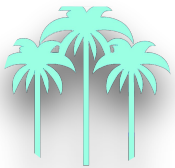
Examining Semantic Change in the Dutch East India Company's Ethnic Categories: When Computational Linguistics Meets Colonial History



Jiaqi Zhu
PhD Candidate
DHLab, Humanities Cluster, KNAW, the Netherlands
jiaqi.zhu@dh.huc.knaw.nl

27 March
ESSHC 2025
Leiden, the Netherlands

DHLAB

KNAW Humanities Cluster

Disclaimer

This presentation contain derogatory words and phrases. They are provided solely as illustrations of the research results and do not reflect the opinions of the authors or her organisation. In-text examples of derogatory and potentially offensive are presented in ***“quotes, boldfaced and italicised”***. The author is aware of the colonial biases in these words and phrases.

The distributional hypothesis in linguistics

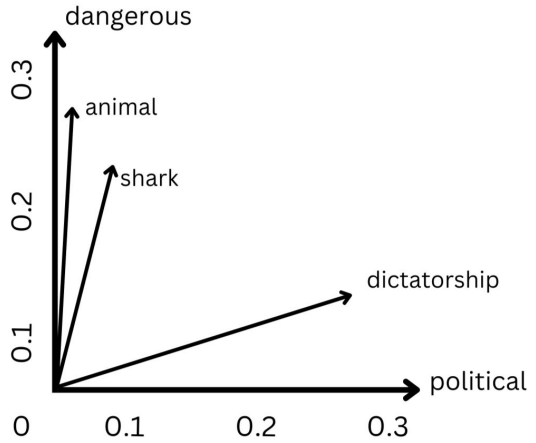
The hypothesis is that words with similar distributions have similar meanings.

Illustrative Example:

Consider the words "puppy" and "kitten."

Both often appear in contexts with words like "cute," "playful," "pet," and "adorable." This shared contextual environment suggests a semantic similarity between "puppy" and "kitten," as both refer to young animals commonly kept as pets.

Conversely, the word "puppy" rarely shares contexts with words like "engine" or "quantum," indicating a lack of semantic similarity.

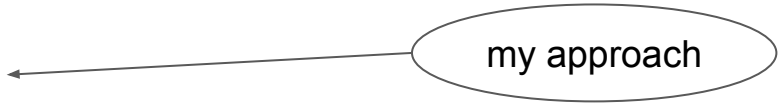


How words are related in a given language is demonstrated in the "semantic space", which mathematically corresponds to the vector space.

What is lexical semantic change in linguistics?

Definitions in linguistics (Geeraerts et al., 2023):

Semasiological change:



looks from a word to its meanings; it studies **polysemy**, like the various senses of *underground*.

Onomasiological change:

reverses the perspective and describes how a given meaning can be expressed by various words, like the **synonymy** of *underground* and *subway* in the ‘subterranean railway’ sense.

What is the VOC?

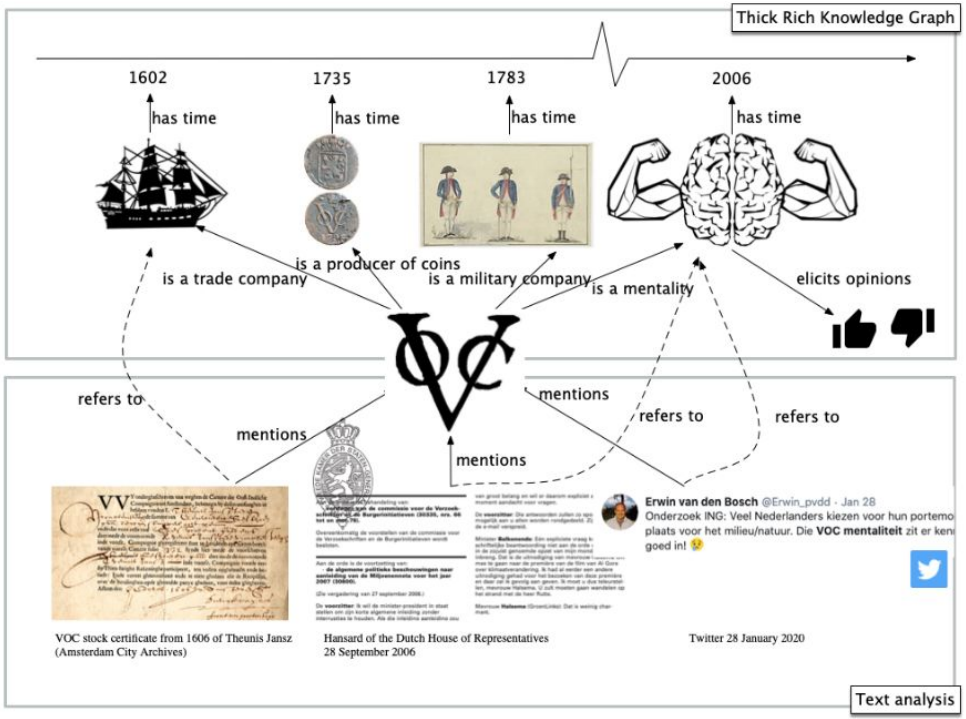
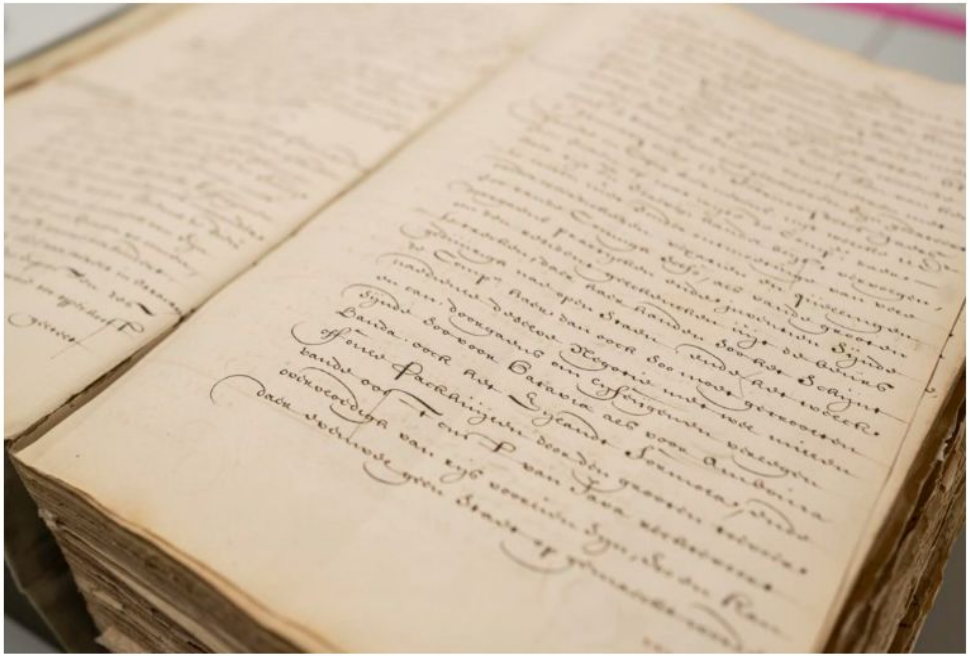


Figure 1: Schematic illustration of meanings VOC can take on over time

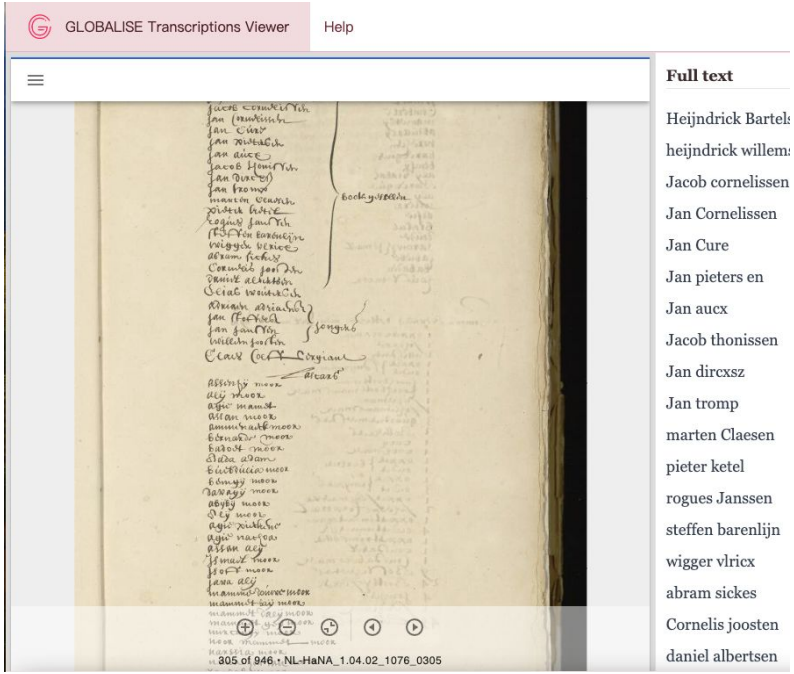
Oost-Indisch Huis

Heren XVII's meeting room in the Oost-Indisch Huis in Amsterdam

What are the VOC archives ?



A page from the so-called 'General Letters', a collection of summarising reports within the OBP.
Photo Dave Straatmeyer



A screenshot of Globalise Transcriptions Viewer

Ethnic, Religious, and Caste Categories in the VOC archives

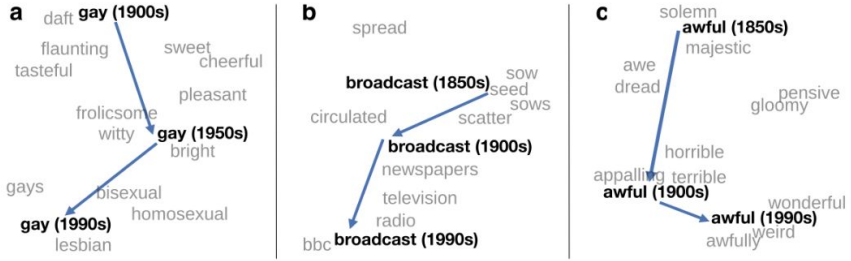
- The VOC archives reflect “the concerns and interests of colonial administrators” (Raben, 2019), in which we find great ethnic diversity and various labels or representations of ethnic, religious, and caste groups in Asia.

- However, these labels are complex and contentious, since
  1) the boundaries between the identity labels used to designate Asian peoples in the archives are ambiguous, which suggests that the criteria by which Company agents grouped people together or drew distinctions between them were varied and situational, instead of clear and consistent. (van Meersbergen, 2021).
  2) objectifying, othering, and Eurocentric perspectives

Why this research?

Hypothesis: Semantic changes often align with historical events or cultural shifts. Detecting semantic change with computational methods helps to contextualize keywords within larger narratives.

Example:



Two-dimensional visualization of semantic change in English (words “gay”, “broadcast”, and “awful”).
See Hamilton et al., 2016.

Aim: provide more historical contexts for these categories within larger narratives by revealing historical shifts in society or culture hidden by changes in word meanings.

How? find key ethnic categories to explore

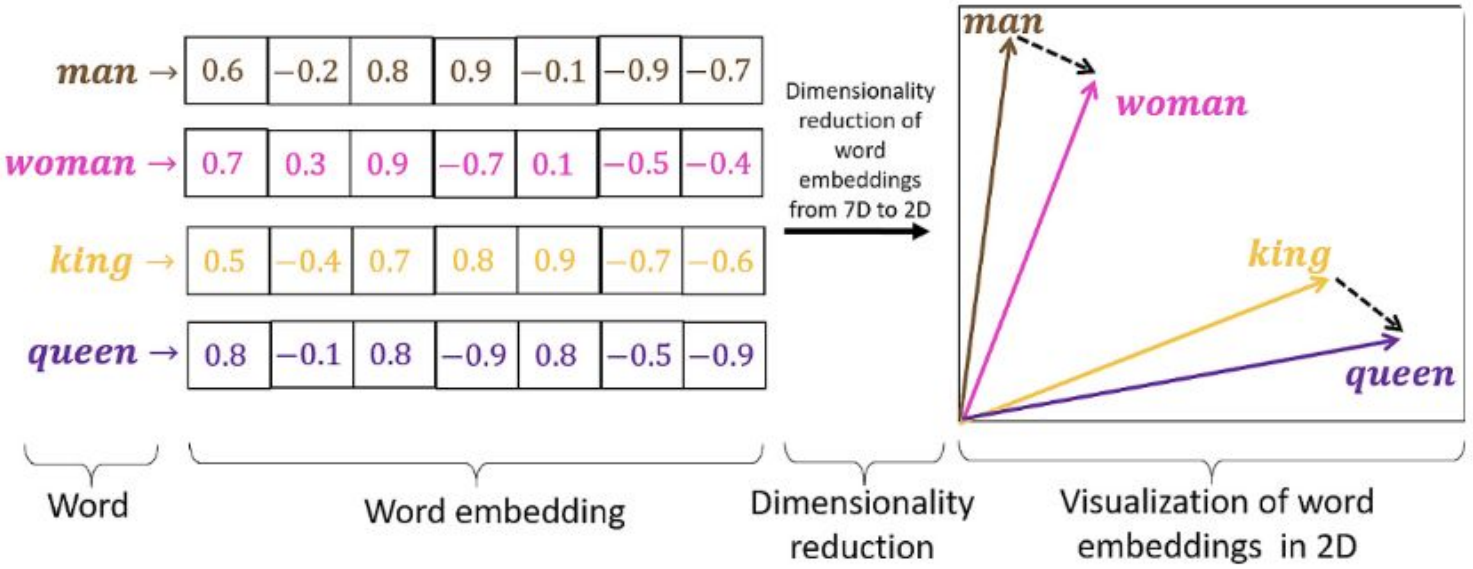
Which ethnic/religious/caste categories have experienced semantic change/are polysemical in the 17th and 18th centuries' VOC context?

We need domain knowledge from a historian

These categories are: ***"moor", "tartar", "mogol", "inboorling", "inlander", "kling", "papoeaas", "toradjeners", burger***

"Inlander"		Ethnic category	a. one who belongs to the country (in question, or thought of), someone from the country itself; a native, resident, "citizen" of the country; the opposite of an alien (foreigner, stranger, foreigner). b. a member of the Indigenous population in foreign countries and overseas territories, as opposed to Europeans residing or settled there

Methods: Word Embeddings



visualisation from: https://airbyte.com/data-engineering-resources/sentence-word-embeddings

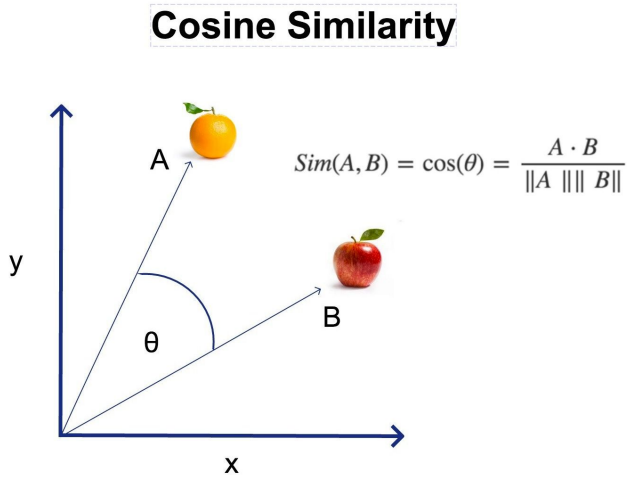
word embeddings=representations of words (as vectors) based on their meaning in contexts

Methods: Cosine Similarity

Cosine similarity is a metric used to measure how similar two vectors are, regardless of their magnitude.

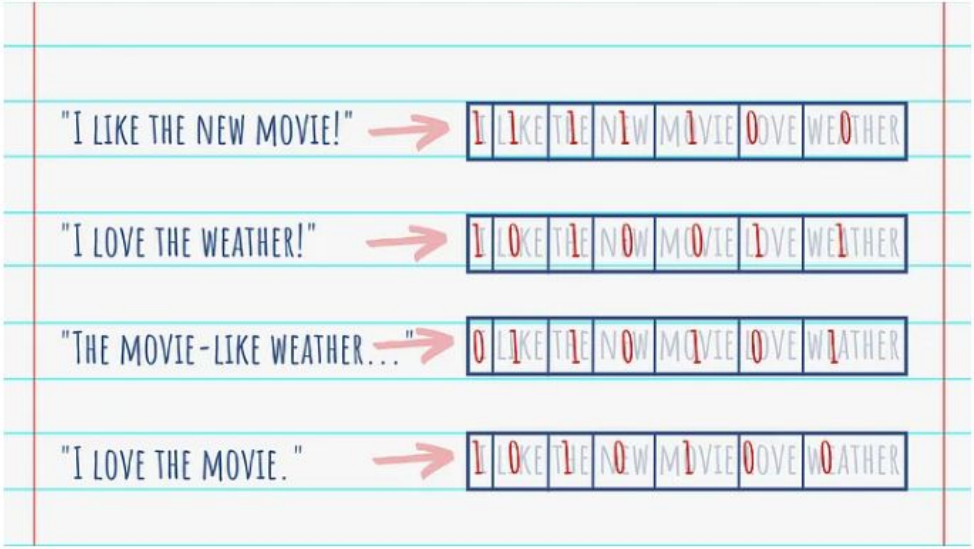
It calculates the cosine of the angle between the vectors, which gives a value between -1 and 1:

- **1** indicates the vectors are identical (i.e., pointing in the same direction).
- **0** indicates the vectors are orthogonal (i.e., no similarity).
- **-1** indicates the vectors are opposite (i.e., pointing in completely opposite directions).



from: https://businessanalytics.substack.com/p/cosine-similarity-explained

Methods: Cosine Similarity



Visualising the Bag-of-Words representation

I [1,0,0,0,0,0,0]

like [0,1,0,0,0,0,0]

the [0,0,1,0,0,0,0]

new [0,0,0,1,0,0,0]

movie [0,0,0,0,1,0,0]

love [0,0,0,0,0,1,0]

weather [0,0,0,0,0,0,1]

The sentences will then be represented as:

[1,1,1,1,1,0,0] and [1,0,1,0,0,1,1]

from https://medium.com/@manansuri/a-dummys-guide-to-word2vec-456444f3c673

Methods: Cosine Similarity

Document 1: "I love data science."
Document 2: "Data science is amazing."

Let's say the word frequency vectors are:

- **A** = [1, 1, 1, 0, 0] (for "I", "love", "data", "science", "amazing")
- **B** = [0, 0, 1, 1, 1] (for "I", "love", "data", "science", "amazing")

1. Dot Product

2. Magnitude of A

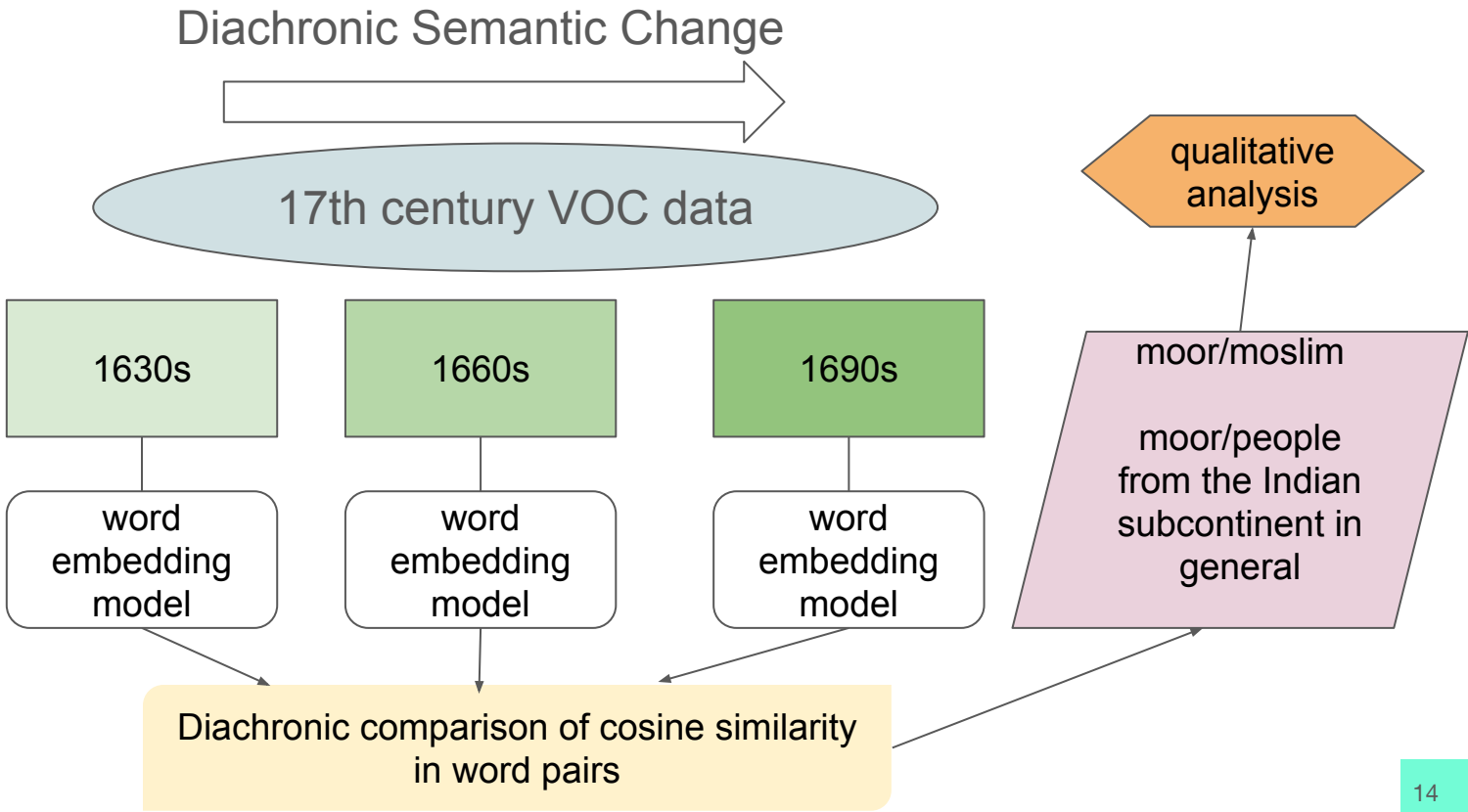
3. Magnitude of B

4. Cosine Similarity

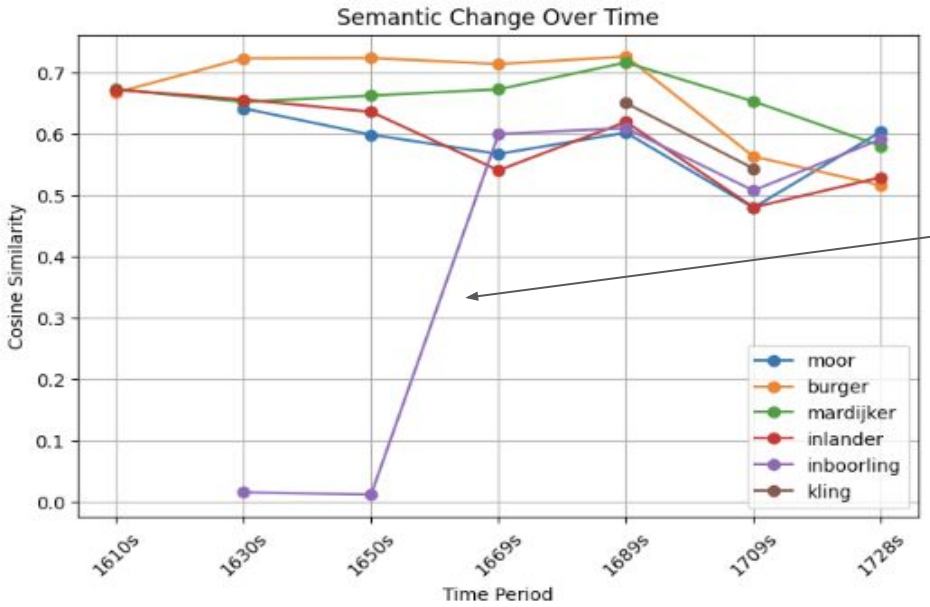
🚩 **Interpretation:**

The cosine similarity of **0.333** indicates that the documents are somewhat similar but not identical. A higher similarity would result if more words overlapped.

Methods: Diachronic Semantic Change



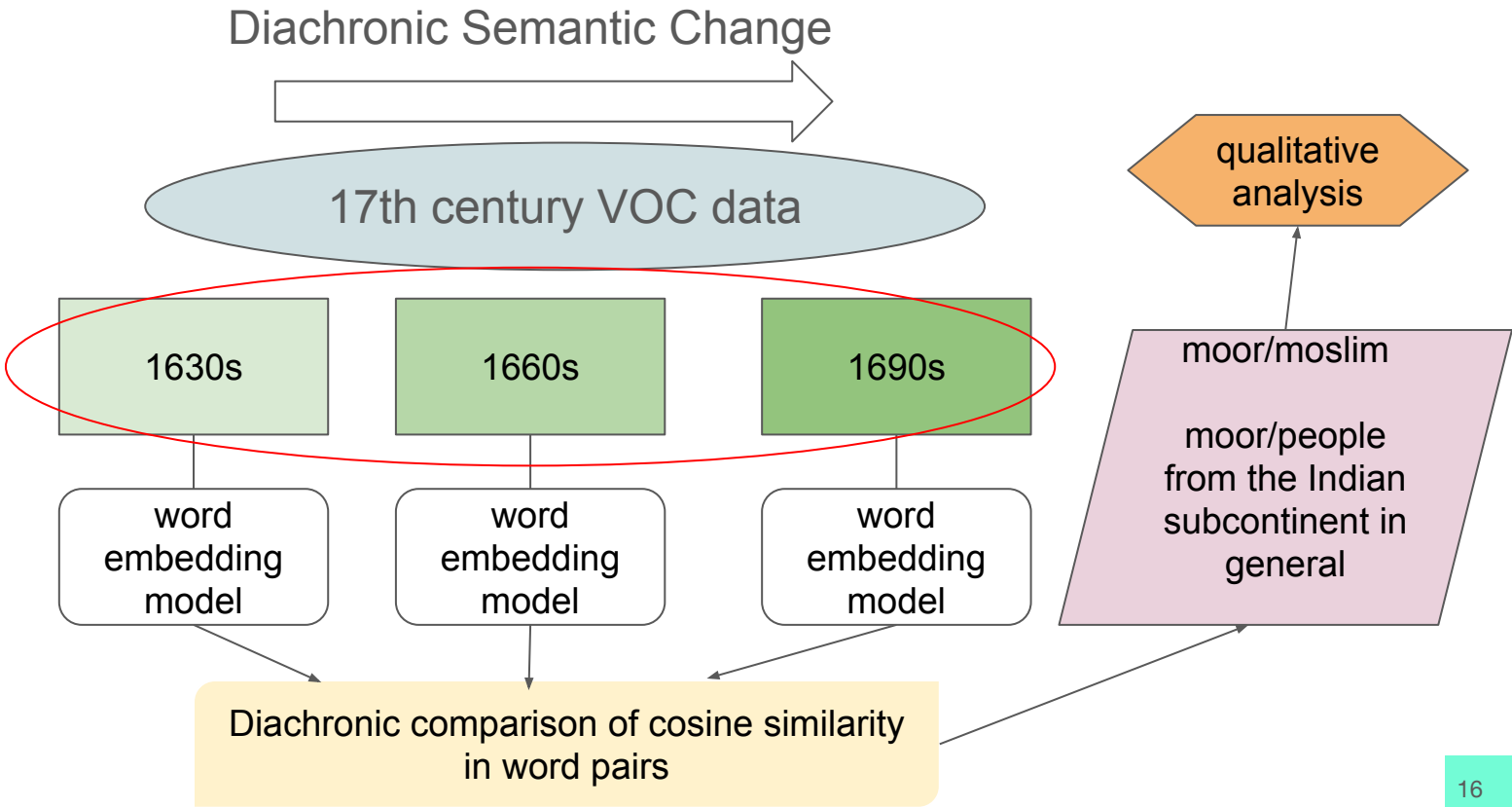
Qualitative analysis: semantic change trajectory of individual words



there might be a changepoint between 1650s and 1669s, what does it tell us about?

Limitation: segmentation of the data (to smaller sub-data)

ideally, divide the data into smaller periods of time based on **major political, societal and religious events**



Summary & Discussion

This presentation introduced:

- Lexical Semantic Change in linguistics
- The VOC, VOC archives, and ethnic, religious, and caste categories in them
- Motivation
- Methods

Discussion points:

- computational methods & historiographical practices of the VOC archives
- bridge qualitative & quantitative methods for humanities research

References

Raben, R. (2019). Ethnic disorder in VOC Asia: A plea for eccentric reading. *BMGN - Low Countries Historical Review, 134*(2), 115–128. https://doi.org/10.18352/bmgn-lchr.10684.

van Meersbergen, G. (2021). Ethnography and encounter. Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004471825.

Geeraerts, D., Speelman, D., Heylen, K., Montes, M., De Pascale, S., Franco, K., & Lang, M. (2023). Lexical Variation and Change: A Distributional Semantic Approach. Oxford University Press. https://doi.org/10.1093/oso/9780198890676.001.0001

Questions?