Linguistica Silesiana nr 45/1, 2024
ISSN 0208-4228
DOI: 10.24425/linsi.2024.150390

REMO APPOLLONI
Faculty of Arts and Humanities
Sapienza University of Rome, Italy
ORCID: 0000-0003-1826-8016
remo.appolloni@uniroma1.it

OPPORTUNITIES AND THREATS OF HISTORICAL EVIDENCE AND DATA VIA CORPUS-BASED SOFTWARE: A CASE STUDY ON GERARD MALYNES

The aim of this paper is to exploit the informative nature of datasets that can be created from corpus-based software to explore specific phenomena in early modern specialized discourse, and to corroborate the adoption of the same software for historical analysis. Particular relevance will be devoted to the special nature of historical evidence, which has caused critical issues in the reliability of the data collected for the purpose of historical investigation of English. Spelling variation, in this sense, is one of the most crucial problems of Early Modern English, and this has often affected the reliability of data to be collected via software, especially when statistical findings are involved. The normalisation of historical texts has contributed enormously to make texts better readable for historical corpus analysis; and, consequently, to improve the accuracy and manipulation of data. Moreover, several tools used in corpus linguistics have benefited from the normalisation of spelling variants in the same terms, e.g. part-of-speech taggers for historical variety. This case study will attempt to explore the data retrievable from corpus-based software like VARD, #LancsBox and CQPweb, and to use them to corroborate a preliminary analysis of early modern economics discourse in two treatises written by Gerard Malynes in 1601 and 1623.

Keywords: historical evidence, corpus-based software, data analysis, Early Modern English, specialized discourse, mercantilism

1. Introduction

This paper aims to propose a preliminary exploration of the data which can be collected via the most frequently used tools available in corpus linguistics applied to the history of English, i.e. VARD[1] (Baron and Rayson 2008), #LancsBox[2] (Brezina, McEnery and Wattam 2015) and CQPweb[3] (Hardie 2012). This is to observe the transformation of the data retrieved via corpus-based software into prospective evidence of language at historical level in specialized discourse or into useful additional metadata to corroborate historical corpus-based analysis via software. In particular, the question to be addressed here is which information is possible to be disclosed from the collection, quantification and aggregation of these data, regardless of them being numerical or categorical. This case study[4] will be conducted on two texts from a larger sample corpus created for the purpose of a PhD research project being conducted at Sapienza University jointly with Silesia University[5]. The combination of these two texts will be of mere exemplificative purpose, to have a couple of small *snapshot corpora* (McEnery and Hardie, 2012: 9) to be sampled as a dataset for the analysis. The ultimate goal is to evaluate which are the opportunities of utilising corpus-based data in the historical analysis of English economics discourse in the early 1600's; and to identify potential threats from historical evidence which may undermine the quality of data analysis.

2. Theoretical and methodological premises

2.1. Early Modern English: some known issues

Early Modern English has been extensively studied and investigated, especially due to the large number of written texts available for research purposes (Baron, Rayson and Archer 2009b: 41). The establishment of the printing press in 1476 increased the amount of textual evidence available and occasioned a widespread diffusion of a standard variety across England (Baugh and Cable 2002; Beal 2016). However, the production of a wide range of texts witnesses the presence of the distinctive creativity and experimentation that the English

[1] VARD (v. 2.5.4) is available online at https://ucrel.lancs.ac.uk/vard/about/.

[2] #LancsBox (v. 6) is available online at http://corpora.lancs.ac.uk/lancsbox/download.php.

[3] CQPweb (v3.3.17) is available online at https://cqpweb.lancs.ac.uk/.

[4] The results presented here are also part of a larger preliminary analysis which has been presented in two conferences (Appolloni 2022a; 2022b), in which the entire period of the project was covered, i.e. 1572-1664.

[5] The project focuses on the creation of a specialized discourse in the EModE period, that is the language of economics as a new empirical science in the period 1572-1664.

language underwent during the same period. This textual evidence for both diachronic or synchronic analyses is frequently contradictory, and researchers experience considerable difficulty in their interpretation (Nevalainen 2006: 12). In fact, Early Modern English (EModE henceforth) went through a high level of orthographic variability and reached a general uniformity virtually only around 1650 (Nevalainen 2006; Culpeper and Archer 2018). It is conventionally agreed that English was established as a commonly accepted uniform variety only by the end of the 17th century (Beal 2016: 313). In this sense, no generally accepted system for the English spelling existed during the Renaissance, and this encouraged idiosyncrasies between writers, despite the level of consistency they had in their own works (Baugh and Cable 2002: 193-194). Spelling has hence proved to be one of the major issues when dealing with historical research and its evidence, especially for the typical focus on the presence of a uniform variety among historical texts (Beal 2016: 303). Likewise, the massive presence of variation has not facilitated the analysis of historical sources such as fragmentary texts or language varieties which are inconsistent in their orthography; this is particularly true when corpus-based software packages are utilized. As a logical consequence, in order to carry out historical investigation on this macro-historical variety, researchers require little or great intervention on texts before examining the evidence available, so that reliable results (or those that are approximately reliable) may be collected and interpreted, and robust theses may be formulated.

2.2. Corpus linguistics in historical investigation: the pros and cons

Many disciplines have benefited from the advances of corpus linguistics techniques, software and methods, with language teaching, discourse analysis or translation as some of the examples. In general, corpus-based sources have amplified the range of analysis from a mere micro-level, i.e. close-reading, to a macro-level of investigation, where empirical evidence can be interpreted quantitatively (Taavitsainen 2016: 272). Historical linguistics has been argued not to fully exploit these opportunities and strengths offered by the technological breakthrough of corpus-based analysis and statistics (Jenset and McGillivray 2017: 17). Electronic corpora have in this sense offered large quantities of data for the historical analysis of language, and improved the level of efficiency in the creation and examination of historical datasets (López-Couso 2016: 129). For example, Nevalainen and Raumolin-Brunberg (2003) showed how corpus linguistics has facilitated enormously the sociohistorical investigation of English during the Tudors era; and how the quantitative approach could shed the light on the heterogeneity of the diachronic changes occurred (as cited in Beal 2016: 310). Beyond these preliminary assumptions, corpus data are not considered here only in terms of evidence to describe linguistic phenomena occurring at a historical level. The framework that will be followed here is the one to exploit the quantitative approach in its application to those tools available for the historical analysis of language. In this respect, corpus data will be assumed as evidence to explore the full potential of the tools utilized in historical corpus linguistics. And hence they will also be adopted to progress the field of historical corpus-based data in its overall perspective (Jenset and McGillivray 2017: 1).

For the purpose of this paper, corpus linguistics is hence employed as a principled method and as a set of tools to collect data to be analysed (López-Couso 2016: 126). This is to try to answer some of the theoretical questions in historical linguistics as a discipline, e.g. the appropriateness of a quantitative approach to historical linguistic data. As a method, corpus linguistics is entirely suitable to study the diachronic change of language, especially to track the evolution of certain aspects of language in terms of their frequency (McEnery and Hardie 2012: 94-95). In particular, it is possible to observe general trends of diachronic change through statistics, and also to establish the relevance of selected linguistic features in the examined variety, e.g. historical or diachronic. Additionally, the adoption of corpus linguistic methods has recently resulted in increasing opportunities to investigate new research questions and, more importantly for the purpose of this paper, to collect data to test them (Davidse and De Smet 2020: 211). In the case of historical and diachronic analyses, old research questions could be reviewed, and new insights have been provided in the field (Beal 2016; López-Couso 2016). The opportunity to exploit data to be collected and interpreted via software[6] is even more important when considered as an attempt to pursue objectivity in the results obtained. Corpus linguistics is thus increasingly deemed to provide useful data related to linguistic phenomena as, for example, diachrony and variation (Davies 2015: 12). In fact, the most relevant aspect for this study is that a corpus is considered here as a dataset in its entirety, that can be searched and investigated via a computer (Brezina and Gablasova 2018: 595), and hence quantified to extract useful information. For example, the usage of annotation as a process – regardless of the linguistic category to study – guarantees additional information and linguistic analysis encoded in the textual data which are readable via corpus-based software (McEnery and Hardie 2012; Newman and Cox 2020). But although widely adopted as a method in the study of historical and diachronic corpora, corpus linguistics reveals some issues related to the impact of spelling variation in historical texts (Rayson and Potts 2020: 128). Diachronic corpora, especially those employed for historical linguistics, have been frequently challenged by the presence of massive orthographic variability, as for example in the case

[6] A corpus is by definition a collection of "machine readable" texts which can be processed and then studied via software (McEnery and Hardie 2012: 1-2).

of EModE. Spelling can cause a significant reduction in the level of accuracy when automatic annotators[7] are adopted or when the dispersion of frequency-related data across different variants of the same lexeme is examined (Rayson 2015: 44).

The historical resources researchers have at their disposal to build historical corpora are furthermore generally incomplete or, more accurately, scarcely granular in statistical terms. This may result in the phases of design and compilation of such corpora to be demanding (Jenset and McGillivray 2017: 8). Historical corpus linguistics can obviously take advantage of the wealth of written texts surviving from the EMod period, which are the primary source for research purposes (Nevalainen 2006; Jenset and McGillivray 2017). But although this evidence can offer a privileged view on the massive variability in that period, it is frequently conflicting and fragmentary, and obviously requires careful evaluation or intervention of researchers (Nevalainen 2006; Kytö and Smitterberg 2015; Jenset and McGillivray 2017). This, as already noted, is mostly due to the considerable level of variation in spelling. In order to deal with this issue, scholars may opt for edited versions of the texts with modernized spelling. A case in point is the study conducted by Culpeper (2002) on *Romeo and Juliet* in which he adopted a modernised version of the text to handle the possible hindrances of spelling variation in applying statistical measures (as cited in Baron, Rayson and Archer 2009b: 47). This strategy obviously reduces the time required to analyse fragmented texts. However, editorial choices must be taken into consideration, for they may possibly affect the linguistic occurrences under investigation (Kytö and Smitterberg 2015: 337). On the other hand, the digitalization of manuscripts – thanks to the Optical Character Recognition (OCR) technique, keying and the hard work of librarians, e.g. the Oxford Text Archive project (OTA)[8] – has provided scholars and researchers with more texts available, especially via digital archives which make them directly accessible for research. However, they may necessitate some little or great intervention to make them employable for scholarly purposes (Horobin 2016: 114).

2.3. A case study on historical evidence: objectives

On the basis of these premises, i.e. the EModE spelling variability, the data repository from corpus-based methods and the special nature of historical data, there are different types of data to be collected here to formulate some preliminary assumptions about the texts examined in quantitative terms. Firstly,

[7] For example, in the case of automatic lemmatization, the level of accuracy is generally attested around 90% (Newman and Cox 2020: 35-36), but this may decrease due to massive spelling variation if not normalised.

[8] The Oxford Text Archive is available online at https://ota.bodleian.ox.ac.uk/repository/xmlui/.

the data in the form of XML tags will be produced via VARD as a result of the normalisation process. This can be automatic, on the basis of an established threshold of quality confidence, or manual, to normalise the remaining variants from the automatic process or, as in this study, to annotate words of a foreign language to quantify its presence in the texts. A wordlist that is automatically generated by VARD will be also employed as a repository of the variants occurring in the corpus to formulate proper query-based search in modern corpus-based software. Secondly, the data produced via annotation and tagging will be processed automatically with #LancsBox to encode part-of-speech information in the texts (i.e. POS-tags), which will be assumed here as a general criterion to evaluate whether the texts are suitable for the research purpose designed. This subsumes one of the two research questions investigated in the PhD project mentioned here, that is the one related to the formation of taxonomies in EModE economics discourse. Thirdly, descriptive statistical data will be collected, i.e. both absolute and relative frequencies of selected terms, for the comparison with a normative corpus. They will be extracted to evaluate the level of deviation in the accuracy of the wordlist and of POS-tags lists generated by corpus-based software between normalised and original versions of the texts. Moreover, the log-likelihood value (LL) (Dunning 1993) and the log ratio score (Hardie 2014) will be calculated to test keyness between the sample and the target corpus (Brezina 2018; McIntyre and Walker 2019). Lastly, a few key metadata of the snapshot corpora will be used as main criteria to extract a reference sub-corpus from the Early English Book Online (EEBO v.3) available via a web-based corpus tool; this will comply with the principled approach to compare homogenous and almost equivalent corpora.

3. Methods and materials

3.1. Texts repository

The texts used here to compile the overall snapshot corpus are downloaded from the Oxford Text Archive (OTA), a project hosted at Bodleian Libraries, University of Oxford. This is a digital repository of texts which provides scholars and researchers with literary and linguistic data, that can be downloaded in various formats, e.g. XML or HTML, and used, as in this case, for corpus linguistics purposes. Most of the texts are copies of those available in the above-mentioned partnership project EEBO-TCP, which aims to encode editions of texts in English in the period 1475-1700[9]. The adoption of digitalised editions of

[9] More details about EEBO and OTA at https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/5.

texts is progressively common, although some text adjustments are often required in terms of punctuation, abbreviations or parts of the text lacking (Horobin 2016: 112).

3.2. Textual data

Two texts are used here to create a small snapshot corpus for the purpose of exemplifying the main objectives of the present case study. Both the texts are works written in the first half of the 17th century by Gerard de Malynes (1586-1627), a commissioner of trade who firmly advanced some crucial positions against mercantilism regarding the exchange control, with particular reference to the necessity to prohibit speculation in the foreign exchanges (Robbins 2000: 49-50). The reason why Malynes is included in the list of texts selected for the creation of the main project's sample corpus relies on his leading presence in the mercantilist debate between the end of the 16th century and the first half of the 17th century. The EMod intellectual debate on economics[10] focused on the "regulation of commerce and money" (Robbins 2000: 48), and was centred also on two others key pamphleteers of the early economics discourse, i.e. Edward Misselden and Thomas Mun. They were both critics of Malynes' positions (Robbins 2000: 50), which were essentially bullionist, and hence in contrast with their early mercantilist positions. Mercantilism, in fact, was based on the idea that the exchange value is positive only when a surplus balance of trade generates a higher flow of precious metals.

The two texts selected to summarise Malynes' positions are as follows:

1. G. Malynes, (1601). *A treatise of the canker of Englands common wealth Deuided into three parts: wherein the author imitating the rule of good phisitions, first, declareth the disease. Secondarily, sheweth the efficient cause thereof. Lastly, a remedy for the same. By Gerrard De Malynes merchant.*
2. G. Malynes, (1623). *The center of The circle of commerce. Or, A refutation of a treatise, intituled The circle of commerce, or The ballance of trade, lately published by E.M. By Gerard Malynes merchant.*

The combination of the two treatises has created a small snapshot corpus of approximately 60,000 tokens, which partly represents the diachronic evolution of the language of Malynes in the first twenty years of the 17th century.

On the basis of the metadata collected during the phase of compilation of this sample, the CQPweb function of "restricted query" (Hardie 2012: 389) has been

[10] The term *economics* is used here conventionally, for it was still not a proper science or discipline at the time, but a mere intellectual debate.

used to create an equivalent reference sub-corpus to be extracted from EEBO v. 3[11]. In details, the query was restricted in terms of *rough-draft genre*, i.e. selecting *treatise* only, and of *years of publication*, i.e. selecting *1601* and *1623* only. The final result is the production of two corpora, one sample and one normative, which are generally equivalent in their structure and in the nature of the language variety presented, that is of specialized discourse in the EMod period.

Table 1. Sampling frame of the corpora used in the case study.

REFERENCE	**YEAR(s)**	**AUTHOR(s)**	**GENRE(s)**	**TEXT(s)**	**TOKENS**	**TYPES**	**TTR**
sample corpus							
A06791	1601	G. Malynes	treatise	*A Treatise of the Canker of England's Common wealth*	20.430	2.456	0,120
A06785	1623	G. Malynes	treatise	*The Center of the Circle of Commerce*	37.625	4.770	0,127
TOT,					**58.055**		
reference corpus							
EEBO v3 (restricted)	1601	aavv.	treatise	Number of texts: 6	171.844	674	0,004
	1623	aavv.	treatise	Number of texts: 7	367.400	1.440	0,004
TOT.					**539.244**		
source corpus							
EEBO v3	1475-1700	aavv.	vv.	Number of texts 44,422	1.202.214.511	4.713.326	0,004

The EEBO reference, e.g. A06791, is highlighted here for pragmatic reasons, i.e. to facilitate the retrieval of the text via EEBO, via OTA and via CQPweb. For the same reason, both texts were categorised as *treatise*, following the genre label available in their metadata on CQPweb. Lastly, the type-token ratio (TTR) is signalled here for the mere purpose of indicating the general homogeneity of lexical variation across the sub-sections of the corpora considered.

3.3. Software packages

VARD is an interactive software package for the standardization of spelling variants in historical texts, producing an XML version of the normalised text (Baron, Rayson and Archer 2009a; Baron and Rayson 2009). Beyond its main task, VARD is commonly employed in corpus linguistics dealing with historical texts[12] to improve the accuracy of automatic annotation, e.g. the POS-tagging (Rayson 2015; Smitterberg 2016), and to improve the accuracy of the texts for the sake of statistical manipulation and findings (Baron and Rayson 2009; Baron, Rayson and Archer 2009b). In this case study, it will be utilized to exploit the quantifiable data resulting from both automatic and manual annotations to observe a general trend of spelling variability in the two texts; to quantify the presence of Latin in the same texts; to evaluate the level of deviation between normalised and original texts in terms of POS-tags and frequency; and to take

[11] This version is available among the several corpora uploaded on CQPweb.

[12] VARD is also employed in the normalisation of diatopic varieties of English having different spelling norms (e.g. British English versus American English).

advantage of the Known Variant List (KVL henceforth) to have a repository of all the variants occurring in the original texts, which will be useful to formulate proper queries for lexemes to be searched via CQPweb[13].

LancsBox is a software package in the form of a desktop tool that is commonly used for corpus linguistics analysis (Brezina, McEnery and Wattam 2015; Brezina and Gablasova 2018). Among its main functionalities and methods, it is adopted for the extraction and analysis of frequency lists; for the analysis of keywords in the related contexts via concordances; for the extraction of collocations and the evaluation of its association measures, and for the evaluation of the frequency of prospective lexical bundles. For the purpose of this case study, #LancsBox will be exploited for the retrieval of data from automatic annotations, i.e. tokenization and POS-tagging, and for the statistical measure of keywords frequency.

CQPweb is a web-based tool adopted for corpus linguistics analysis (Hardie 2012; Brezina and Gablasova 2018). This tool is mainly based on the Corpus Query Processor (CQP) server, which allows it to search large corpora efficiently with the adoption of sophisticated queries (Hardie 2012; Rayson 2015). In this analysis, CQPweb will be exploited in the restricted query function to create an approximate equivalent reference sub-corpus, and for the comparison with the sample in terms of trends of general linguistic phenomena and of keyword analysis representation.

Lastly, Excel[14] will be used here to create a dataset of numerical and categorical data from the textual, linguistic and statistical data collected via the above-mentioned software. In this sense, Excel will be employed for the aggregation, interpretation and visualization of data.

3.4. Data extraction

In this section, all the phases involved in the extraction of the relevant data will be described. They mainly refer to three main macro-phases: preparation of the sources, data generation and extraction, data comparison.

In the first phase, the selected texts were downloaded from the OTA website in HTML format, and hence transformed into .txt files – that is a conventionally used format for the purpose of corpus-based software reading. All the non-textual elements generated were promptly removed from the text, e.g. any formatting characters. Additionally, all those footnotes with mere bibliographic references were cancelled to make the proper textual data available for software-based reading and analysis.

[13] The EEBO v3 uploaded on CQPweb is partially not normalised.

[14] Excel v. 2210 (build 157226.20174).

Secondly, both source texts were singularly uploaded to VARD to generate a first quantification of the spelling *variants* existing in the files, which had to be normalised, and of those labelled as *not variants*. The KVL was set according to the frequency of the spotted variants. On the basis of the italics signalled in the HTML version of the texts on OTA, all the Latin words in both the texts were manually annotated via VARD, in order for them not to be normalised, and to quantify their presence against the total of variants and non-variants identified by the software. The first most frequent variants were qualitatively scrutinized together with their proposed pairs for normalisation together with the percentage of confidence. This was to establish a minimum threshold of confidence to carry out a reliable normalisation of the texts automatically: the value was set at 81%, based on the correctness of the pairs suggested by the tool. Once normalisation was completed, the KVL was downloaded into an Excel spreadsheet to collect useful data regarding all the token variants paired with their normalised type, and to quantify the size of spelling variability in the texts in terms of occurrences. The list of remaining variants was pasted in a .txt file to be automatically annotated via #LancsBox: this was to exploit POS-tags also for the identification of the language categories of the remaining variants, e.g. proper nouns which are very commonly identified as *real-word errors* (Baron, Rayson and Archer 2009b: 53.). The normalised texts were imported into #LancsBox for the automatic annotation, i.e. the tokenization[15] of the texts and POS-tagging. Once annotated, an automatic wordlist was generated for the sample in the section *Words* of #LancsBox. The setting related to the form of the unit displayed in the list was changed from *type* to *lemma* to include all the possible inflectional endings of the queried lexemes; and to filter the list in search of nouns via a specific wildcard, i.e. *_n[16]. After a qualitative scrutiny of the most frequent nouns related to the economics discourse, 15 terms were selected. They have been extracted from the snapshot corpus on the basis of statistical (absolute frequency), syntactic (lemmatized nouns) and domain-related criteria (economics). The same terms were then searched in the KVL downloaded from VARD, in order to retrieve all the spelling variants existing in the sample for each term. Consequently, a specific syntax for querying CQPweb[17] via wildcards was written for each term, in order to prospectively retrieve all the variants of the terms in the concordances produced by the tool. This was expected to deal with the spelling variation of the original texts in the EEBO v3 available on CQPweb.

[15] Tokenization is a prerequisite for corpus annotation, for linguistic items must be divided into units which are suitable for the analysis and annotation via software (Newman and Cox 2020: 26).

[16] The main project from which this case study is extracted mainly focuses on the formation of taxonomies in EModE economics discourse; hence, the extraction of terms was focused on nouns.

[17] See section 2.2 for the extraction of a reference sub-corpus from EEBO with the restricted query.

Finally, each term was searched through the corpus, and frequencies were collected.

Lastly, sample and reference corpora were compared in terms of frequency to produce additional data in the form of log-likelihood values and log-ratio scores, to establish whether prospective differences merely resulted from chance or whether they had a statistical significance (Brezina 2018: 83). A further sample was created and uploaded via #LancsBox with the original versions of the texts, i.e. non-normalised, to compare the deviation of frequency and of the POS-tags examined as a result of the adoption of VARD as a software package.

4. Results discussion

4.1. Data extracted from VARD

The following stacked bar charts represent the quantification of the automatic annotations produced by VARD. In particular, fig. 1 shows the amount of spelling variants when uploading the two texts by Malynes in their raw form, i. e. .txt, and in their original language variety – hence with the typical EModE spelling variability. In general, the two texts reveal no major difference in terms of presence of spelling variability diachronically, although a slight increase is observable in the text published in 1623. In particular, both the texts present almost 18% of spelling variants as tagged by VARD before normalisation (i.e. the text A0791 17,71%, and the text A06785 17,98%, i.e. in 1601 and 1623 respectively). This clearly shows a reduction of massive spelling variation in EMod economics discourse in the early 17th century, especially if compared to other texts published in the second half of the 16th century, where the percentage was approximately 30% (Appolloni, 2022a). Furthermore, fig. 2 displays the effect of automatic normalisation via VARD (confidence threshold set at 81%): both the texts have the amount of spelling variants being reduced of almost 15% in the sample. As is observable from the bars, the number of remaining variants (i.e. between 700 and 1300 respectively) allows for a manual normalisation without excessive time-consuming effort; this hopefully confirms the feasibility of VARD's automatic normalisation also in EMod economics discourse.

Part of the remaining variants in the KVL are possible real-word errors, e.g. proper nouns. Therefore, text A06791 has been used again as an example to formulate prospective assumptions about the presence of real-word errors in the variants still to be normalised[18]. The following doughnut chart in fig. 3 illustrates the percentage of word categories under which the remaining variants can be

[18] The version of the text adopted at this point is the one which has been also normalised manually after the automatic annotation via VARD, and it is a part of the PhD project.

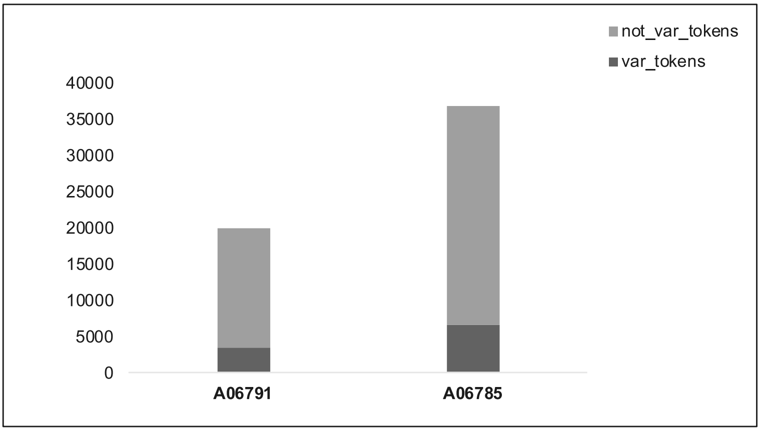

Fig. 1 Quantification of spelling variants in Malynes’ selected texts (original version).

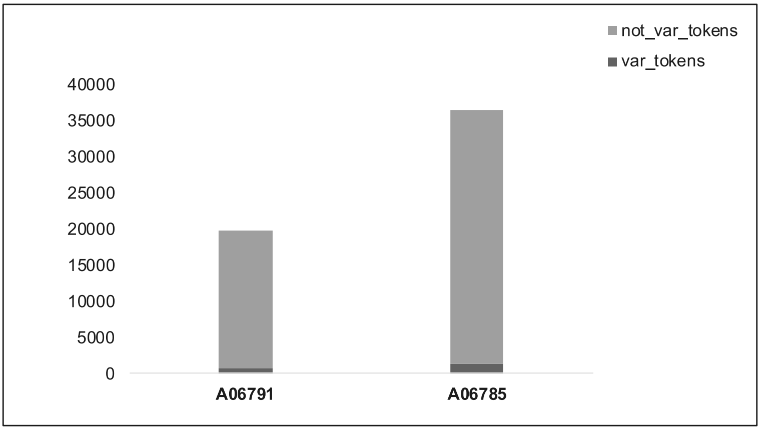

Fig. 2 Quantification of spelling variants in Malynes’ selected texts after normalisation via VARD (threshold 81%).

classified with the automatic POS annotation of the KVL conducted via #LancsBox. From an overall perspective, it is evident that almost all the results are nouns, regardless of their distinctive declinations. In particular, the majority of the lexical items tagged are apparently proper nouns; and this must be taken with great caution, for the massive use of capital letters in EModE may have produced additional errors (Baron, Rayson and Archer 2009b: 53). A few examples taken from the list will offer a clear idea of the items which the tool still tag as variants after normalisation occurred: *starling, caregacaon, Embden, Zeilan, Beatilhas, Cassas, lin, Mangelin, Middleborough, comun*[19].

[19] The words are listed on the basis of their frequency.

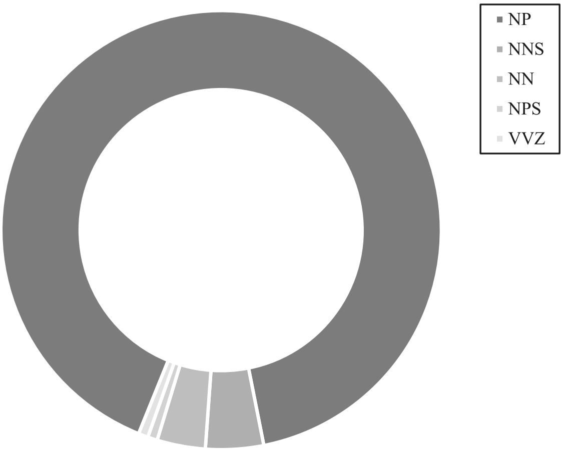

Fig. 3 POS-tagging[20] of remaining variants after automatic and manual normalisations of Malynes's text A06791 via VARD.

Lastly, the manual annotation of the texts via VARD, although not providing dramatic results, has revealed that the usage of Latin as a language increased in Malynes' production from 1601 to 1623 (tagged items showed 0,34% and 1,03% of the texts being in Latin, respectively). This obviously requires more qualitative analysis to establish whether it may be a stylistic markedness of the time or not: other preliminary results from different texts showed a different scenario, which is subject to higher fluctuations, e.g. 0,12% in 1621 and 2,28% in 1623 (Appolloni 2022b). However, this preliminary quantitative approach, that is based on manual annotation, offers an overall picture for the identification of prospective trends of the diachronic development of the examined variety.

4.2. Data extracted from #LancsBox

The following doughnut charts in fig. 4 show the amount of specific word classes in the attempt to formulate some hypotheses regarding the appropriateness of the selected texts to be investigated as representative of a specialized discourse, that is of EModE economics. The PhD project from which the texts are taken is mainly oriented to taxonomies, and hence predominantly nouns. For this reason, both the data resulting from the tags NN and NNS were aggregated as representing both singular and plural nouns. Although this was the main focus, i.e. to establish whether the texts were sufficiently taxonomic in their nature,

[20] Singular (NP) and plural (NPs) proper nouns; singular (NN) and plural (NNS) nouns; verbs, 3rd person singular present (VVZ).

other tags have proved to be of interest. In general, the normalised snapshot corpus revealed the predominance of nouns as a grammar category in the texts (i.e. 19% of the lexical items in the sample are nouns). The same query searched via CQPWeb in the restricted sub-corpus revealed a similar result: that circa 17% of the lexical items in 13 treatises being published in 1601 and 1623 are nouns, both singular and plural[21]. Additionally, other elements appeared to be of relevance for the purpose of evaluating the suitability of texts as specialized in their nature. For example, the presence of determiners (DT) and existential theres (EX), which are responsible for the referential nature of specialized discourse, e.g. in the construction of the informative structure of the text, which together represent almost the 12% of the sample. Another example is the presence of modal verbs (MD), generally used for rethorical or speculation purposes, that covers approximately half of the percentage in the sample as that covered by lexical verbs[22] (i.e. 2% and 5% respectively). This is to say that just over the 30% of the sample is made of linguistic elements which may be categorised for their function in treatises under the field of specialized discourse[23]. They additionally offer an overall picture of the deviation in POS categories frequency between normalised and original versions of the texts. In particular, it is possible to notice that more lexemes have been tagged as NN/NNS in the original texts than in the normalised ones (22% versus 19% respectively); and this confirms a sufficient degree of deviation which requires normalisation before processing POS-tagging.

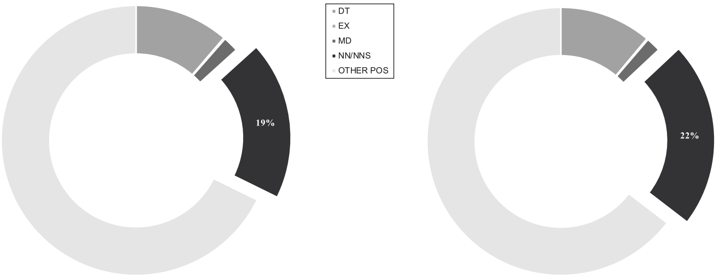

Fig. 4 Quantification of POS-tags to determine the suitability of Malynes's texts for the research purpose (normalised VS original).

[21] The syntax adopted for this query in CQPweb is _NN[1,2].

[22] The POS-tags considered for comparison are: VV, VVD and VVZ, following the POS-tag system incorporated in #LancsBox.

[23] Following these figures, future research could be conducted on possible patterns of language in such texts. Due to the high percentage of prepositions in the texts (the tag IN resulted in 14% of the sample), the following pattern may here of great interest for deeper investigation: prepositions or subordinating conjunctions together with determiners and nouns [Prep + Det + Noun], e.g. *of the realm*, *of the money*, etc.

These stacked area charts in fig. 5 are plotted here to analyse the effect of normalisation via VARD in the extraction of frequency-based wordlist via #LancsBox. And hence, they evaluate the level of deviation in terms of statistical accuracy of the normalised versions of Malynes' texts from the original ones. A first glance can suggest a relevant variance from the frequency values detected by the tool – the measure displayed here is the absolute frequency, i.e. any single occurrence of tokens in the sample. In detail, lexical items such as *money*, *commodity*, *trade* and *merchant* appear for example as not being properly detectable by the tool when lemmatizing non-normalised versions. Respectively, 201 out of 744, 18 out of 372, 29 out of 247 and 81 out of 321 occurrences of the lexemes cited were not measured by the tool in the creation of a frequency-based wordlist. Comprehensively, almost 15% of the occurrences quantified in the absolute frequency values of the selected terms were not detected, especially when lemmatized as in the case proposed here. This possibly confirms the necessity to normalise a historical text regarding economics to be accurately processed via #LancsBox or potentially any other modern tagger for historical purposes.

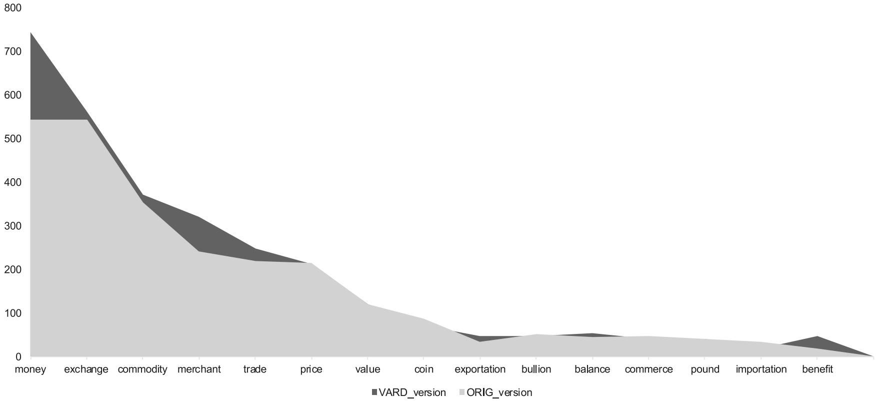

Fig. 5 Quantification of frequency deviation between normalised and original versions of Malynes' texts.

4.3. Data extracted from CQPweb

The statistical tests carried out here to compare the list of lexemes between the two corpora adopted in this case study aim to explore the data-based nature of corpora beyond their linguistic level. In order to pursue proper statistical comparison, corpora must hence be assumed as datasets which require some kind of relationship in terms of data in order to be fully comparable (Culpeper and

Demmen 2015: 96). Hence, since this may influence the acquisition of keyword results in the framework of specific research questions (Culpeper and Demmen 2015: 96), the years of publication and genre[24] were assumed as consistent categories of metadata to create an equivalent reference corpus for the purpose of statistical testing of keywords between the two datasets. Log-likelihood and log ratio, as respectively statistical significance and effect size measures, have been calculated for keyness analysis in the snapshot corpus. This has identified the most representative terms of EModE economics discourse in Malynes' texts.

As a result, the extracted terms are all over-represented in the snapshot corpus, being potentially considered as keywords. However, the data show that *commodity, exchange* and *money* are mostly distinctive of the snapshot corpus for their statistical significance. Additionally, considering the size of the two datasets, the relative frequency of usage of *trade, merchant* and *exchange* is substantially different in the snapshot corpus; this is particularly true for *exportation, commerce* and *bullion*, which by no means occur in the reference sub-corpus. Both these measures also offer an objective reduction of the amount of data for qualitative analysis (McIntyre and Walker 2019: 163).

Table 2. Contingency table for calculating keyness in the snapshot corpus (LL and log ratio).

TERM(s)	**OBSERVED_freq**		**EXPECTED_freq**		**RELATIVE_freq**		**over/under-use**	**LL**	**Log ratio**
	SC	REF	SC	REF	SC	REF	keywords	*p* 0,01	
money	744	23	74,5	692,5	0,012815	0,000043	+	**3266,6**	8,23
exchange	561	5	55,0	511,0	0,009663	0,000009	+	**2559,2**	**10,03**
commodity	372	11	37,2	345,8	0,006408	0,000020	+	**1636,7**	8,30
merchant	321	2	31,4	291,6	0,005529	0,000004	+	1472,6	**10,54**
trade	247	1	24	224	0,004255	0,000002	+	1138,7	**11,16**
price	213	11	21,8	202,2	0,003669	0,000020	+	907,5	7,49
value	119	38	15,3	141,7	0,002050	0,000070	+	388,8	4,86
coin	71	5	7,4	68,6	0,001223	0,000009	+	295,2	7,04
exportation	46	0	4,5	41,5	0,000792	0	+	214,5	**9,74**
bullion	46	0	4,5	41,5	0,000792	0	+	214,5	**9,74**
balance	53	6	5,7	53,3	0,000913	0,000011	+	209,5	6,36
commerce	41	0	4,0	37,0	0,000706	0	+	191,1	**9,57**
pound	40	52	8,9	83,1	0,000689	0,000096	+	71,1	2,84
importation	19	13	3,1	28,9	0,000327	0,000024	+	48	3,76
benefit	46	118	15,9	148,1	0,000792	0,000219	+	43,9	1,9

The highlighted keywords confirm the ideological positions of Malynes (see section 3.2) as linguistically represented in the sample, statistically retrievable with tests and appropriately comparable with an equivalent normative corpus.

Lastly, table 3 shows some examples on the utility of the data from VARD list of known variants post-normalisation to formulate ad-hoc syntaxes to search non-normalised EEBO texts available via CQPweb. Once normalised, VARD automatically produces tags to be downloaded from the KVL in the form of linguistic data, i.e. the occurrences of spelling variants in the texts paired with the

[24] Genre is hypothesized as being a relevant criterion to determine the distance of the relationship of the data to be investigated with a specific corpus (Culpeper and Demmen 2015: 96).

appropriate standardized forms occurring in the normalised version. This dataset proves to be a useful repository of data, for the complete list of variants of a type suggests all the possible spelling alternatives to be included in the syntax of CQPweb queries; this is to retrieve as many occurrences as possible of the same type in a non-normalised corpus. For example, *commodity* may occur in the forms of *commoditie* or *cōmoditie*; thus, such information is of great significance to write a syntax which allow the type *commodity* to be retrieved in a non-normalised corpus. This is the case if the lexeme occurs with single or double *m* (square brackets indicate both the options) or with final *-y* or *-ie* (the asterisk includes whatever inflectional endings), and if it is searched as a noun both as singular or plural (_NN retrieves nouns in singular [1] or plural [2] forms).

Table 3. Examples of variants occurring in the KVL post-normalisation via VARD as data to create proper query via CQPweb.

TEXT(s)	**PdE TERM(s)**	**EModE SPELL_VAR**	**CQPweb_SYNTAX**
A06791	*balance*	ballance	bal[l,]ance*_NN[1,2]
A06785	*commerce*	cōmerce	com[m]erce*_NN[1,2]
A06791	*commodity*	commoditie, cōmoditie	com[m]odit*_NN[1,2]
A06791	*pound*	poūd	pou[n,]d*_NN[1,2]

5. Conclusions

To summarise, this analysis is a preliminary attempt to investigate the language of EModE economics in terms of the opportunities provided by quantitative evidence in historical linguistics, considering the threats from historical texts analysed with corpora. Although quantitative evidence is deemed as solely suitable for the identification of trends (Jenset and McGillivray 2017: 50), this case study has attempted to suggest that quantitative evidence can also provide even further information regarding the text and the dataset created as well.

Historical evidence is complex in its nature, due to the EModE massive variation which requires appropriate tools to be tackled in order to explore language change (Hilpert and Gries 2016: 52). Being most of the annotations and analyses carried out automatically via corpus-based software, a certain margin of error is to be accepted here, particularly when using NLP tools to process historical data (Jenset and McGillivray 2017: 102). This margin of error may represent a major threat to face: the statistical manipulation of data (Baron and Rayson 2009: 1,4,9) and the automatic data retrieval in historical linguistics (Smitterberg 2016: 197) can be seriously compromised. A certain amount of deviation, in fact, appears between the data collected via #Lancsbox when comparing the original version of the sample with the normalised one. Firstly, non-normalised texts have reduced the level of accuracy of the POS-tags being annotated automatically by 3%. Secondly, a considerable number of occurrences, i.e. 15% of the extracted terms, has not been detected in the original form of the sample when being searched in their lemmatized form. In this sense, both data sufficiently suggest the necessity to normalise EModE treatises of economics.

The three corpus-based software utilised here can nonetheless create several opportunities for the benefit of linguistic analysis. Firstly, VARD has been shown not to be a mere variant detector for the normalisation of texts, but also a repository of additional data which can be used to quantify spelling variation and Latin in EModE texts, to detect possible real-word errors, and to ameliorate the syntax for queries in other corpus-based software (e.g. CQPWeb). Secondly, #Lancsbox can be adopted to investigate POS-tagging as a general quantitative indicator of appropriateness of corpus texts for the research questions: being approximately 30% of the texts categorised under the field of specialized discourse, this has proved the suitability of the sample with the research question. Moreover, data from VARD and #LancsBox can be combined to calculate deviation in the frequency profile between normalised and non-normalised texts, and to categorize real-word errors with POS-tagging (in this case, most of them were proper nouns). Lastly, metadata from CQPweb (i.e. years of publication and genre) can help extract a sub-section of EEBO v3 to be used as a reference sub-corpus for the comparison of the frequency profile with the sample. Statistical significance and effect size measures can thus be calculated to reduce the extracted keyword list (McIntyre and Walker 2019: 163), i.e. from 15 extracted terms to 8 key terms.

Primary sources:

Malynes, Gerard, fl. 1586-1641., 2004, *A treatise of the canker of Englands common wealth Deuided into three parts: wherein the author imitating the rule of good phisitions, first, declareth the disease. Secondarily, sheweth the efficient cause thereof. Lastly, a remedy for the same. By Gerrard De Malynes merchant.*, Oxford Text Archive, http://hdl.handle.net/20.500.12024/A06791.

Malynes, Gerard, fl. 1586-1641., 2004, *The center of The circle of commerce. Or, A refutation of a treatise, intituled The circle of commerce, or The ballance of trade, lately published by E.M. By Gerard Malynes merchant*, Oxford Text Archive, http://hdl.handle.net/20.500.12024/A06785.

References:

Appolloni, R. 2022a. The role of English in the circulation of early modern economic discourse (1550-1600): a preliminary corpus-based analysis. Paper presented at AIA 30. Experiment and Innovation: Branching Forwards and Backwards, University of Catania, September 15-17.

Appolloni, R. 2022b. Morphological Aspects of English in Early Modern Economic Discourse (1600-1650): a Corpus-based Examination. Paper presented at II SEDERI International Conference for Junior Researchers of Early Modern English Studies, Universidad Autónoma de Madrid, October 5-7.

Baron, A., and P. Rayson 2008. VARD2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the postgraduate conference in corpus linguistics*, Aston University, Birmingham.

Baron, A., and P. Rayson 2009. Automatic standardization of texts containing spelling variation. How much training do you need? In M. Mahlberg, V. González-Díaz and C. Smith (eds.), *Proceedings of the Corpus Linguistics Conference, CL2009*, 1-25. University of Liverpool, Liverpool.

Baron, A., Rayson, P., and D. Archer 2009a. The extent of spelling variation in Early Modern English. Paper presented at ICAME30, Lancaster University, May 27-31.

Baron, A., Rayson, P., and D. Archer 2009b. Word frequency and key word statistics in historical corpus linguistics. *Anglistik*: 41-67.

Baugh, A.C., and T. Cable 2002. *A History of the English Language (5th ed.)*. London: Routledge.

Beal, J.C. 2016. Standardization. In M. Kytö, and P. Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, 301-317. Cambridge: Cambridge University Press.

Brezina, V. 2018. *Statistics in Corpus Linguistics. A Practical Guide*. Cambridge: Cambridge University Press.

Brezina, V., and D. Gablasova 2018. The Corpus Method. In J.V. Culpeper, P. Kerswill, R. Wodak, T. McEnery and F. Katamba (eds.), *English Language: Description, Variation and Context (2nd ed.)*, 595-609. London: Palgrave Macmillan.

Brezina, V., T. McEnery, and S. Wattam 2015. Collocations in context. A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 139-173.

Culpeper, J. 2002. Computers, language and characterisation: An analysis of six characters in Romeo and Juliet. In U. Merlander-Marttala, C. Ostman and M. Kytö (eds.), *Conversation in Life and in Literature: Papers from the ASLA Symposium* 15, 11-30. Uppsala Universitetstryckeriet.

Culpeper, J., and D. Archer 2018. The History of English Spelling. In J.V. Culpeper, P. Kerswill, R. Wodak, T. McEnery and F. Katamba (eds.), *English Language: Description, Variation and Context (2nd ed)*, 186-199. London: Palgrave Macmillan.

Culpeper, J., and J. Demmen 2015. Keywords. In D. Biber and R. Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 90-105. Cambridge: Cambridge University Press.

Davidse, K., and H. De Smet 2020. Diachronic Corpora. In M. Paquot and S.T. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 211-233. Cham: Springer.

Davies, M. 2015. Corpora: an introduction. In D. Biber and R. Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 11-31. Cambridge: Cambridge University Press.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1): 61-74.

Hardie, A. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380-409.

Hardie, A. 2014. Log ratio: an informal introduction. Lancaster: ESRC Centre for Corpus Approaches to Social Science (CASS). Available at: https://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

Hilpert, M., and S.T. Gries 2016. Quantitative approaches to diachronic corpus linguistics. In M. Kytö and P. Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, 36-53. Cambridge: Cambridge University Press.

Horobin, S. 2016. Manuscripts and early printed books. In M. Kytö and P. Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, 111-126. Cambridge: Cambridge University Press.

Jenset, G.B., and B. McGillivray 2017. *Quantitative Historical Linguistics. A Corpus Framework*. Oxford: Oxford University Press.

Kytö, M., and E. Smitterberg 2015. Diachronic registers. In D. Biber and R. Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 330-345. Cambridge: Cambridge University Press.

López-Couso, M. 2016. Corpora and online resources in English historical linguistics. In M. Kytö, and P. Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, 127-145. Cambridge: Cambridge University Press.

McEnery, T., and A. Hardie 2012. *Corpus Linguistics. Method, Theory, Practice*. Cambridge: Cambridge University Press.

McIntyre, D., and B. Walker 2019. *Corpus Stylistics: Theory and Practice*. Edinburgh: Edinburg University Press.

Nevalainen, T. 2006. *An Introduction to Early Modern English*. Edinburgh: Edinburgh University Press.

Nevalainen, T., and H. Raumolin-Brunberg 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Harlow: Pearson.

Newman, J., and C. Cox 2020. Corpus Annotation. In M. Paquot and S.T. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 25-48. Cham: Springer.

Rayson, P. 2015. Computational tools and methods for corpus compilation and analysis. In D. Biber and R. Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 32-49. Cambridge: Cambridge University Press.

Rayson, P., and A. Potts 2020. Analysing Keyword Lists. In M. Paquot and S.T. Gries (eds.), *A Practical Handbook of Corpus Linguistics*, 119-139. Cham: Springer.

Robbins, L. 2000. *A History of Economic Thought. The LSE Lectures*. Princeton and Oxford: Princeton University Press.

Smitterberg, E. 2016. Extracting data from historical material. In M. Kytö and P. Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, 181-200. Cambridge: Cambridge University Press.

Taavitsainen, I. 2016. Genre dynamics in the history of English. In M. Kytö, and P. Pahta (eds.), *The Cambridge Handbook of English Historical Linguistics*, 271-285. Cambridge: Cambridge University Press.