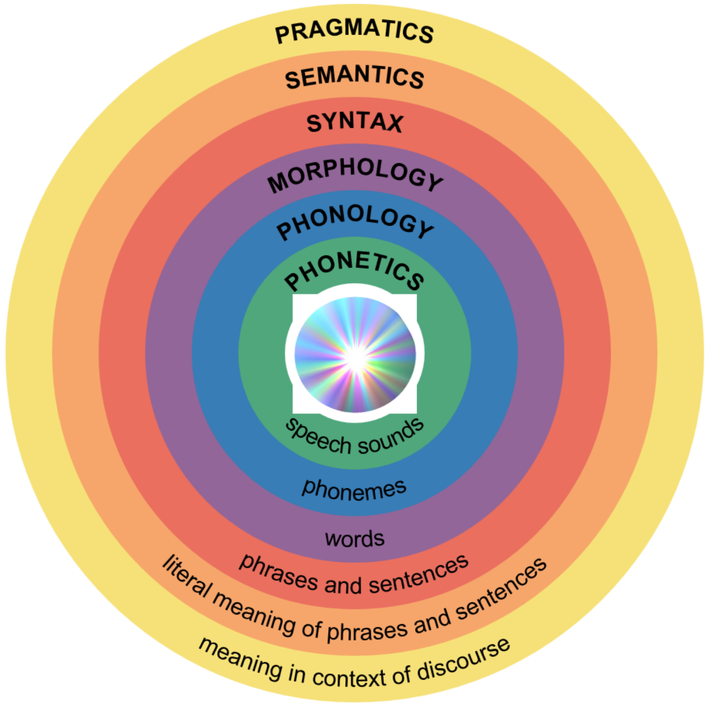
Diachronic Language Models

Lecture 1 - Introduction to language change & language models

Dr. Wei Zhao 15/10/2023

Outline

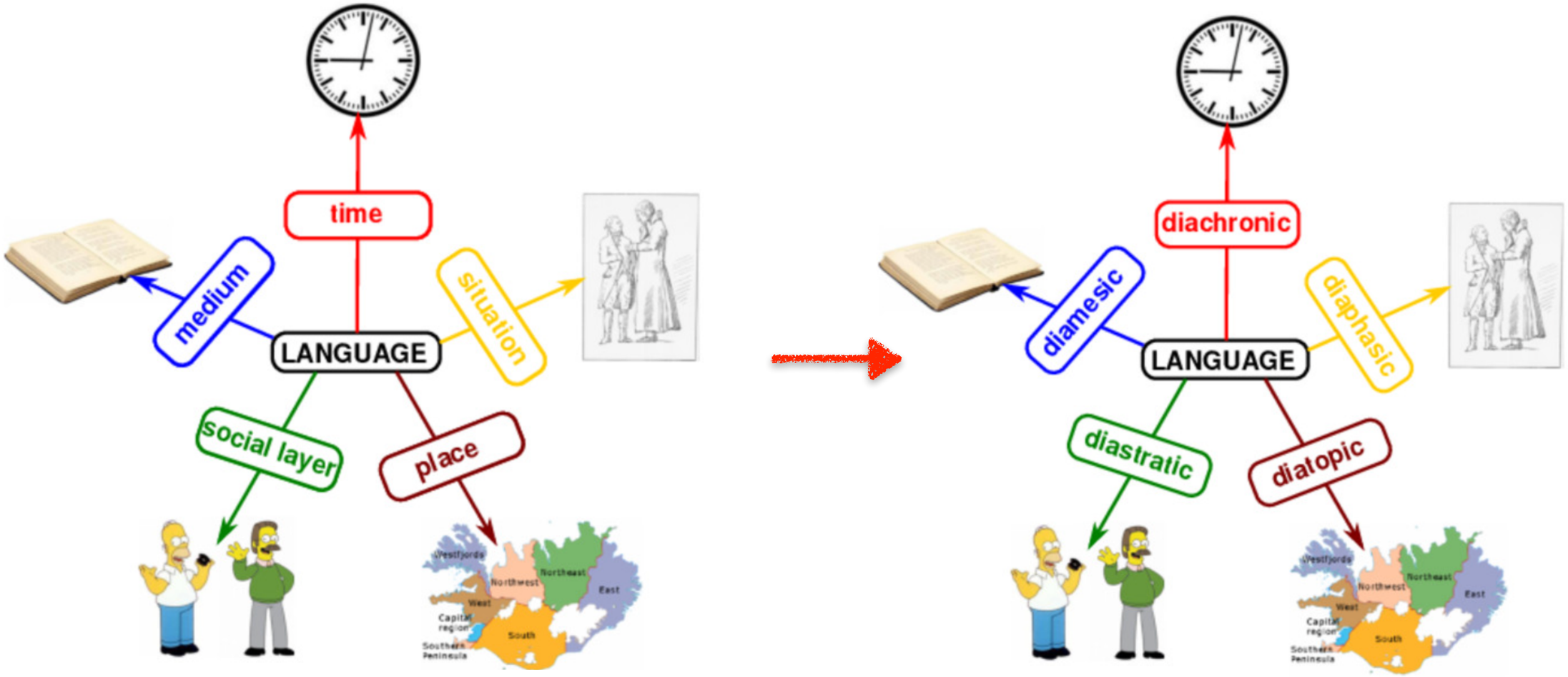
Foundations of diachronic language models



- Language Change - a subject of study in linguistic change over time
- Language Modeling
  - Embedding: a computer algorithm that translates dictionary to coordinates in vector space
  - Generative: a time-series-ish algorithm to generate human-like text.

Language Change

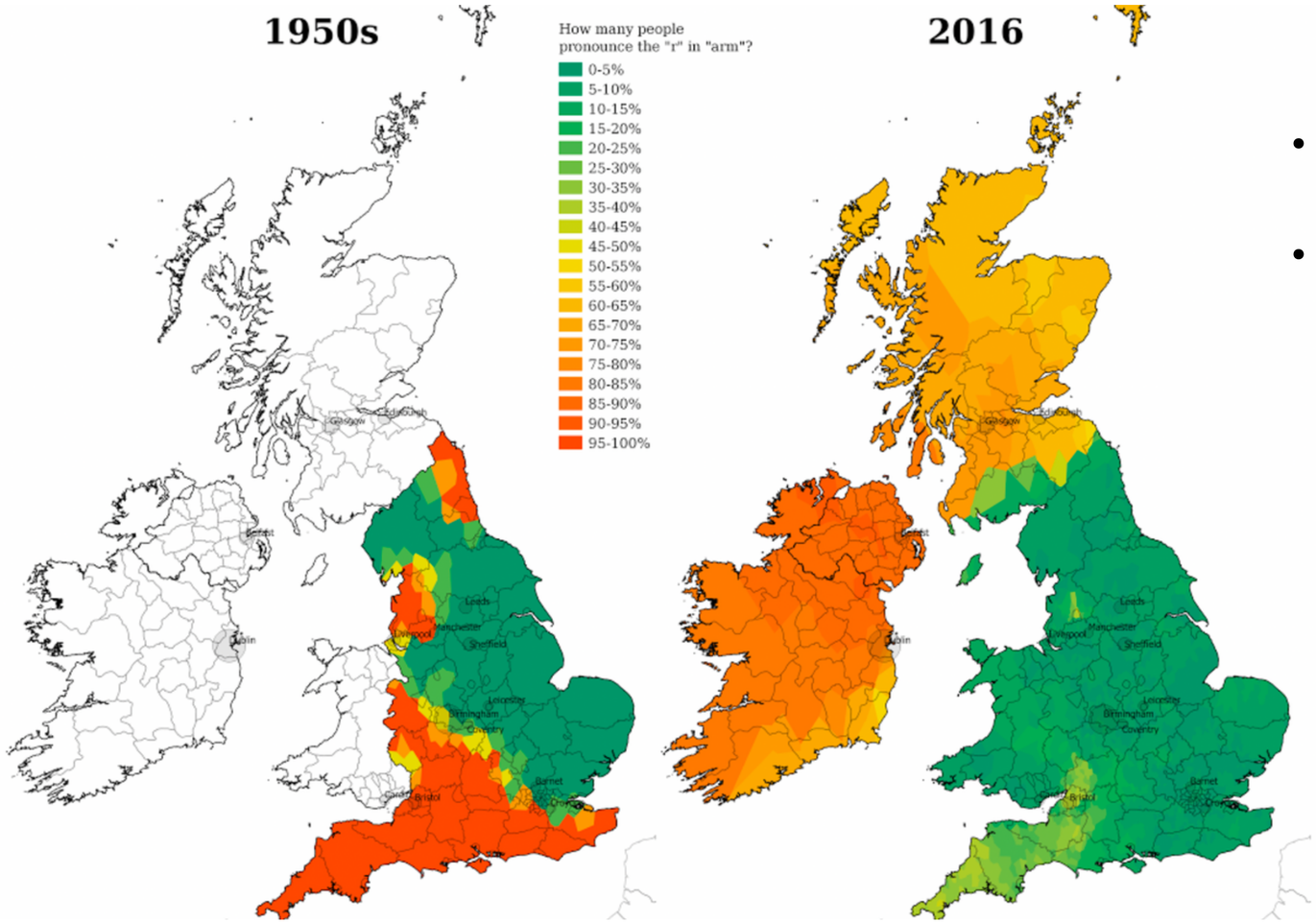
Five dimensions by Eugenio Coșeriu



http://phylonetworks.blogspot.com/2015/06/the-diasystematic-structure-of.html

Language Change

Examples: changes in sounds

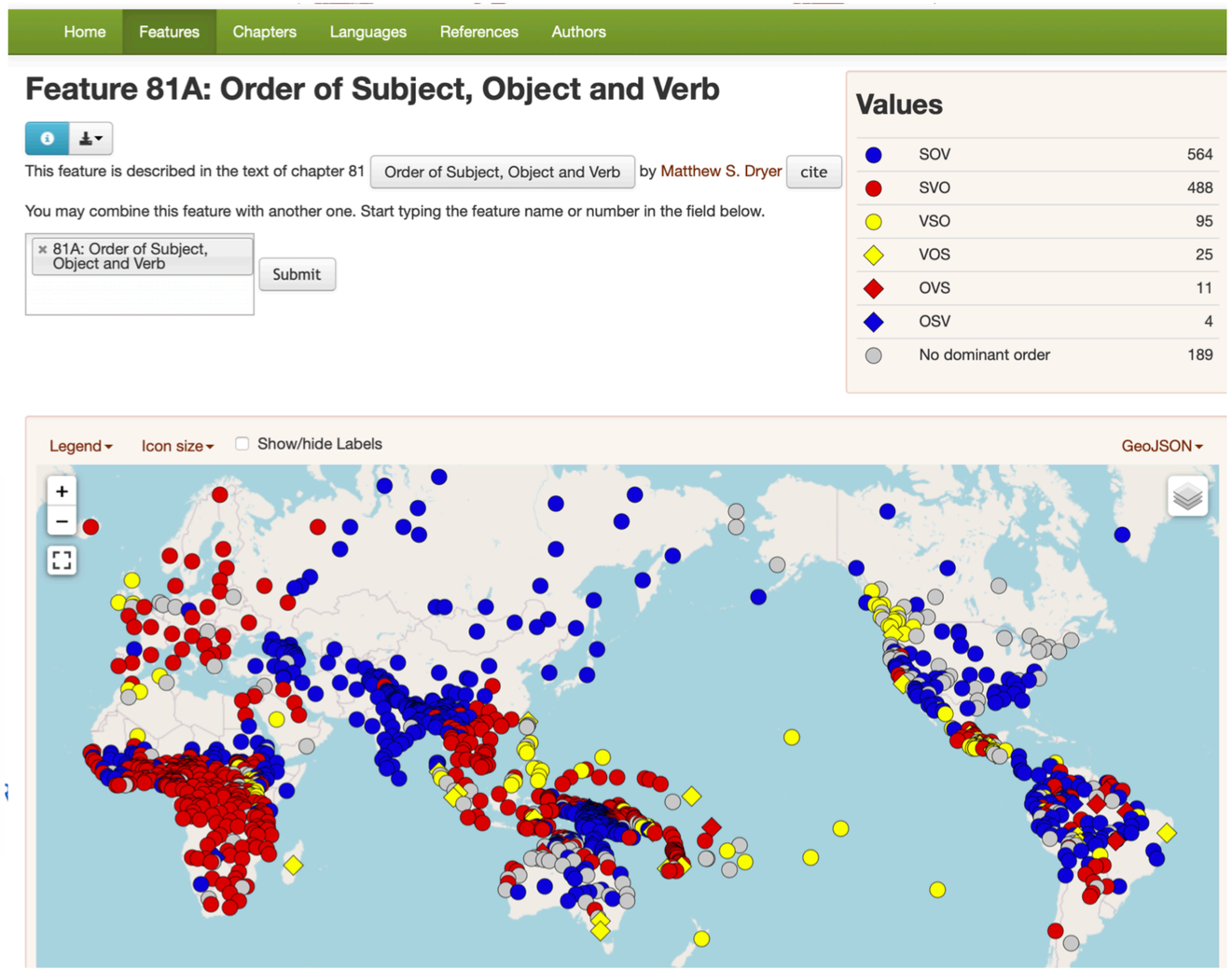


- TIME, PLACE (DIACHRONIC, DIATOPIC)
- How many people pronounce ‘r’ ? Transformation of `r’ from regional dialect in England to a national standard in the UK

https://www.reddit.com/r/MapPorn/comments/occ0c3/how_many_people_pronounce_the_r_in_arm/

Language Change

Examples: difference in word order



https://wals.info/

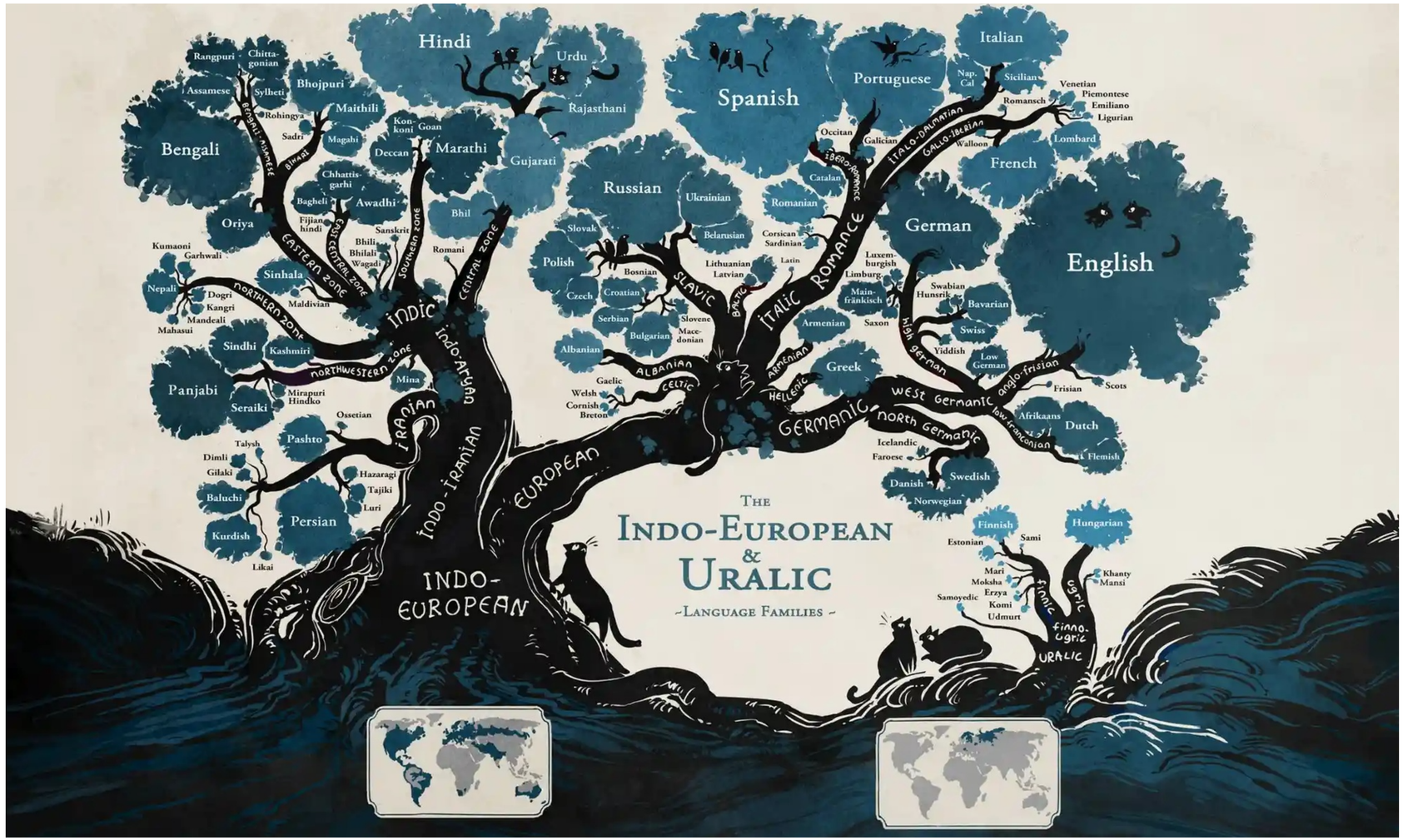
Statistics: the number of languages for each word order.

- PLACE (DIATOPIC)

- Africa: SVO
- Central Europe: SVO
- Asia: SOV
- Australia: SOV, VOS, SVO, No dominant order
- South America: SOV, SVO
- North America: SOV, VSO, No dominant order

Language Change

A phylogenetic tree is a result of language change



https://www.theguardian.com/education/gallery/2015/jan/23/a-language-family-tree-in-pictures

- How languages have diverged from common ancestral languages at roots?
- The tree is constructed on linguistic similarities in words, sounds and grammar.
- The size of leaves indicate the population of a language.
- Despite being geographically close, Finnish is quite different from Swedish.

Language Change

Weinreich, Labov & Herzog (1968)

Focus: Inform an empirical theory of language change that should address the following problems:

- Constraints Problem
- Transition Problem
- Embedding Problem
- Evaluation Problem
- Actuation Problem

Language Change

Weinreich, Labov & Herzog (1968)

- Constraints Problem
  - What are constraints that underlie the changes of a language (laws of change)?
- Examples (laws)
  - Morphology (rich -> simple)
  - Word order (flexible -> fixed)
  - But why? (Communication efficiency)

Language Change

Weinreich, Labov & Herzog (1968)

- Language models for constraints problem
  - Detect the laws of linguistic change (Hamilton, Leskovec, and Jurafsky, 2016)
    - Law of conformity: Meanings of words are more likely to change if the words are used less frequently.
    - Law of innovation: Meanings of words change faster with more senses.

Language Change

Weinreich, Labov & Herzog (1968)

- Transition Problem
  - How a language evolves over time across different stages?

Old English
500 - 1150

A. Flexible word order
B. Rich inflectional system

Middle English
1150 - 1500

A. Loss of inflectional endings
B. Reliance on word order

Modern English
1500 - Present

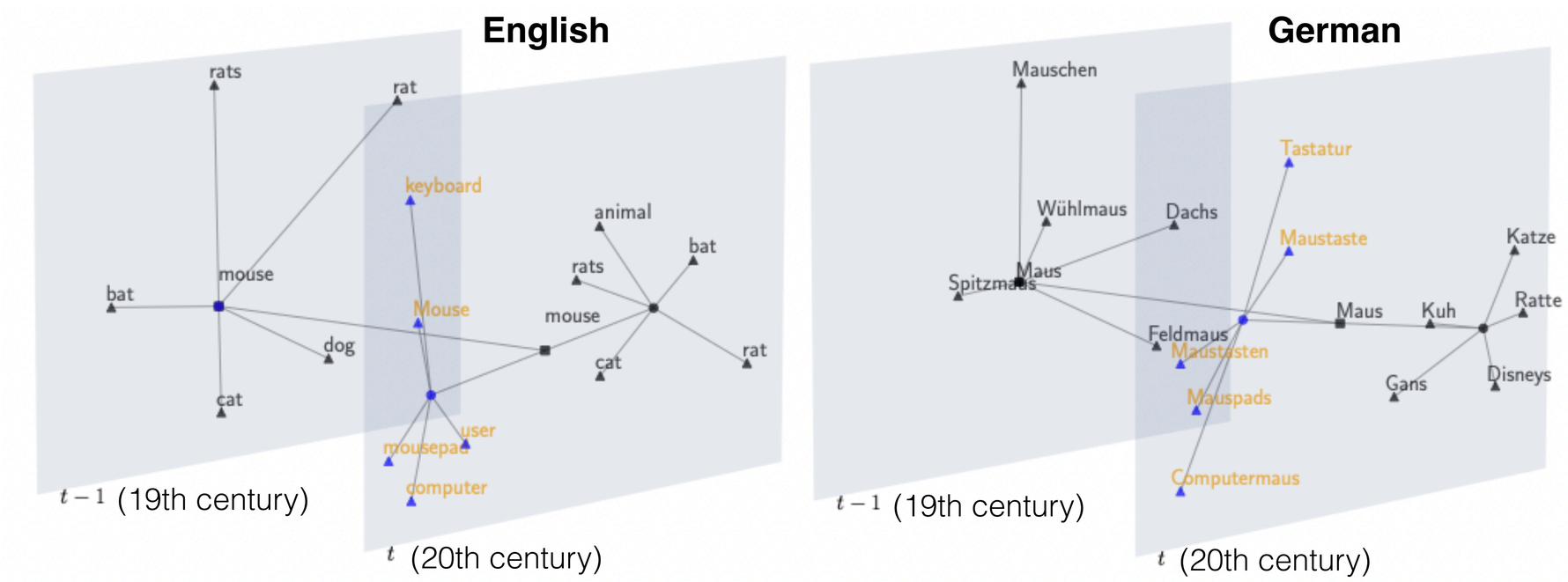
A. Subject-Verb-Object
B. Simplified morphology

See changes in words, sounds and grammar, such as the loss of gender system and dative case over time.

Language Change

Weinreich, Labov & Herzog (1968)

- Language models for transition problem
  - Detect linguistic changes such as in semantics across different stages

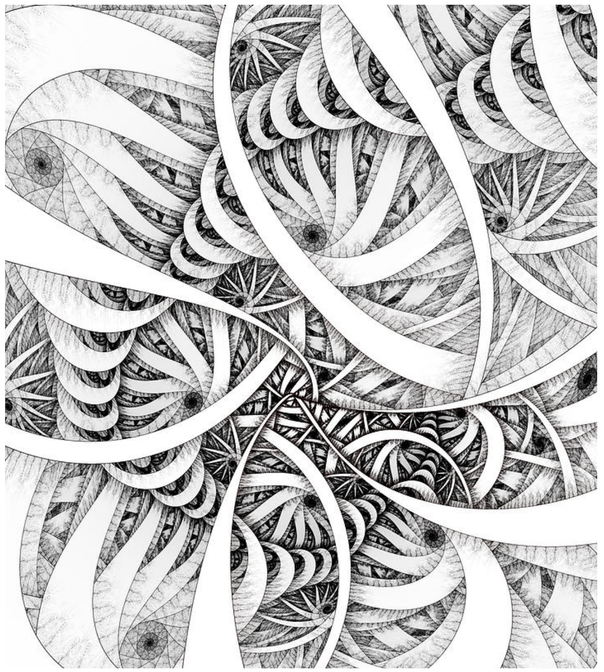


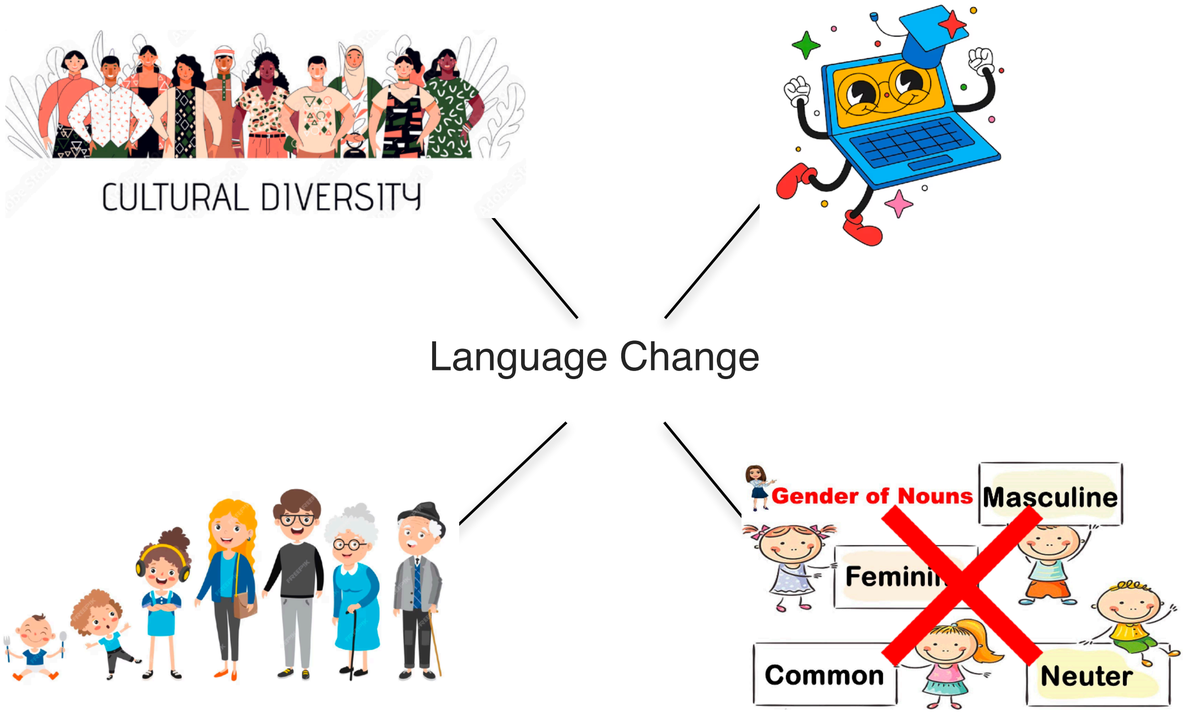
Ongoing work

Language Change

Weinreich, Labov & Herzog (1968)

- Embedding Problem
  - What linguistic and extralinguistic factors influence the change embedded in the language system?





Language Change

Weinreich, Labov & Herzog (1968)

- Language models for embedding problem
  - Detect the factors that influence linguistic change
  - Quantify the degree to which MOR, SYN and SEM changes interact?

Language Change

Weinreich, Labov & Herzog (1968)

- Evaluation Problem
  - What are the impacts of language change on communication efficiency?
- Examples
  - How to evaluate the impact of the loss of gender system for Modern English?

Language Change

Weinreich, Labov & Herzog (1968)

- Language models for evaluation problem

Language Change

Weinreich, Labov & Herzog (1968)

- Actuation Problem
  - Why a change occurs at a particular time in a particular language?
- Examples
  - Why the new meaning of mouse is gained in 1960s in English rather than in Asian languages?

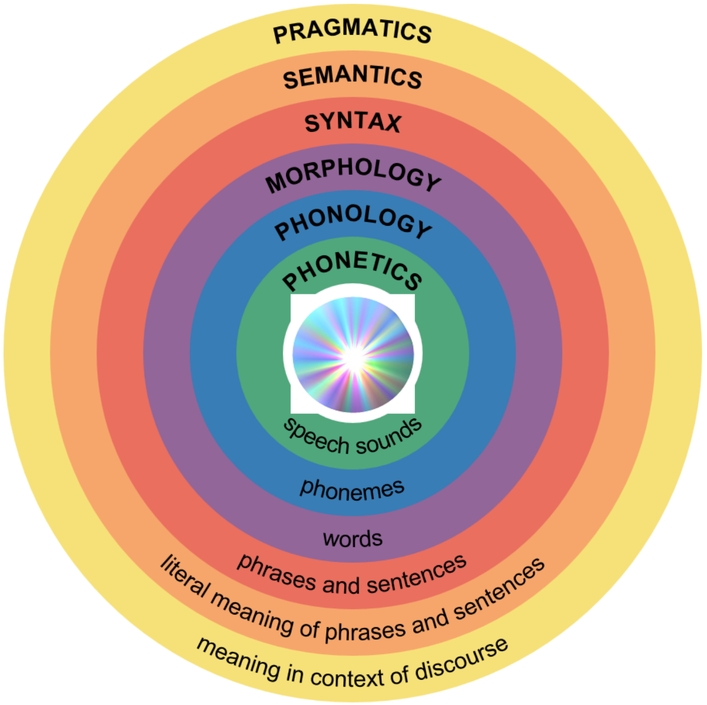
Language Change

Weinreich, Labov & Herzog (1968)

- Language models for actuation problem
  - Detect & compare the speed of change across languages

Outline

Foundations of diachronic language models



- Language Change - a subject of study in linguistic change over time
- Language Modeling
  - Embedding: a computer algorithm that translates dictionary to coordinates in vector space
  - Generative: a time-series-ish algorithm to generate human-like text.

Language Embedding

What this computer algorithm means

Computer Scientists

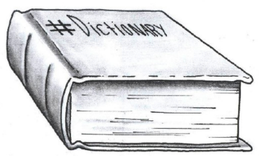
Translate language into a specific format that is comprehensible to computers, enabling them to perform tasks on a large scale with much greater efficiency compared to linguists.

Language Embedding

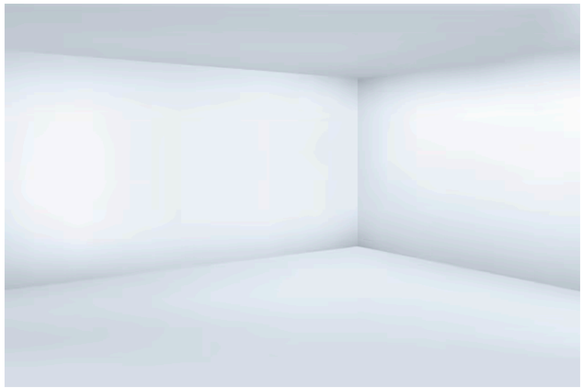
How it works

Translate lexicons into coordinates in space.

Computer Scientists



Lexicons



Coordinates in space

Where should coordinates be positioned in space?

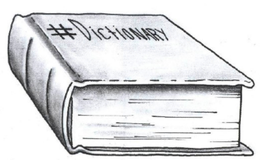
Language Embedding

How it works

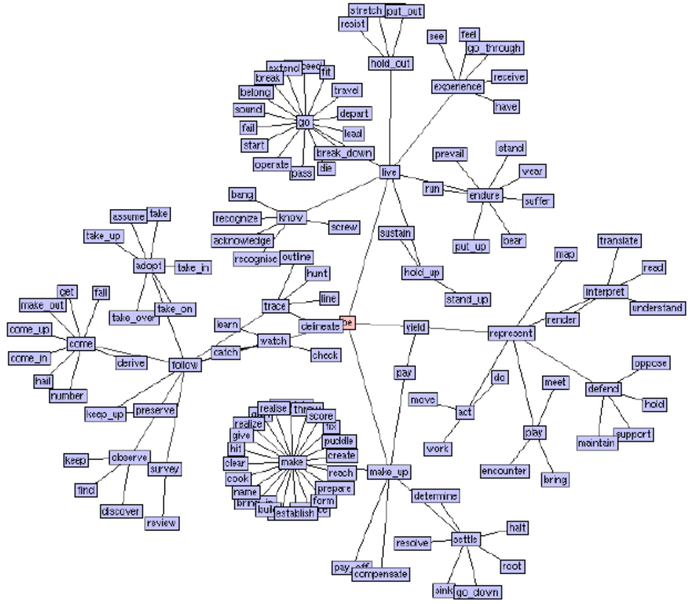
Translate lexicons into coordinates in space while maintaining lexical-semantic relations.

- Meronymy (has-a): part-whole relationship
- Hyponym and hypernym: is-a relationship

Computer Scientists



Lexicons



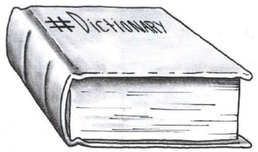
Coordinates in space

Language Embedding

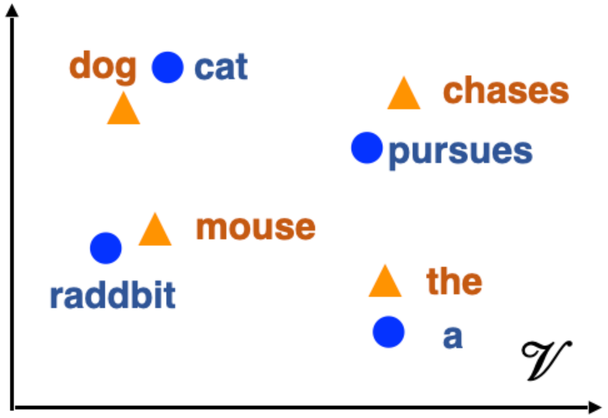
How it works

Translate lexicons into coordinates in space while maintaining functional relations.

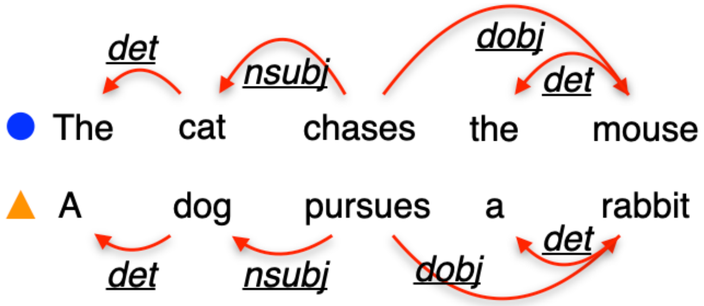
Computer Scientists



Lexicons



Coordinates in space



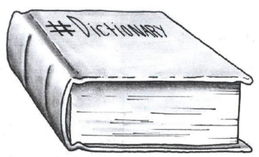
- cat (subj) = dog (subj)
- mouse (obj) = rabbit (obj)

Language Embedding

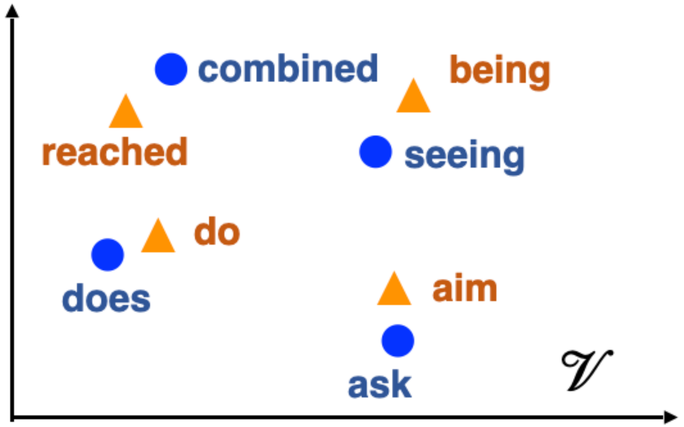
How it works

Translate lexicons into coordinates in space while maintaining morphological relations.

Computer Scientists



Lexicons



Coordinates in space

reached	combined	Tense=Past, VerbFrom=Part
aim	ask	VerbForm=Infinitive
do	does	VerbFrom=Finite
seeing	being	VerbForm=Gerund

- reached (pp.) = combined (pp.)
- being (gerund) = seeing (gerund)

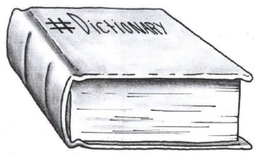
Language Embedding

How it works

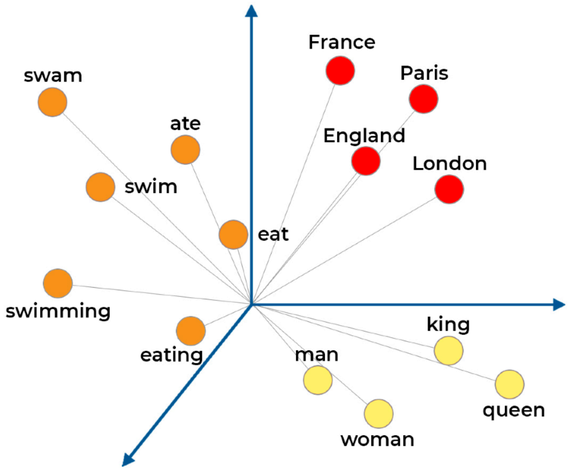
You shall know a word by the company it keeps.

— John Rupert Firth

Computer Scientists



Lexicons



Coordinates in space

Example: London ~ Paris

John lived in London.

John lived in Paris.

John lived in England.

Motif pattern: [person] lived in [city]

Two words are said to be similar if they frequently appear in similar contexts

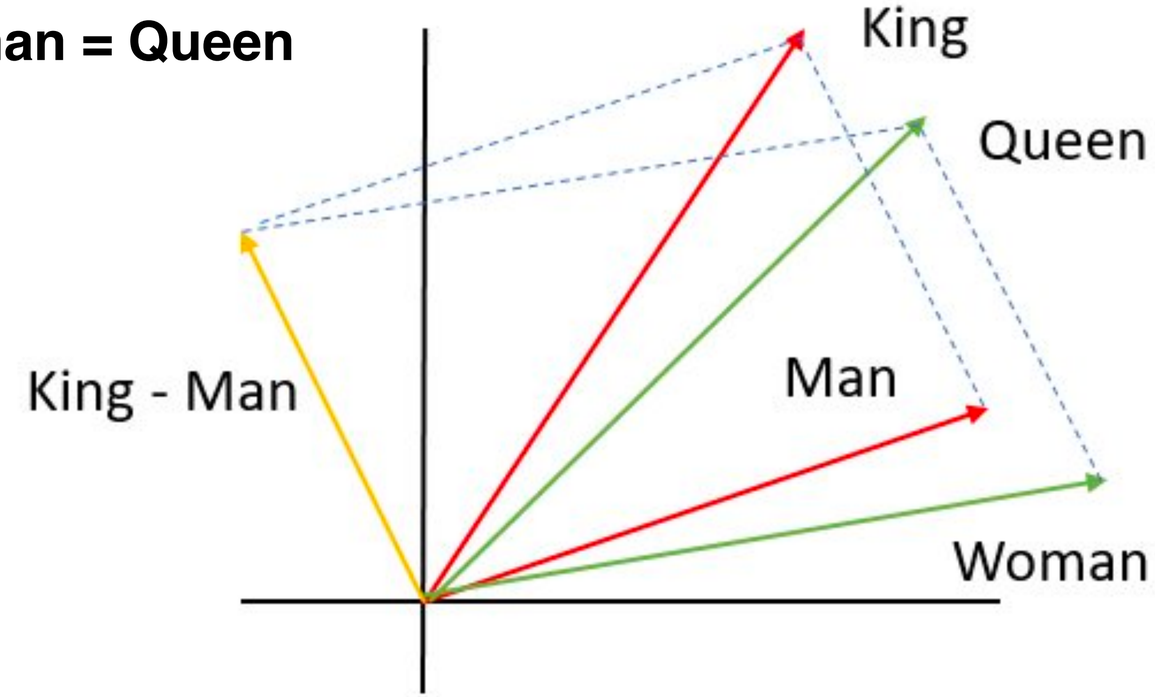
Language Embedding

How it interprets word meanings

Translate lexical semantics into the geometric relations over coordinates.

King - Man + Woman = Queen

Computer Scientists

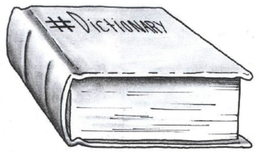


Language Embedding

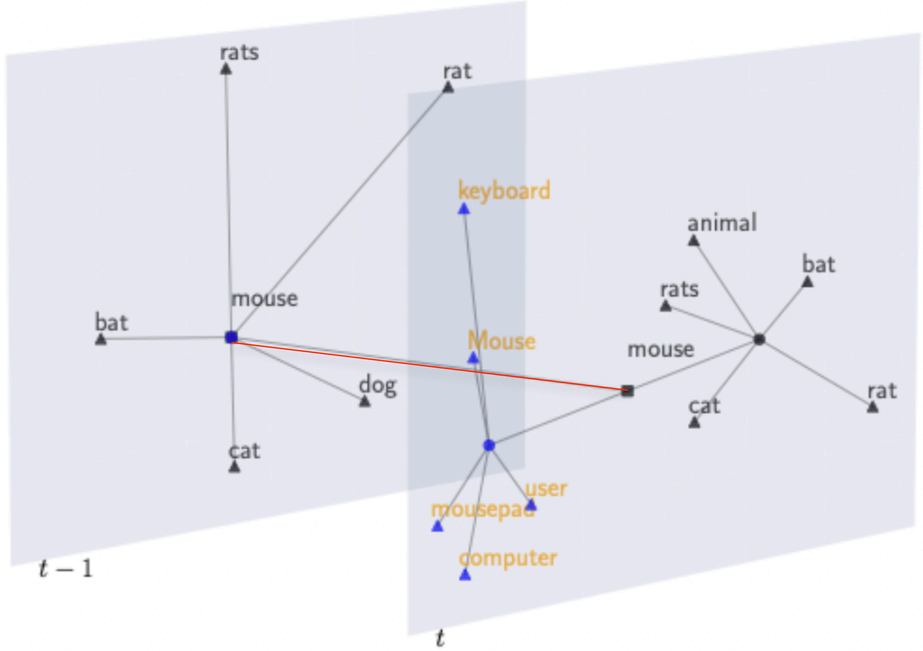
When it comes to diachrony

- Lexicons undergo changes in semantics, syntactic roles, inflectional forms.
- Translate lexicons into coordinates in a sequence of interconnected spaces that account for the changes over time.

Computer Scientists



Lexicons



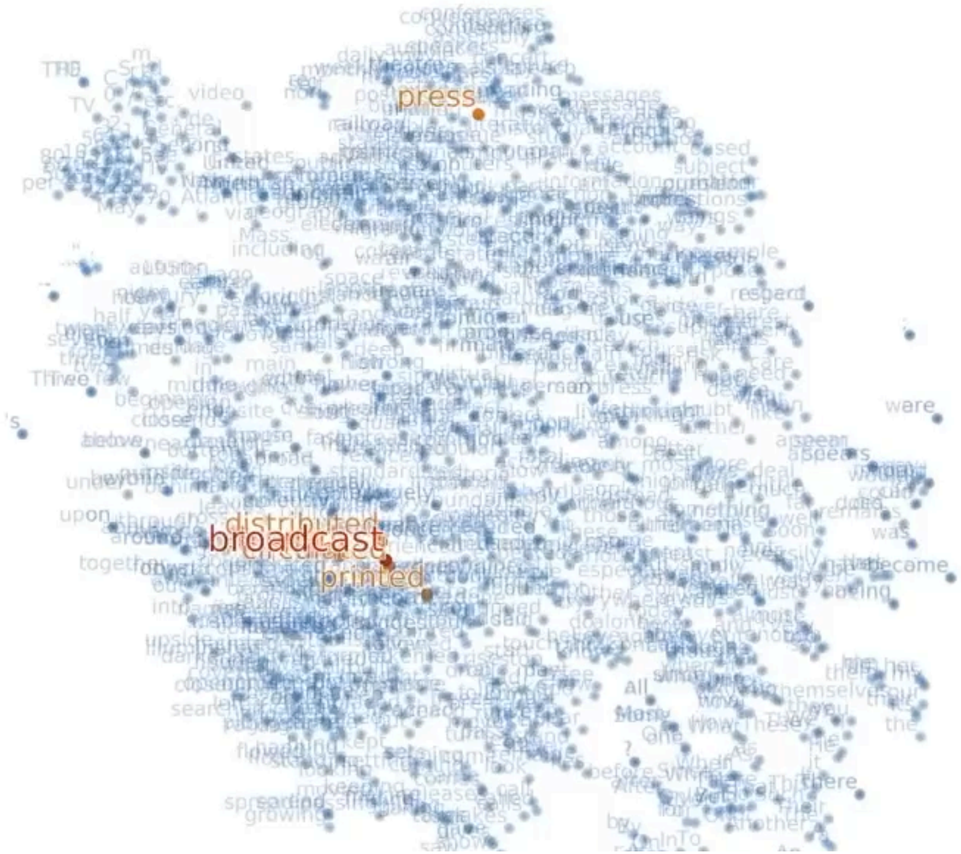
Coordinates in space

Language Embedding

When it comes to diachrony

1930

broadcast
circulated
press
distributed
shipped
printed



© Disney

Robert Bamler and Stephan Mandt, Dynamic Word Embeddings, ICML 2017.

Take-aways

Things to remember

- Language change is a complex development.
- Language varies across time periods, social groups, registers, geographical locations, and mediums.
- Linguistic and extralinguistic factors are impactful for language change.
- Linguistic change does follow laws.
- A linguistic change may be consistent or diverge across languages.
- Language models: embedding and generative parts.