Baltic Journal of English Language, Literature and Culture Vol. 13, 2023: 48–68
https://doi.org/10.22364/BJELLC.13.2023.04

THE APPLICATION OF DIACHRONIC CORPUS COMPILATION PRINCIPLES IN A PILOT STUDY OF SUBJECTIVITY

KRISTĪNA KORNELIUSA and ZIGRĪDA VINČELA
University of Latvia, Latvia

Abstract. Researchers claim (see Egbert, 2018) that, irrespective of the growing amount of corpora, there is insufficient focus on the research and discussion of corpus creation and analysis challenges. The ongoing international project LEXECON (2021-2024) raises awareness about these kinds of issues. The goal of this study is twofold: firstly, to explore corpus creation stages in relation to compilation criteria; and secondly, to pilot the functionality of the created subcorpus by researching first-person pronoun variations to uncover the subjectivity across the subcorpus genres. The pronouns were explored by observing their relative frequency, context, and surplus-deficit index. Two corpus analysis tools—Sketch Engine and Hyperbase 10—were applied. The corpus creation results confirm that balance is the most challenging corpus criterion to fulfil, whereas corpus editing is the most time-consuming corpus creation stage. The results obtained via first-person pronoun extraction confirm that the context and surplus-deficit index contribute to the research results no less than the relative frequency data. The analysis of personal pronoun data variations shows that essays contain the fewest first-person singular pronouns; however, in other genres, they often do not convey an authorial stance. Moreover, a greater surplus of possessive case reflects a more active authorial stance as opposed to objective case.

Key words: corpus creation, frequency, surplus-deficit index, subjectivity, first-person pronouns

INTRODUCTION

The creation of specialised diachronic corpora allows for the tracking of the development of various scientific and professional fields. It is assumed that the linguistic changes reflect the changes in the respective discipline. Such corpora also ensure the development of interdisciplinary research. However, as Egbert (2018: 35) admits, despite the number of corpora compiled for linguistic analysis, 'there has been very little discussion in the corpus linguistics literature about the process of corpus design and creation'.

In the current research, a corpus of texts on political economy published between 1841 and 1850 is presented as a case study of personal pronoun variations in its texts to reveal their subjectivity in different genres of texts on political economy. The corpus has been compiled within the framework of an ongoing international project, *LEXECON. The Economic Teacher: a transnational and diachronic study of treatises and textbooks of economics (18th to 20th century). Intra- and interlingual corpus-driven and corpus-based analysis with a focus on lexicon and argumentation* (project code 2020X24S9N), implemented by a joint research team of the University of Pisa, the University of Padua and the University of Palermo in collaboration with junior contributors from other countries. The aim of the LEXECON project is to create a corpus of texts on political economy spanning from 1750 to 1970 and thus enhance the interdisciplinary research connecting the fields of economics and applied linguistics. The project is funded by the Italian Ministry of University and Research for the period 2021-2024 as research of national interest. However, since the corpus includes six European languages—Italian, English, French, German, Spanish and Portuguese—and involves native and/or proficient speakers of these languages from the whole of Europe in the creation of the lexicon, the project is of high importance for establishing an international research network (Guidi et al., 2021). The involvement with the project became possible thanks to the Erasmus+ programme and has continued since March 2021.

The goal of this study is twofold: firstly, to explore corpus creation stages in relation to corpus compilation criteria; and secondly, to pilot the functionality of the created subcorpus by researching first-person pronoun variations to uncover the subjectivity across the subcorpus genres. In order to achieve this goal, the following research questions were asked: (1) how are the corpus criteria complied with during the subcorpus creation process for the LEXECON database; (2) what are the challenges of each subcorpus formation stage; (3) what functionalities do the corpus analysis tools provide in the subjectivity analysis of the genres included in the subcorpus; and (4) does the frequency of the first-person pronouns correlate with the expectations regarding the subjectivity of each selected genre?

Due to the two-fold goal of the study, this article is arranged into two major sections. The first section is devoted to the subcorpus of LEXECON creation and uncovers the theoretical background behind corpus definition, its principles, the procedure, and the results of the subcorpus compilation. The second section is

devoted to the created subcorpus piloting and uncovers the theoretical background on subjectivity and first-person pronouns, the procedure of first-person pronoun extraction, and the obtained results.

SUBCORPUS CREATION

1 THEORETICAL BACKGROUND

Definitions of a corpus are based on corpora principles that are formulated and proposed by linguists as criteria frameworks for corpora creation to address goal-oriented research questions. Even if researchers agree that contemporary corpora are computerised, 'unless otherwise stated' (Weisser, 2016: 23), i.e., machine-readable (McEnery and Wilson, 1996: 29), and that their criteria features typically refer to such interrelated aspects as authenticity, representativeness, balance, sampling and size of corpora (Kilgarriff and Grefenstette, 2003; Gatto, 2014: 8-15; McEnery and Brookes, 2022: 35-47), the considerations vary depending on the researchers' theoretical frameworks concerning the role of these corpora features.

Since corpus linguistics is the 'study of language based on examples of real use' (McEnery and Wilson, 1996: 1), the authenticity of corpora texts is a crucial feature. Authenticity puts into practice 'the empirical trend' (Sampson, 2013: 281) of corpus linguistics since and before the first computer-generated concordances were used. Authenticity is also the key argument made by Kilgarriff and Grefensterte (2003) in their discussion of the World Wide Web in the context of corpus linguistics.

Corpora representativeness is another core criterion of a corpus. According to Biber (1993: 243), representativeness 'depends, first of all, on the extent to which [the sample] is selected from the range of text types in the target population; […] and the techniques used to select the sample from population'. Representative corpora in linguistic research are 'a source for extracting instances of a particular linguistic feature' (Egbert, 2018: 28), the value of which, as stated by Gatto (2014: 12), is that the uncovering of repeated use of linguistic features can lead to the formation of generalisations as far as these samples are representative. Therefore, a commonly applied approach for the creation of representative corpora (see Gablasova, Brezina and McEnery, 2019: 131) includes the explicit presentation of information about language samples that are included in a corpus (or metadata), information about the method of their collection, and the availability of the documentation that details corpus design criteria (see McEnery, Xiao and Tono, 2006: 18).

Corpora balance and sampling (see Gatto, 2014: 12-13) directly contribute to their representativeness. According to McEnery et al. (2006: 16), balance refers to the feature of corpora to comprise 'a range of text categories' included in a corpus, or according to Biber (1993: 243), the 'extent to which it includes the range of text samples in a language'. However, Gablasova et al. (2019: 134) argue that balance, or the range of text categories in a corpus, is secondary to the information about language samples and corpus structure because 'the balance of a dataset

is often defined individually in different studies by selecting an appropriate subset of the corpus that can answer a particular research question'. In addition, specialised corpus compilers might address text availability challenges. Thus, even if Weisser (2016: 45) considers that balance 'may be more easily achievable, especially for domain-specific corpora or limited fields of investigation, because often there are relatively definable criteria for what represents a certain genre of text or domain', Weisser (ibid.) also admits that the availability of specific texts depends on 'a number of legal points you ought to consider when making decisions about which data to incorporate'. These challenges might refer to the choice of text categories and sampling approaches. For example, random sampling within the defined 'text categories' (McEnery et al., 2006: 16) might be unfeasible due to the text availability constraints. Consequently, the definition of these categories and metadata detailing are vital to the creation of balanced specialised corpora (McEnery et al., 2006: 16). Sampling (the size of individual texts in a corpus), like balance, is widely discussed by corpus linguists because linguistic features 'are not distributed equally in a language' (Gablasova et al., 2019: 132), and hence texts. Biber (1993) explains that the size of each text sample accounts for the capacity of linguistic data extraction. For example, Biber (1993: 249) details that in the case of extracting frequent linguistic features, the inclusion of short text parts of the same size (e.g. 1000 or 2000 running words covering consistently selected text portions) can be sufficient in a corpus. However, he reminds that careful testing of text samples is required to find out how far the selections would represent the linguistic characteristics of the whole text. Sacrificing text integrity in corpora was strongly criticised by Sinclair (2005: n.p.), who claimed that 'there is no virtue from a linguistic point of view in selecting samples all of the same size'. The inclusion of full texts in specialised corpora is also supported by the capacity of contemporary concordance tools, for example, *SketchEngine*, that enable accessing linguistic features in a wider context (Kilgariff and Rychly, 2008; Brunet, 2011).

Corpus size is a common feature that contributes to its representativeness. According to linguists (Gatto, 2014: 14), corpora size refers to their finiteness, the number of texts and running words that can serve as departure points for quantitative research. Egbert, Larsson and Biber (2020: 4) remind us that 'corpus size has been a major goal within corpus linguistics throughout its history'. However, two approaches to this significant feature are observable. One approach puts a strong emphasis on corpus size advocated by, for example, Sinclair (1991) and Hanks (2012). The other approach puts specific emphasis on the role of corpora representativeness (along with the size of a corpus) in the context of a corpus potential to address research-goal-oriented questions (Biber, 1993; McEnery et al., 2006; Egbert, 2019). McEnery and Brookes (2022: 41) conclude that corpus size depends on the correlation of research goals and ambitions with 'practical considerations and limitations regarding what is possible'. Therefore, the use of corpora built manually by a single researcher is normally considerably smaller and produces less generalisable results than corpus-based studies examining a linguistic feature across built-in online general corpora or corpora in which teams of multiple researchers

and corpus-building assistants are involved. For example, in the large corpus-based study conducted by Vinčela in 2017, the queries were examined and compared across corpora, the largest of which amounted to 1.9 billion words (Vinčela, 2017: 162).

Finally, corpus editing, also called *text normalisation* by some scholars (Dash and Ramamoorthy, 2019: 35), 'involves diverse tasks of text adjustment and standardization' which, in a study using computer software, is necessary 'to improve utility of the texts' (ibid.). This process is further described in detail in Section *3.3 EDITING*.

2 METHODOLOGY

The LEXECON subcorpus creation methodology is based on the theoretical framework underlying corpus creation principles—authenticity, representativeness, balance and sampling, and size—proposed and discussed by linguists (McEnery et al., 2006; Gatto, 2014; Egbert et al., 2020; Reppen, 2022).

The sequencing methodology of the subcorpus compilation stages complies with and aims at the corpus creation principles explained by the previously mentioned corpus linguists and supported by the LEXECON project team (Guidi et al., 2021): (1) bibliographical research and the selection of corpus texts by exploring the available databases aim at subcorpus authenticity and representativeness; (2) narrowing down the primary search aims at sampling, balance, and size; (3) corpus editing, which aims at subcorpus balance, refers to the elimination of the mistakes occurring in the texts due to their conversion, decision-making about the paratext removal, mark-up solutions, and the correction of the authentic mistakes in the corpus texts (e.g. orthographic errors); and (4) corpus structuring, i.e., the creation of its architecture underlying subcorpus balance.

3 PROCEDURE

The subcorpus compilation stages in the following sections comply with the corpus criteria.

3.1 AUTHENTICITY AND REPRESENTATIVENESS

The samples included in the subcorpus were authentic texts on political economy selected using such databases as the HathiTrust Digital Library, the WorldCat, the Library of Congress, and Google Books. The representativeness of the sample was ensured by the application of the following search filters: language, year of publication, and keywords related to the field of political economy. Additionally, the full-text items from the public domain were preferred over the titles with limited access. For the subcorpus designed for the current research, only those titles that were first published between 1841 and 1850 were selected. The re-issues of earlier works published within the decade were not considered representative of the language of the time. The information about the first published edition was extracted from the bibliographical entries, mostly on Google Books. The selection

of the texts for analysis using the search filters by language, field, publication year, number of the edition, and accessibility is also closely related to the sampling principle. This can be regarded as the first stage of sampling. The further selection is described in the following section *3.2 SAMPLING, BALANCE, AND SIZE*.

3.2 SAMPLING, BALANCE, AND SIZE

The primary selection of 74 titles was narrowed down to 10 items only, for the sake of balance (i.e., so that all the corpus categories, in this case, genres, would be covered equally) and, most importantly, due to the time constraints, which also impacted the corpus size. It was decided to focus on four genres—essay, academic lecture, textbook and treatise—as they were the most represented ones in the list of the primary selection. Hence genre—just like language, theme, and publishing year—became another filter for text sampling. The texts were selected so that they would comply with the balance criterion. It was decided to select the texts so that all four genres would be represented by two to three items. Nevertheless, it has to be noted that the balance was not fully achieved, since academic lectures were found to be considerably shorter than texts belonging to other genres, for example, treatises. An equal number of items belonging to the genres of academic lectures and treatises, respectively, did not ensure an equal size of these corpus segments.

3.3 EDITING

Despite the fact that some scholars believe that 'corpora representing the written form of a language usually present the smallest technical challenge to construct' (McEnery and Hardie, 2012: 3-4), thanks to the introduction of Unicode, this is an overstatement. The texts available in TXT format, especially those converted from the PDF versions of 19th-century editions, require extensive editing due to the errors caused by the imperfections of optical character recognition, which have to be removed manually by the researcher (see Figure 1).

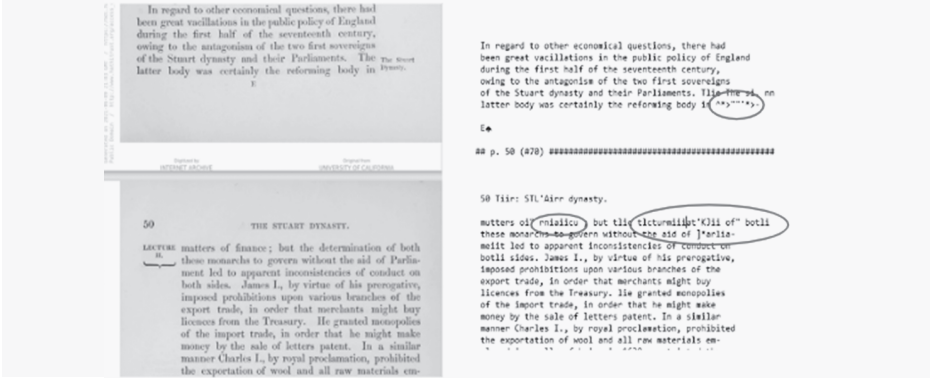


Figure 1 **The original PDF file and the imperfections of the converted TXT file**

Even if the text does not contain any errors, it is often necessary to remove any paratext that would influence the data processing (e.g., the chapter titles and footnotes, if they were not considered in a particular study).

For this research, the footnotes were preserved because they were believed to contain valuable additional information provided by the author of the text, thus representing his style (the authors of all the texts analysed are men, hence the use of the pronoun). They were included as part of the main text, marked by an asterisk (or several asterisks, depending on the order of the footnote on the page), and separated by curly brackets (see Example 1).

[1] Each step forward in the exertion of this power lays a foundation for future progress {* Mr. Senior founds the whole science of political economy on a moral propensity in man, in his first axiom: "Every man desires to obtain additional wealth with as little sacrifice as possible." — Outline, p. 139.}. [A 1]

Also, additional markup was added in the form of page numbers to preserve the structure of the text, as shown in Example 2.

[2] Without impugning, therefore, the general belief, that wealth consisted exclusively of gold and silver, //8// the earliest writers ventured to question the wisdom of prohibiting the exportation of the precious metals. [A 9]

In order to ensure that the page numbers do not influence the collocations, they were inserted at the end of a clause or a sentence.

Lastly, according to the conventions set by the LEXECON project team, the orthographic and syntactic errors, as well as typos, were preserved to distinguish the specific edition from other editions of the same work and to not remove the linguistic peculiarities of the decade unnecessarily. Consider the examples below (the original spelling and syntax were preserved; italics were added by the authors):

[3] *In so far* as it contributes to give increased activity to industry, it is properly among the causes which it is the aim of our science to *develope*. [A 6]

[4] The imperfection of our senses, even when assisted by the most elaborate *intruments* of art, must ever present an obstacle to the attainment of mathematical exactness. [A 6]

Example 3 provides a case of an orthographic peculiarity that would have been considered an error from the point of view of modern conventions: *develope* instead of *develop*; it is a consistent spelling throughout the text. Additionally, in modern American English, *in so far* as is spelled *insofar as*, but the American-born Henry Middleton uses the British variant of the spelling in the text. Example 4 shows a typo: *intruments* instead of *instruments*; the typo is kept to distinguish this edition from the newer ones, in which, most likely, the error is removed.

One should bear in mind that, since the tools used for the analysis are designed for the current norms and conventions of the English language, the preservation of the orthography of the 1840s, as well as the typos, leads to faulty part-of-speech tagging and parsing. This is also confirmed in theoretical sources; For example, Bollman (2019: 3885) admits that 'spelling variation is one of the key challenges for NLP on historical texts, affecting the performance of tools such as part-of-speech taggers or parsers and complicating users' search queries on a corpus'. In order to make the corpus reusable for analysing a wider scope of linguistic features, the LEXECON team is currently working on possible solutions for this issue as the project is in progress (C. Flinz, 2022, personal communication, October 24; M. E. L. Guidi, 2022, personal communication, October 24). However, the corpus can already be used for linguistic analysis of features that are not affected by outdated spelling or typos. First-person singular pronouns, selected for this study, are one such option. Moreover, as it is explained in Sections *3.1 SKETCH ENGINE* and *3.2 HYPERBASE 10*, the authors have used the function of inserting a list of forms in the query or checking separate forms instead of relying on automatic part-of-speech tagging.

3.4 STRUCTURING

The structure of the subcorpus is presented in Appendix 1. The subcorpus of the LEXECON database was created and named after the language and the decade researched—ENG 1841-1850—to distinguish it from other subcorpora of this multilingual diachronic database. As it can be seen, the corpus has been structured according to genre because the case study selected for the research on subjectivity deals with the comparison of the use of first-person pronouns across the genres of texts on political economy included in this subcorpus. A different research topic would have prompted a different framework for this subcorpus structure (e.g., a division according to the author or his origin). Ten texts were selected in total (see Table 1).

Table 1 **The information about the texts analysed**

Genre	Surname	Name	Title	Year
Textbook	Banfield	Thomas Charles	Four lectures on the organization of industry; being part of a course delivered in the University of Cambridge in Easter term 1844	1845
Treatise	Burton	John Hill	Political and Social Economy: its practical applications	1849
General Essay	Duncombe	Charles	Duncombe's free banking: an essay on banking, currency, finance, exchanges, and political economy	1841

Genre	Surname	Name	Title	Year
Textbook	Gilbart	James William	Lectures on the history and principles of ancient commerce	1847
Academic Lecture	Hancock	William Neilson	Three lectures [...] delivered in the theatre of Trinity College, Dublin, in Hilary term, 1847, by W. Neilson Hancock	1847
General Essay	Middleton	Henry	Four Essays	1847
General Essay	Mill	John Stuart	Essays on Some Unsettled Questions of Political Economy	1844
Academic Lecture	Smith	Herbert	A lecture on the capability of Great Britain and Ireland to give employment, and provide a sufficient maintenance for the whole population	1846
Textbook	Twiss	Travers	View of the progress of political economy in Europe since the sixteenth century	1847
Treatise	Ware	Nathaniel A.	Notes on political economy, as applicable to the United States. By a southern planter	1844

4 DISCUSSION OF THE RESULTS

During the corpus compilation stage, it was found that, although the text sample was formed so that the balance criterion would be fulfilled, the objective was not entirely achieved. This was mainly due to the fact that academic lectures are much shorter than textbooks, treatises and collections of essays. In terms of the number of tokens, the subcorpus of lectures amounted to 4 percent only (see Table 2).

Table 2 **The size of each set of texts by genre**

Name	Tokens	%
Essay	213,284	28.5
Lecture	30,069	4
Textbook	230,478	30.8
Treatise	273,739	36.6

Since 'the proportions of different kinds of text […] [are expected to] correspond with informed and intuitive judgements' (McEnery and Xiao, 2005: n. p.), there is no uniform way to approach corpus balance. In the current analysis, two ways were possible: balance the corpus based on either the number of tokens or the number of represented texts. The first approach was selected due to the limited number of available texts (only two lectures published in the 1840s were available in full-text format). The second option would require increasing the number of texts represented in order to achieve a more or less equal volume in all the subcorpora. This could have been done had the research object been different.

The corpus editing stage represented another challenge—the amount of time to be spent on it. This can be seen in the corpus compilation timeline (Figure 2).

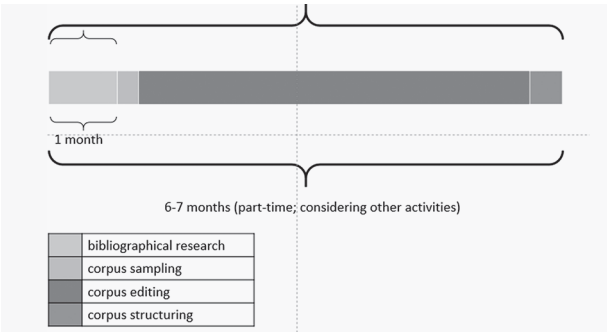


Figure 2 **Corpus compilation timeline**

As it can be seen, one month was dedicated to bibliographical search. This was the approximate time period assigned by the LEXECON project team for the first stage of the corpus formation; the period could be longer if the research decade were more fruitful in terms of publications on political economy. The time period included not only the bibliographical research itself but also the period during which the selected entries had to be approved by the researchers in the field of political economy. Corpus sampling took roughly a week. The corpus editing stage was the most time-consuming, taking up to five months, given that the removal of errors and the insertion of markup (see Section 3.3 *EDITING*) had to be performed manually. Finally, the corpus ready for analysis was structured in roughly two weeks, taking into account double-checking and approval of the genre attribution.

This shows that the corpus compilation and linguistic analysis timelines should be planned accordingly, allowing enough time for corpus compilation before performing the actual linguistic analysis and taking into account the possible approval periods if the corpus is a part of a project or any other type of teamwork. The timeline may be considerably different, depending on the size of the corpus and the number of errors to be removed; also, it is important to consider how many hours per day the researcher spends on corpus compilation and whether he or

she has other commitments. The current timeline reflects the corpus compilation process for a roughly 640,000-token corpus with errors caused by the natural damage to pages that occurred over the course of more than 150 years. Around 20 hours a week were spent on corpus editing.

As a result, the created subcorpus complied with LEXECON corpus creation criteria and was piloted for the research of first-person pronoun variations feasible without corpus annotation. Meanwhile, it is worth noting that the LEXECON team is addressing the issue of preserving the authentic spelling and avoiding the resulting inaccuracies of contemporary tools to enable a more varied linguistic diachronic analysis.

SUBCORPUS PILOTING: PRONOUN VARIATIONS

1 THEORETICAL BACKGROUND

In linguistics, subjectivity is seen as 'self-expression in language' (Fina, 2009: 121-122; Baumgarten, Du Bois and House, 2012: 1). While subjectivity can be expressed in a variety of ways, both lexically (including all parts of speech) and syntactically, the limitations of the research require narrowing down the selection of subjectivity markers to just a few features.

For this reason, the view of subjectivity as an authorial stance and the extent to which the author reveals it (Pho, 2012: 97) is taken. House (2012: 140) defines stance as 'the cognitive and affective attitude of a speaker towards the events and states of affairs he or she is describing or using in an utterance as well as the attitude towards the language used in the interaction'. Fina avoids the term *stance* and indicates instead that subjectivity is 'the presence of the speaker in language' (Fina, 2009: 171); however, this definition conveys the same idea.

The first-person singular pronoun was chosen over other linguistic features that uncover subjectivity, as it overtly conveys an authorial stance. While other pronouns, particularly the first-person plural pronoun *we* and the second-person pronoun *you* (both singular and plural), can convey the authors' attitude towards the reader or hearer (Langacker, 2009: 122), their sense of belonging to and/or exclusion from social groups (ibid.; Tantucci, 2021: 16), and possibly proximity to or distancing from the ideas expressed, they would additionally refer to the exchange of attitudes and feelings (Tantucci, 2021: 7) rather than simply stating them. In this case, the scope of this research would expand to include intersubjectivity (ibid.). This could be done through further research. Currently, it is decided to focus on the authorial stance and personal involvement, i.e., the first-person pronouns in subcorpus texts.

The function of first-person pronouns has not changed since the 1840s. The 1847 edition of *The Principles of English Grammar* (first published in 1834) by Bullions (1847: 22) states that '*I* […] denotes the speaker' in the same way that Biber et al. (2021: 41) state that 'first person pronouns "function" to refer to the speaker/writer'. The differences refer only to the classification of the personal pronouns. Bullions (1847: 22) lists *I*, *mine* and *me* as nominative, genitive and accusative

declensions of the first-person singular pronoun. The form *my* is listed in a separate chapter on *adjective pronouns* (ibid.: 25) and included in their subgroup, possessive pronouns (ibid.: 26). The reflexive pronouns, however, are included in the chapter on personal pronouns, though marked separately as reflexive (ibid.: 22). These differences should not be seen as a reason to modify the list of forms examined.

It is presumed that the frequency of first-person pronoun use may be different in various genres included in the subcorpus, as some of them may be more involved and personal than others. The analysis of the previous research on contemporary genres forms a series of expectations regarding the subjectivity of the texts of the 1840s. In essays, the authors express their personal views and arguments regarding the matter discussed and thus may seem to involve the reader in a reflexive dialogue (Chadbourne, 1983: 50). Textbooks and academic lectures serve the purpose of informing and educating students (Malavska, 2016: 64); the only difference between the two is the mode of delivery—written and spoken, respectively. Finally, a treatise can be seen as a mixture of the genres discussed above, since it combines argumentation and methodological discussion.

Based on these considerations, it is expected that essays published in the 1840s, similarly to contemporary essays, might contain the most first-person pronouns, while textbooks contain the least. Academic lectures could contain more of the feature since it is an example of spoken discourse.

2 METHODOLOGY

A case study of the first-person singular pronoun variations across the subcorpus texts of four genres (essay, academic lecture, textbook and treatise) was applied to illustrate the subcorpus applicability in subjectivity research of its texts. The electronic format, one of the corpus criteria (Gatto, 2014), allows the researcher to use computer software for data processing and analysis as a research methodology. The data retrieved using the selected tools for the pilot were both quantitative and qualitative. The applied methodology included elements of both qualitative and quantitative methods: (1) the extraction of the first-person pronoun relative frequency (RF) across the texts of the four genres and the analysis of the extracted concordance lines by the application of the corpus analysis tool Sketch Engine; and (2) the obtaining of the surplus-deficit index with the help of Hyperbase 10. While the extraction of RF and the surplus-deficit index are examples of quantitative research methodology, concordance extraction is an example of qualitative research methodology.

It should be noted that the term *surplus-deficit index* that refers to linguistic distribution is specific to the Hyperbase 10 software. Moreover, it was derived by the authors of this article by translating the original French terms *excédent* (surplus) and *déficit* (deficit) used in the Hyperbase 10 interface and the user manual by Brunet (2011: 40). He states that this index allows to measure *distribution*; the French term would correspond to *linguistic distribution* and *dispersion* in English (Baker, Hardie and McEnery, 2006: 61). The index is used to see how much more or less frequently the first-person singular pronouns are found in the texts analysed than in

the general corpus. Brezina (2018: 49), discussing standard deviation as a measure of dispersion, writes about the 'distance from the mean', i.e., the difference in RF values between the target corpus and the reference corpus. Using his terminology, a surplus would correspond to a *positive distance* (i.e., the value of the RF is greater than the mean), while a deficit corresponds to a *negative distance*, i.e., 'the values [are] smaller than the mean' (ibid.). The reference corpus' RF values in the Hyperbase 10 terminology correspond to Brezina's concept of the mean.

3 DATA EXTRACTION PROCEDURE

3.1 SKETCH ENGINE

Sketch Engine is an online text analysis tool that is convenient to use for linguistic data extraction purposes (Kilgarriff and Rychly, 2008); therefore, the following functions were used in the process of the first-pronoun variation research across the subcorpus texts:

For the extraction of words from a particular list, the function 'from this list' was used for separating personal pronouns from other items belonging to the same part of speech. Due to the fact that pronouns are a closed class of words, and the list of personal pronouns is clearly defined in the theoretical sources, the list was short and feasible to use.

For the extraction of concordance, the query was set in the way it has been described above. It allowed the extraction of absolute frequency (AF), relative frequency (RF) and contextual information (concordance lines). Due to the differing sizes of the subcorpus segments, AF was discarded and it was decided to focus on RF of the pronouns.

The query can be formed using wildcards—special formulae to extract complex constructions or sum up multiple query options. The query formed for the pronoun extraction can be seen in Figure 3.

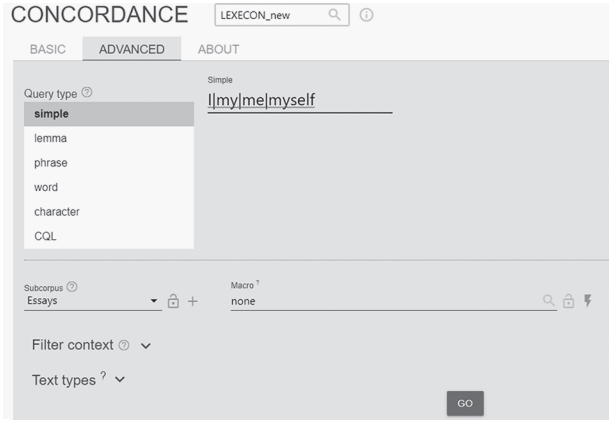


Figure 3 **The query formulated for the concordance extraction**

The reason each form was not checked separately in Sketch Engine for frequency is that the output would only let one compare the frequencies across the four genres rather than to the general reference corpus. This option is available in Hyperbase 10, which is described in the next section. What is more, the form in which the results are automatically presented by Hyperbase 10 is faster and easier to extract than from Sketch Engine, where each RF value needs to be registered manually from the window above the concordance lines (see Figure 4).

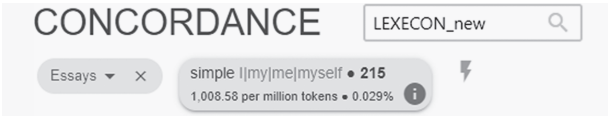


Figure 4 **The AF (215) and RF (1,008.58 per million) frequency of the first-person pronoun forms in essays**

Hence, it was more convenient to sum up the frequencies for all the points and draw conclusions based on these numbers.

3.2 HYPERBASE 10

Hyperbase 10 is a software created by the French linguist Etienne Brunet in 1989, and it is used for textometric analysis (Brunet, 2011). One of the functions available is the extraction of the surplus-deficit index for the selected linguistic feature. It shows how much more or less frequently a linguistic feature is found in the text than in the reference corpus. The reference corpus used by default is the British National Corpus.

The texts were uploaded into the system with the following names: TEX1 (essays), TEX2 (academic lectures), TEX3 (textbooks), and TEX4 (treatises). Next, the search term was entered into the concordance tool, and a function of showing a histogram of surplus and deficit was selected. An example of the histogram is shown in Figure 5.

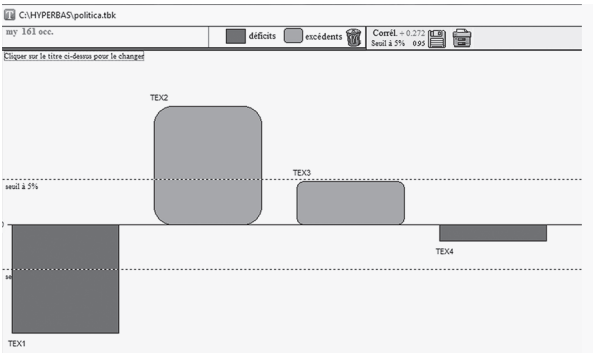


Figure 5 **Histogram for the form my**

4 DISCUSSION OF THE RESULTS

The RF per million tokens, summed up for all forms of the first-person singular pronoun, was retrieved from Sketch Engine. The results can be seen in Table 3.

Table 3 **RF of first-person pronoun forms across genres**

Genre	*Essays*	*Lectures*	*Textbooks*	*Treatises*
RF per million	1009	4556	1663	1725

As it can be seen, contrary to expectations, the essays contain fewer first-person singular pronouns per million than the other genres, while in academic lectures the linguistic feature is found four times more often than in the other texts.

Instead of concluding that the expectations set for the genres are faulty, one has to bear in mind that personal judgement and attitude can be expressed in other ways than using first-person singular pronouns. Also, the quantitative results cannot be fully interpreted without concordance that allows viewing the context in which the pronoun is found and shows what it refers to. To exemplify this, the excerpts containing direct quotations in a treatise by Burton (1849; A 2) are shown below (original spelling and syntax preserved, italics added by the authors).

[5] "*I* am amazed," he said in a level tone of voice," at the attack the noble duke has made on *me*. Yes, *my* lords" – considerably raising his voice – "*I* am amazed at his Grace's speech. [A 2]

[6] '*I* have purposely,' he said, 'omitted any endowment to keep the Arboretum in order, as *I* know by experience that *I* shall best provide for its future preservation by intrusting it to those who will enjoy and profit by it, and who will take an interest in its permanence.' [A 2]

These quotations were taken by Burton from other works; therefore, the first-person singular pronouns found in them in no way reflect his authorial stance. Whether to subtract them from the absolute frequency and recalculate the relative frequency for more accuracy is up to the researcher. The aim of the current paper is simply to demonstrate that relying only on quantitative information does not allow one to see all the nuances and may lead to faulty conclusions. In most of the cases, however, the use of first-person pronouns directly indicated the authorial stance:

[7] *I* doubt the truth of what they say, and *I* will tell you *my* reasons. [A 8]

[8] It occurred to *me* that *I* could not fix upon a subject more important or more interesting. [A 4]

Apart from frequency and concordance, this analysis intends to show the results of surplus-deficit extraction (see Table 4).

Table 4 **Surplus-deficit of first-person pronoun forms across genres**

Pronoun	Essay	Lecture	Textbook	Treatise
I	-7.6	8.7	3.2	-1
my	-5.1	5.5	2	0.8
me	-2.2	2.7	-3.7	3.8
myself	-0.9	2.4	-0.8	-1.1

The results largely correspond to those extracted in terms of frequency. Essays have a deficit in all forms of the first-person personal pronouns, while in the academic lectures, there is a consistent surplus. The results for textbooks and treatises reveal that there is a surplus in some forms and a deficit in others. The summed-up relative frequency of all the forms is roughly the same for both genres; however, the surplus-deficit analysis reveals that in textbooks there is a prevalence of the nominative and possessive forms *I* and *my*, while in treatises, the first-person pronoun forms are dominated by the objective form *me*. This allows one to presume that the personal stance in textbooks is more prominent because the author's role in the text is more active. However, in order to draw more precise conclusions, as has been exemplified earlier, concordance has to be taken into account as well.

CONCLUSIONS

The study demonstrates that corpus creation is a complex process that requires careful consideration of the corpus criteria discussed by researchers, the decision-making concerning selection and sequencing of corpus creation stages, as well as the corpus structuring details to address research questions. The study results also revealed the topicality of time management due to the unpredictable challenges during the corpus creation process.

The research results revealed that each stage of corpus compilation allowed for compliance with the corpus criteria: bibliographical search ensures authenticity and representativeness; corpus structuring ensures sampling, balance and size; and corpus editing and further linguistic analysis are possible due to the electronic format of the corpus, which provides opportunities for optical character recognition and the use of computer software.

The corpus compilation revealed that balance was found to be the most challenging corpus criterion to fulfil because the text size variations pertaining to the specific features of the selected genres can cause token count disproportion across their texts, hence corpus. In addition, the limited availability of the texts representing these genres can add another challenge to the creation of a balanced corpus.

Corpus editing was discovered to be the most time-consuming stage of corpus creation. However, this may change depending on the corpus size, the visual quality of the text, and the research goal.

The application of linguistic data extraction software revealed that terminology variations can occur in the interface of research tools, particularly in Hyperbase 10. It also revealed that the concept of surplus and deficit can be aligned with the terms *positive distance* and *negative distance*, respectively, used by Vaclav Brezina in describing the standard deviation. The reference corpus values correspond to the concept of the mean.

The analysis of the use of first-person pronouns across four genres of texts on political economy, conducted for illustration, has revealed that the concordance and surplus-deficiency extraction contribute to the research results no less than the relative frequency data. Contrary to genre expectations, the essays were found to contain the fewest first-person singular pronouns; at the same time, the surplus of all forms of first-person pronouns in academic lectures was not surprising as, unlike the rest of the selected genres, this one belongs to spoken discourse. Still, one has to bear in mind that first-person pronouns are not the only linguistic features to express subjectivity. Moreover, as concordance has revealed, a considerable amount of the first-person singular pronouns in other genres were found in direct quotations. The surplus-deficiency analysis of separate forms of first-person pronouns allows for more precise conclusions on whether the role of the author in the text is active or passive.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the LEXECON team for initiating this interdisciplinary project, which, as early as its development and implementation stages, inspires new research topics and tackles various corpus linguistics issues. The authors are thanking the project coordinator, Prof. Marco Enrico Luigi Guidi, for giving them permission to use the project theme and materials while performing this research.

REFERENCES

Baker, P., Hardie, A. and McEnery, T. (2006) *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Biber, D. (1993) Representativeness in corpus design. *Literary and Linguistic Computing*, 6 (4): 243-259.

Biber, D., Johansson, S., Leech, G. N., Conrad, S. and Finegan, E. (2021) *Grammar of Spoken and Written English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Baumgarten, N., Du Bois, I. and House, J. (2012) Introduction. In N. Baumgarten, I. Du Bois and J. House (eds.) *Subjectivity in Language and in Discourse* (pp. 1-14). Bingley: Emerald Group Publishing Limited.

Bollmann, M. (2019) A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long and Short Papers), (pp. 3885-3898). Minneapolis, Minnesota: Association for Computational Linguistics.

Brezina, V. (2018) *Statistics for Corpus Linguistics*. Cambridge University Press.

Brunet, E. (2011) *Hyperbase. Manuel de référence*. Available from http://hyperbase.unice.fr/hyperbase/doc/manuel.pdf [Accessed on 7 September 2022].

Bullions, P. (1847) *The Principles of English Grammar*. New York: Pratt, Woodford & Co. Available from https://catalog.hathitrust.org/Record/100619837 [Accessed on 1 November 2022].

Chadbourne, R. M. (1983) A puzzling literary genre: Comparative views of the essay. *Comparative Literature Studies*, 20 (2): 133-153. Available from https://www.jstor.org/stable/40246392 [Accessed on 2 October 2021].

Dash, N. S. and Ramamoorthy, L. (2019) Corpus editing and text normalization. In *Utility and Application of Language Corpora* (pp. 35-56). Singapore: Springer.

Egbert, J. (2018) Corpus design and representativeness. In T. B. Sardinha and M. V. Pinto (eds.) *Multi-Dimensional Analysis: research methods and current issues* (pp. 27-42). London: Bloomsbury Academic.

Egbert, J. (2019) Usage-based theories of Construction Grammar: triangulating corpus linguistics and psycholinguistics. In J. Egbert and P. Baker (eds.) *Using Corpus Methods to Triangulate Linguistic Analysis*. New York: Routledge.

Egbert, J., Larsson, T. and Biber, D. (2020) *Doing Linguistics with Corpus*. Cambridge: Cambridge University Press.

Fina, A. (2009) Language and subjectivity. *Estudios de Lingüística Aplicada*, 27 (50): 117-176.

Gablasova, D., Brezina, V. and McEnery, T. (2019) The Trinity Lancaster Corpus Development, description and application. *International Journal of Learner Corpus Research*, 5 (2): 126-158.

Gatto, M. (2014) *Web as Corpus: theory and practice*. London: A&C Black. Available from https://books.google.lv/books?id=J4NnAgAAQBAJ&source=gbs_navlinks_s [Accessed on 23 August 2022].

Guidi, M. E. L., Tusset, G., Guccione, C., Lupetti, M., Carpi, E., Henrot Sostero, G., Musacchio, M. T., Simon, F., Flinz, C., Beghini, F., Bientinesi, F., Caldari, K., Cammalleri, C. M., Migliorelli, M., Morleo, M., Pagliai, L., Pomini, M., Quinci, C., Romeo, M., Rosati, F., Sclafani, M. D., Vaccarelli, F. and Vezzani., F. (2021) LEXECON. The international research network on the economics lexicon. The Economic Teacher: A transnational and diachronic study of treatises and textbooks of economics (18th to 20th century). Project: *Translations and the Circulation of Economic Ideas*. Available from https://www.shorturl.at/luyT2 [Accessed on 23 August 2022].

Hanks, P. (2012) The corpus revolution in lexicography. *International Journal of Lexicography*, 25 (4): 398-436.

House, J. (2012) Subjectivity in English lingua franca interactions. In N. Baumgarten, I. Du Bois and J. House (eds.) *Subjectivity in Language and in Discourse* (pp. 139-156). Bingley: Emerald Group Publishing Limited.

Kilgarriff, A. and Grefenstette, G. (2003) Introduction to a special issue on the web as corpus. *Computational Linguistics*, 29 (3): 333-347. Available from https://aclanthology.org/J03-3001.pdf [Accessed on 3 November 2022].

Kilgarriff, A. and Rychly, P. (2008) *Finding the Words Which Are Most X*. Available from https://www.sketchengine.eu/wp-content/uploads/2015/05/Finding_the_words_2008.pdf[Accessed on 7 September 2022].

Langacker, R. W. (2009) *Investigations in Cognitive Grammar*. Berlin: Mouton de Gruyter.

Malavska, V. (2016) Genre of an academic lecture. *International Journal on Language Literature and Culture in Education*, 3 (2): 56-84. Available from https://www.researchgate.net/publication/310815820_Genre_of_an_Academic_Lecture [Accessed on 5 May 2022].

McEnery, T. and Brooks, G. (2022) Building a written corpus: what are the basics? In A. O'Keeffe and M. J. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics,* 2nd ed. (n. p.). New York: Routledge. Available from https://www.taylorfrancis.com/chapters/edit/10.4324/9780367076399-2/building-corpus-key-considerations-randi-reppen [Accessed on 15 September 2022].

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: method, theory and practice*. New York: Cambridge University Press.

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T. and Xiao, R. (2005) *Character encoding in corpus construction.* In *Developing Linguistic Corpora: a guide to good practice*. Available from http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf [Accessed on 1 September 2022].

McEnery, T., Xiao, R. and Tono, R. (2006) Corpus-Based Language Studies: an advanced resource book. New York: Routledge. Available from https://www.jbe-platform.com/content/journals/10.1075/ijcl.11.4.09w [Accessed on 23 August 2022].

Pho, P. D. (2012) Authorial stance in research article abstracts and introductions from two disciplines. In N. Baumgarten, I. Du Bois and J. House (eds.) *Subjectivity in Language and in Discourse* (pp. 97-114). Bingley: Emerald Group Publishing Limited.

Reppen, R. (2022) Building a corpus: what are the key considerations? In A. O'Keeffe and M. J. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*, 2nd ed. (n. p.). New York: Routledge. Available from https://www.taylorfrancis.com/chapters/edit/10.4324/9780367076399-2/building-corpus-key-considerations-randi-reppen [Accessed on 15 September 2022].

Sampson, G. (2013) The empirical trend. *International Journal of Corpus Linguistics*, 18 (2): 281-289.

Sinclair, J. (2005) Corpus and text – basic principles. In M. Wynne (ed.) *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books. Available from: https://users.ox.ac.uk/~martinw/dlc/index.htm [Accessed on 6 November 2022].

Sinclair, J. (1991) *Corpus. Concordance, Collocation*. Oxford: Oxford University Press.

Tantucci, V. (2021) *Language and Social Minds: the semantics and pragmatics of intersubjectivity*. Cambridge University Press.

Vinčela, Z. (2017). Canadian dollar in the English language varieties: corpus-based study. *Baltic Journal of English Language, Literature and Culture*, 7: 161-171. https://doi.org/10.22364/BJELLC.07.2017.10

Weisser, M. (2016) *Practical Corpus Linguistics*. Hoboken, New Jersey: Wiley Blackwell.

TOOLS USED

Hyperbase 10. Available from http://ancilla.unice.fr/pages/logiciel/ [Accessed on 7 September 2022].

Sketch Engine. Available from https://www.sketchengine.eu/ [Accessed on 7 September 2022].

TEXTS ANALYSED

[A 1] Banfield, T. C. (1845) *Four Lectures on the Organization of Industry; Being Part of a Course Delivered in the University of Cambridge in Easter Term 1844*. London: R. and J. E. Taylor. Available from https://catalog.hathitrust.org/Record/001886951 [Accessed on 29 March 2022].

[A 2] Burton, J. H. (1849) *Political and Social Economy: its practical applications*. Edinburgh: William and Robert Chambers. Available from https://catalog.hathitrust.org/Record/006494040 [Accessed on 29 March 2022].

[A 3] Duncombe, C. (1841) *Duncombe's Free Banking: An Essay on Banking, Currency, Finance, Exchanges, and Political Economy*. Cleveland, OH: Sanford & Co. Available from https://catalog.hathitrust.org/Record/001739852 [Accessed on 29 March 2022].

[A 4] Gilbart, J. W. (1847) *Lectures on the History and Principles of Ancient Commerce*. London: Smith, Elder and Co. Available from https://catalog.hathitrust.org/Record/006512048 [Accessed on 29 March 2022].

[A 5] Hancock, W. N. (1847) *Three Lectures on the Questions: Should the Principles of Political Economy Be Disregarded at the Present Crisis?, And if Not, How Can They Be Applied Towards the Discovery of Measures of Relief?* Dublin: Hodges and Smith. Available from https://catalog.hathitrust.org/Record/011606170 [Accessed on 29 March 2022].

[A 6] Middleton, H. (1847) *Four Essays*. Philadelphia: King & Baird. Available from https://catalog.hathitrust.org/Record/102812246 [Accessed on 29 March 2022].

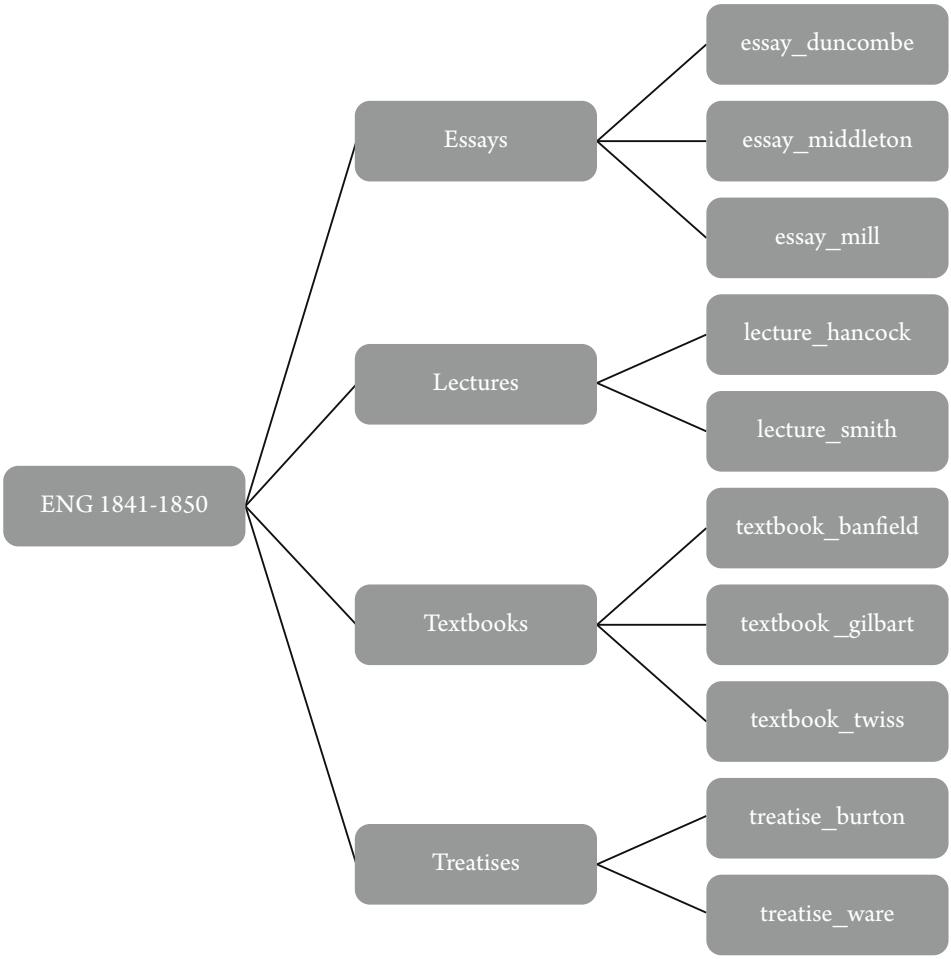
[A 7] Mill, J. S. (1844) *Essays on Some Unsettled Questions of Political Economy*. London: J. W. Parker. Available from https://catalog.hathitrust.org/Record/001308503 [Accessed on 29 March 2022].

[A 8] Smith, H. (1846) *A Lecture on the Capability of Great Britain and Ireland to Give Employment, and Provide a Sufficient Maintenance for the Whole Population; With Some Introductory Remarks on the Science of Political Economy*. Southampton: J. Tucker. Available from https://catalog.hathitrust.org/Record/102812240 [Accessed on 29 March 2022].

[A 9] Twiss, T. (1847) *View of the Progress of Political Economy in Europe since the Sixteenth Century*. London: Longman, Brown, Green, and Longmans. Available from https://catalog.hathitrust.org/Record/001307696 [Accessed on 29 March 2022].

[A 10] Ware, N. (1844) *Notes on Political Economy, as Applicable to the United States. By a Southern Planter*. New York: Leavitt, Trow, and Co. Available from https://catalog.hathitrust.org/Record/006126197 [Accessed on 29 March 2022].

APPENDIX 1 STRUCTURE OF THE SUBCORPUS



Kristīna Korneliusa is currently a Ph.D. student working at the University of Latvia. Her research interests include corpus linguistics, systemic functional linguistics, and stylistics.
https://orcid.org/0000-0001-5003-5445
Email: kristina.korneliusa@lu.lv

Zigrīda Vinčela (Dr. philol., Assoc. prof. in Applied Linguistics) is currently working at the University of Latvia. Her research interests include corpus-based and corpus-driven studies of written and spoken texts as well as some aspects of phonetics and phonology.
https://orcid.org/0000-0003-1930-0970
Email: zigrida.vincela@lu.lv