Not something to gloss over: identifying foreign loanwords and their understood meaning in the corpus of the Dutch East India Company

Kay Pepping (ORCID), Huygens Institute for History and Culture of the Netherlands, Amsterdam, The Netherlands

In recent years, several archival institutions and research infrastructure projects have applied handwritten text recognition (HTR) techniques to early modern historical archives. This has led to an unprecedented availability of historical texts, on a scale previously unattainable.[1] Further contributing to this trend, the GLOBALISE project[2], which aims to enhance the researchability and availability of the Dutch East India Company (VOC) archives, has released transcriptions of almost 5 million scans in June 2023.[3]

The GLOBALISE transcriptions comprise the *Overgekomen Brieven en Papieren* (OBP) section of the VOC archives. The VOC, which operated from 1602 to 1798, set up long-distance trading between Europe and Asia. The OBP consists of documents sent from the Company's Asian headquarters in Batavia to the Dutch Republic. They detail the company's activities, ranging from trade to diplomacy and violent conquest which resulted in colonial power relations between the company and the various local powers they exploited for profit.

The availability of historical documents as machine-readable text is of tremendous value to historians, who can use tools such as the GLOBALISE Transcriptions Viewer to search gigantic corpora of text through simple keyword searches.[4] The machine-readable text also creates room for new approaches to analysis, similar to how automated research into authorial attribution or collocations is now established practice on OCR-corpora (Hill & Hengchen, 2019).

In the case of Asian-European interaction, HTR provides access to material much older than OCR does. This is not just of interest to historians: digital methods applied to large historical datasets also offer possibilities for sociolinguistic research on a new scale (Liimatta et al., 2023). In Asia, the VOC employees encountered a world of new flora, fauna, cultures and languages. To describe these in their correspondence, they often had to rely on words borrowed from local languages.

This forms the first stage of adopting loanwords: a specific group of people loaning words for 'exotics', concepts not native to their world of experience (Van der Sijs, 1996). Thanks to these loans, the primarily Dutch language archive is peppered with words and phrases drawn from Asian languages. Existing automated methods to identify loanwords have two requirements: identifying what borrowing and loaning

[1] See e.g. https://amsterdam-city-archives.transkribus.eu/, https://zoekintranscripties.nl/, https://republic.huygens.knaw.nl/.
[2] https://globalise.huygens.knaw.nl/
[3] https://hdl.handle.net/10622/JCTCJ2
[4] https://transcriptions.globalise.huygens.knaw.nl/

language, and having extensive reference data (e.g. dictionaries) for both (Zhang et al., 2021).

Neither can be assumed for the archival VOC-corpus. When large amounts of scans are transcribed automatically, we cannot know beforehand what languages will be included (Liu & Smith, 2020). Although reference data exists for some of the languages the VOC interacted with (such as Japanese), other languages did not survive their encounter with the VOC. An example is *kelang*, once spoken on the Indonesian island of the same name, whose speakers were violently displaced and killed by the VOC (Collins, 2003). This tragedy makes it more difficult to find their language in the VOC archives due to lack of reference data but also more important, as it is one of the last places where it can potentially be found.

A different manner to extract potential appearances of Asian languages in the archives is needed. This abstract proposes a simple method to start creating data on loaned words appearing in large text corpora: recognising glossing. A gloss is a brief notation of the meaning of a word. This idea has been explored on a small scale with manually transcribed material from the archives of the English East India Company (Kaislaniemi, 2017). The release of the GLOBALISE transcriptions now allows doing the same on a much grander, semi-automated scale. Within the VOC archive, a gloss is often indicated by the use of the word *of* ('or'), as seen in this example:

> here in the village, this was researched by the whole Adaleth, or Moorish court of justice[5]

Utilizing a Word2Vec representation of the corpus, which helps find words that are used in a similar manner (Ayyadevara, 2018), to find words used in the same way as *of* enables easy expansion of a list of gloss indicators with words like *offe* and *ofwel.* This list of gloss indicators can then be applied to the tokenized OBP corpus to extract two pieces of information. First, the token preceding the indicator (the 'term' that is glossed) and the tokens following it (the 'explanation'). Second, a count of the total occurrences of the term in the corpus, which enables a rough percentage of which percentage of a word's occurrences is glossed.

In my paper, I will present an exploratory dataset of loanword terms and their explanation. This dataset contains noise: not every appearance of an indicator is an actual gloss, and not every term and explanation consists of the same amount of tokens. Despite this noise, the data provides a start for research on the intersection of history and sociolinguistics. Using various examples from the data, I will show how the extracted information can be of use:

The **terms** indicate what words were considered 'foreign': this information could be used to improve accurate language tagging in reference thesauri such as the ones GLOBALISE is making (Nijman & Pepping, 2023). For builders of lexica, it can provide

[5] Original: ''hier in ons dorp door den gantschen Adaleth of het Moorse hof van Iustitie onderzogt'. Nationaal Archief, Verenigde Oost-Indische Compagnie (VOC) 1.04.02, inv. nr. 2920 fo. 1334 verso.

early examples of the use of a word. Moreover, it also has the potential to help highlight words in languages that were wiped out or heavily endangered by the VOC.

The **explanations** show the way the author understood a term, and the way this understanding might change over time. These various interpretations could, with some curation to handle the noise, be used to generate a bottom-up version of the VOC-glossary (Kooijmans & Schooneveld-Oosterling, 2000). The 196 year span of the OBP-based dataset means the varying explanations represent the shifting understanding of a term, aiding the much-needed 'unflattening' of the definitions of terms (Van Erp, 2023).

Finally, the amount of times a word is glossed in comparison to the number of times it occurs gives insight into the adoption of loanwords when viewed over time. If a word is glossed in every documental unit (documents meant to be read together) it appears in, it is considered foreign to the recipient. The percentage of glosses going down over time could indicate increasing adoption of the word.

This publication is part of the project GLOBALISE (with project number 175.2019.003) of the research programme Research Infrastructure which is (partly) financed by the Dutch Research Council (NWO).

Bibliography

Ayyadevara, V. K. (2018). Word2vec. In V. K. Ayyadevara, *Pro Machine Learning Algorithms* (pp. 167–178). Apress. https://doi.org/10.1007/978-1-4842-3564-5_8

Collins, J. T. (2003). Language death in Maluku: The impact of the VOC. *Bijdragen Tot de Taal-, Land- En Volkenkunde*, *159*(2/3), 247–289. JSTOR.

Hill, M. J., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, *34*(4), 825–843. https://doi.org/10.1093/llc/fqz024

Kaislaniemi, S. (2017). 7. The early English East India Company as a community of practice: Evidence of multilingualism. In E.-M. Wagner, B. Beinhoff, & B. Outhwaite (Eds.), *Merchants of Innovation* (pp. 132–157). De Gruyter. https://doi.org/10.1515/9781501503542-007

Kooijmans, M., & Schooneveld-Oosterling, J. (2000). *VOC-glossarium, Verklaring van Termen, Verzameld uit de Rijks Geschiedkundige Publicatien die Betrekking hebben op de Verenigde Oost Indische Compagnie*. Instituut voor Nederlandse Geschiedenis.

Liimatta, A., Ryan, Y., Säily, T., & Tolonen, M. (2023). Results from rough data? The large-scale study of early modern historiography with multi-dimensional register analysis. *Digital Humanities in the Nordic and Baltic Countries Publications*, *5*(1), 297–312. https://doi.org/10.5617/dhnbpub.10668

Liu, S., & Smith, D. (2020). Detecting de minimis Code-Switching in Historical German Books. *Proceedings of the 28th International Conference on Computational Linguistics*, 1808–1814. https://doi.org/10.18653/v1/2020.coling-main.163

Nijman, B., & Pepping, K. (2023). *Building a VOCabulary: The uses and challenges of thesauri for working with early modern recognized entities*. https://doi.org/10.5281/ZENODO.7973694

Sijs, N. van der. (1996). *Leenwoordenboek*.

Van Erp, M. (2023). Unflattening Knowledge Graphs. *Proceedings of the 12th Knowledge Capture Conference 2023*, 223–224. https://doi.org/10.1145/3587259.3630082

Zhang, L., Fabri, R., Nerbonne, J., & Nerbonne, J. (2021). Detecting loan words computationally. In E. O. Aboh & C. B. Vigouroux (Eds.), *Contact Language Library* (Vol. 59, pp. 269–288). John Benjamins Publishing Company. https://doi.org/10.1075/coll.59.11zha