Unpacking the weight of spices: a preliminary exploration of long-tail contexts in the VOC trade

Gauri Bhagwat,[1,2] Teresa Paccosi,[1] and Marieke van Erp[1]
[1]KNAW Humanities Cluster, DHLab
[2]Radboud University Nijmegen

{gauri.bhagwat,teresa.paccosi,marieke.van.erp}@dh.huc.knaw.nl

December 2024

1 Introduction

Most tools developed for Digital Humanities (DH) tend to prioritize a small group of highly frequent entities, often called "head entities", while largely neglecting the less frequently mentioned ones, known as the "long tail" [8]. The predominantly quantitative approach adopted by most studies using computational methods, while valuable for "long distance reading" analyses, may overlook phenomena that hold substantial importance from a humanistic perspective. This study explores the less frequent uses of spices in historical trade-related corpora, to uncover possible insights that could inform the automatic detection of long-tail contexts for future research. Indeed, while spices were predominantly valued as trade commodities especially for culinary use, their roles as diplomatic gifts [5] and in medicine [3] underscore their broader historical significance.

2 Data

This research uses two primary sources, the *General Missives* of the Vereenigde Oostindische Compagnie (VOC)[1] (GM)[2] and the *Bookkeeper-General Batavia* (BGB).[3] GM contains the reports to the VOC board of directors sent from Batavia (now Jakarta) to the Dutch Republic between 1618 and 1793. BGB records the transportation of goods and associated ship movements, both between Europe and Asia and within Asia in the 18$^{\text{th}}$ century. We also used word

[1] `https://resources.huygens.knaw.nl/retroboeken/generalemissiven/`

[2] We specifically use the curated dataset provided in `https://github.com/trifecta-project/DHB2024/blob/main/%5B0%5D%20Parsing%20the%20GM%20letters.ipynb` to analyze the distribution of spices over time.

[3] `https://bgb.resources.huygens.knaw.nl/search`

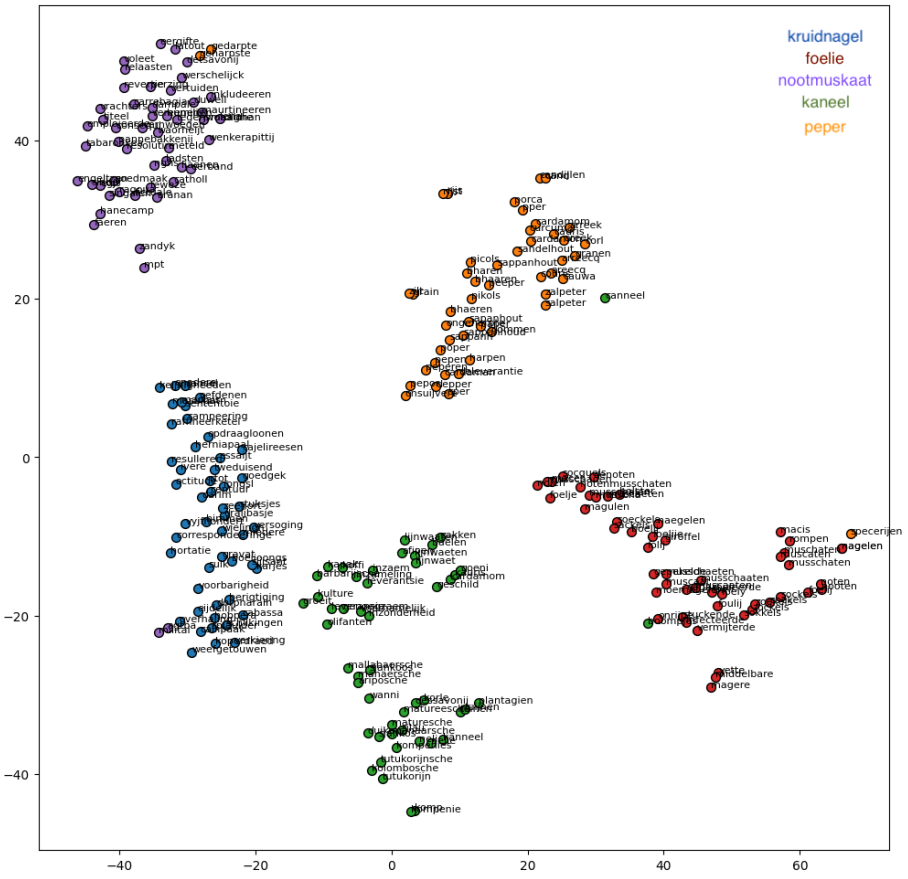

Figure 1: Word embeddings with Globalise Word2Vec model

embedding information from the collection of machine-generated transcriptions of the *Overgekomen brieven en papieren* of the VOC [4], of which the GM is a subset.

3 Methods

We started our analysis by extracting spice mentions from the GLOBALISE Thesaurus of Commodities [4], providing a standardized lexicon essential for consistent spice nomenclature. Using this as a reference, data from BGB was preprocessed to resolve data inconsistencies and to standardize formats, including variations in weight units and regional numerical formatting [2]. Incomplete or ambiguous entries, such as missing quantities or locations, were systemat-

[4] `https://lab.globalise.huygens.knaw.nl/experiments/GLOBALISE_Word2Vec_Lab/`

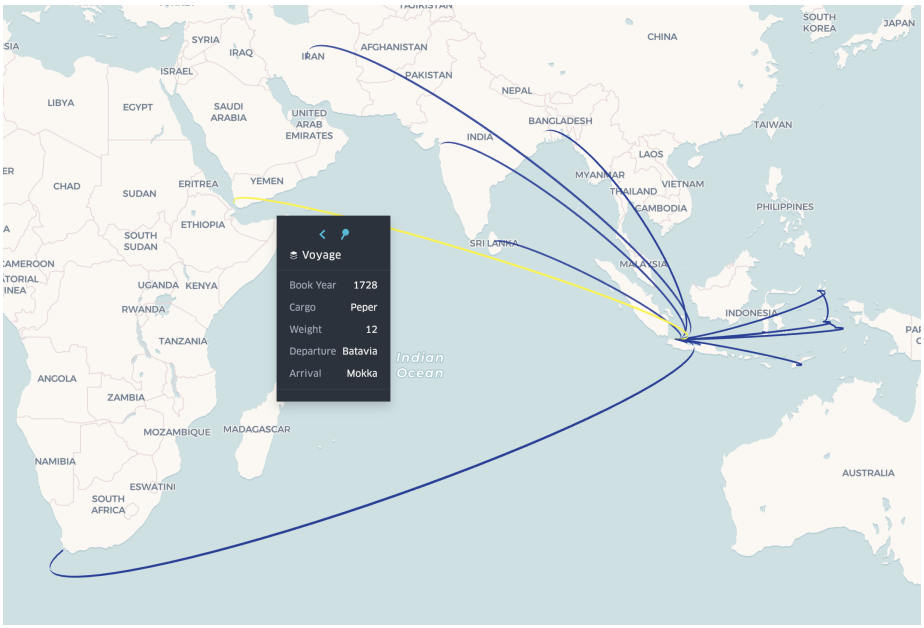

Figure 2: Visualization with Kepler.gl of the 18$^{\text{th}}$-century voyages with less than 100 pounds of pepper

ically filtered. Key data attributes such as spice names, shipment quantity, book year, and port names were normalized into a unified tabular structure for analytical processing. A quantitative and geospatial analysis (see Fig. 2) identified low-quantity shipments, initially assumed to be internal trade, even on long-distance voyages, suggesting alternative uses such as medicinal purposes. A 100-pound threshold was established to differentiate bulk trade from smaller, anomalous shipments, with outliers likely suggesting alternative uses. Temporal analysis pinpointed specific years and spices with recurring low-volume shipments, which we cross-referenced with the GM to explore possible insights in their descriptions within the selected contexts.

For the GM, we used the Word2Vec model from the GLOBALISE project to model the semantic space of selected spices and examine the distribution of target word embeddings in the corpus. First, we calculated the optimal number of clusters using the silhouette score [6], then applied t-SNE [7] for dimensionality reduction and k-means clustering to the 50 most similar words for each target word. These results were visualized as scatter plots, with clusters differentiated by color (see Fig. 1). Furthermore, we analyzed long tail contexts focusing on alternative uses of spices, particularly as medicinal items or valuable gifts, less common than their culinary use at the time. For the medical context, we created a term list starting with “medicijn” (medicine) and “ziekte” (disease), expanded using a historical Dutch lexical repository [1]. Similarly, for the gift context,

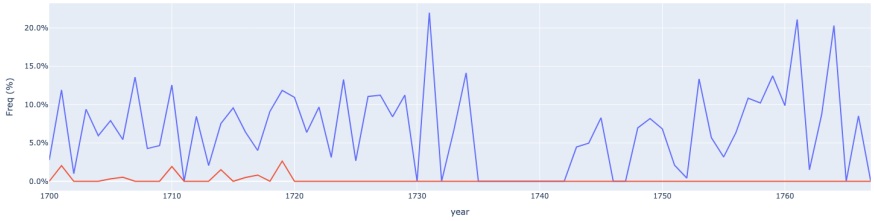

Figure 3: Distribution of 'peper' and its variants in terms of frequency in general and in gift-related context

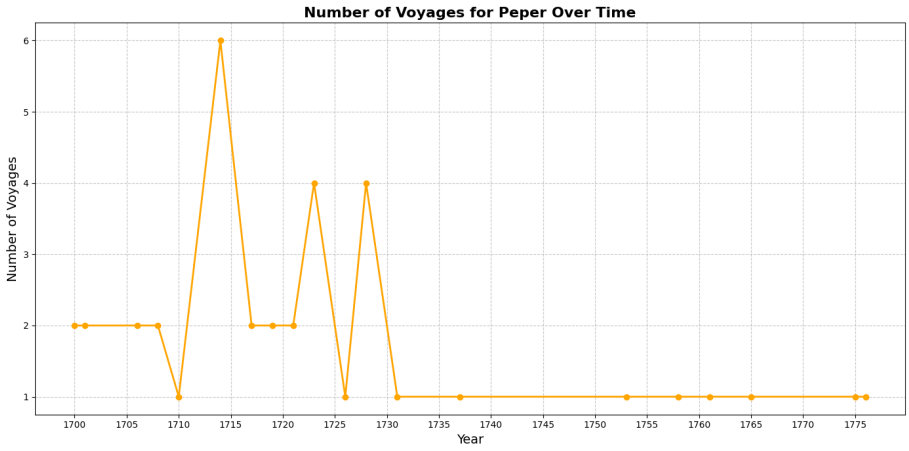

Figure 4: Voyages with less than 100 pounds of shipment of 'peper'

we began with "geschenk" (gift) and its synonyms. Using a 10-word window around target spices, we identified their use within these contexts. To enable a comparison with the possible markers, we selected texts in the GM from the 18^{th} century, the period covered by the BGB, calculating the normalized frequency of target words by dividing their occurrences by the total word count in each text. We also determined the percentage of occurrences in specific contexts and calculated context-to-total ratios. Fig. 3 illustrates the general and gift-related frequencies of the target word "peper".

4 Discussion and Conclusions

Our analysis explores the possible markers of alternative uses of spices, providing a visualization for "peper" and its spelling variants in the GM. Results indicate that "peper" is described as gift-related predominantly in the early

18^{th} century, aligning with our analysis of the BGB, where multiple voyages with low-weight shipments of "peper" were observed in that temporal span (see Fig. 4). This alignment suggests a possible correlation between lower weights and gift usage, implying that gifting could be considered a long-tail use for "peper". Such specific analyses uncover alternative uses often overlooked by methods like word embeddings, which focus on general patterns but miss less frequent, context-specific trends crucial for historical insights. To conclude, we provided a preliminary exploration of specific contexts in the VOC trade, identifying potential markers for long-tail uses. In the future, we aim to expand our focus to other uses, developing a methodology to generalize the approach to such contexts.

Our code is available at:
`https://github.com/trifecta-project/Spices-long-tail`

Author Contributions

According to CreDiT ontology. Conceptualization: GB, Data Curation: GB, TP, Formal analysis and visualization: GB, TP, Methodology: GB, TP, Supervision: ME, Writing (original draft): GB, TP, Writing (review and editing): GB, ME, TP.

Acknowledgements

Funded by the European Union under grant agreement 101088548 - TRIFECTA. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of these funding bodies.

References

[1] Katrien Depuydt and Jesse De Does. "The diachronic semantic lexicon of dutch as linked open data". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France, May. European Language Resources Association (ELRA)*. 2018.

[2] Pascal Konings. *ESTA Database Locations*. Version V5. 2023. DOI: 10622/IWJMHH. URL: https://hdl.handle.net/10622/IWJMHH.

[3] Jennifer Milam. *A cultural history of plants in the seventeenth and eighteenth centuries*. Bloomsbury Publishing, 2023. Chap. *Plants and Medicine*, pp.120–137.

[4] K. Pepping et al. *GLOBALISE Thesaurus - Commodities*. Version V1. 2023.

[5] Rinaldo Adi Pratama, Suparman Arif, and M. Syaiful. "Pepper Diplomacy: Lampung International Network in the Bargaining Position of the Banten Sultanate". In: *Proceedings of the 3rd Universitas Lampung International Conference on Social Sciences (ULICoSS 2022)*. Atlantis Press, 2023, pp. 731–744. ISBN: 978-2-38476-046-6. DOI: 10.2991/978-2-38476-046-6_71. URL: https://doi.org/10.2991/978-2-38476-046-6_71.

[6] Ketan Rajshekhar Shahapure and Charles Nicholas. "Cluster quality analysis using silhouette score". In: *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE. 2020, pp. 747–748.

[7] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[8] A. Vlachidis and D. Tudhope. "A knowledge-based approach to Information Extraction for semantic interoperability in the archaeology domain". In: *Journal of the association for information science and technology* 67.5 (2016), pp. 1138–1152.