This is the accepted manuscript of the article, which has been published in Nurmi A., Rütten T. & Pahta P. (eds)
Challenging the Myth of Monolingual Corpora. Amsterdam: Brill. Language and Computers no. 80.
ISBN: 978-90-04-27668-0. ISSN: 0921-5034. https://doi.org/10.1163/9789004276697_012

Multilingualism and quotations from a corpus linguistic perspective: a case study of Samuel Taylor Coleridge's *Biographia Literaria*

Mark Kaunisto, University of Tampere

1. Introduction

The study of corpora and corpus data on the whole often presents linguists with a variety of practical challenges that need to be taken into account while examining the data. Representativeness in general is the key concept: how well does the material represent what it is supposed to represent? Problems in this regard can be perceived on different levels. From a broader perspective, one could ask, for example, how well does a particular set of texts represent a genre. Challenges of this type have often been observed in connection with corpus study, and an often quoted comment was made by Mukherjee (2004: 114), who said that "absolute representativeness is an unattainable aim." In addition to macro-level questions on representativeness, one could also pay attention to more specific items which contribute to representativeness. For instance, texts may include items such as proper nouns, quotations, and passages in other languages. Such items can be found in different texts in varying degrees, and one question that could be examined more closely is the degree of originality of texts in a corpus. This chapter raises the question of examining the significance of multilingualism and quotations in corpus data, and presents a close-up study of such elements in one book. This is regarded as noteworthy, as such passages contribute to the overall word count of the corpus while in many cases, strictly speaking, such items would not be reflective of the language choices of a particular author at a particular point in time.

The original impetus for the present study comes from examining corpora such as the *Corpus of Late Modern English Texts* (henceforth abbreviated as CLMET) and the *Corpus of Historical American English* (COHA), which are relatively recent examples of diachronic corpora consisting of several million words of texts. The 3.0 version of CLMET contains approximately 35 million words of British English texts published between 1710 and 1920, and COHA, representing written American English, includes over 400 million words of texts published between 1810 and 2009. These corpora have answered the demand for larger corpora, but an essential difference between these corpora and many of the earlier, smaller corpora is that they have largely been compiled by making use of already existing digitized editions of texts created for projects such as *Making of America* or *Project Gutenberg*. Examining concordance lines from a corpus of any size typically involves excluding irrelevant tokens, but searches on corpora of the magnitude of CLMET or COHA occasionally make one cautious about hits from particular texts. For example, edited volumes of letters may be problematic if one wishes to keep close track of the numbers of authors using specific words or expressions. The question of multiple voices or authorship may also come into question in cases where a text clearly has plenty of quotations from different sources, some of which may date even centuries prior to the corpus text in which they appear[1].

This paper takes a closer look into Samuel Taylor Coleridge's *Biographia Literaria*, which is included in the 3.0 version of the *Corpus of Late Modern English Texts*. The famous work on literary criticism, or "a literary biography", as Coleridge himself puts it in the very title of the book, was first published in 1817. Even superficial browsing through the text gives one an impression that it contains several passages in Latin, Greek, and other languages, as well as sometimes rather lengthy quotations of John Milton, William Shakespeare, William Wordsworth, and others. Such passages give rise to contemplating their possible significance when analyzing results gleaned from corpus data. As mentioned, quotations and passages of other languages than English contribute to the word count of the entire corpus, so it might be useful to examine exactly what is the amount of text that one might think is not entirely reflective of English as used by Coleridge himself around the time of the publication of the book. As regards quotations, how old were they in 1817, i.e., how much did Coleridge quote his contemporaries and how much did he quote older classics? What is furthermore of interest is the nature of the foreign language passages: are they original instances of Coleridge's own multilingualism and codeswitching or quotes of foreign authors (cf. Kohnen, this volume)? The results are also discussed from the perspective of corpus linguistics. Considering that literary criticism is very likely a genre characterized by a relatively high proportion of quotations, should corpus compilers then take precautions in connection with such works, and in what way?

For this paper, the entire text of *Biographia Literaria* was examined as regards the foreign language passages, quotations of other authors, quotations of Coleridge's own earlier texts, and passages of texts translated into English by Coleridge. The analysis was based on

[1] For a discussion on multilingual practices in CLMET, see Tyrkkö et al, this volume.

the text file included in CLMET 3.0, but to help in the identification of the sources of quoted passages, the annotated edition by James Engell and W. Jackson Bate, published in 1983, was consulted. To provide some background information on *Biographia Literaria* as well as CLMET 3.0, Section 2 first presents an overview of the book and its reputation as a classic work on literary criticism, followed by a concise description of the structure and compilation policies of CLMET 3.0. The methods and results of the analysis are presented in Section 3. It is observed that as much as 10 per cent of the entire word count of the book is actually text quoted from other sources and/or written in a language other than English. Quotations form a major portion of the 10 per cent, while foreign language items constitute about 2,600 words, that is, approximately 2.6 per cent of the entire word count of the book. The potential multilingualism and quotations may have in skewing corpus results is then studied by examining the usage of personal pronouns in *Biographia Literaria*, with a comparison between the numbers of the pronouns found in quotations and Coleridge's own text. It is noted that some pronouns are notably more common in quotations, and considering the total number of quoted text and foreign language items, it is possible that unless due caution is given on the special character of individual texts in a corpus, distortions in corpus results are possible. The final concluding chapter summarizes and discusses the main findings.

2. Background

Before the examination of the various elements in *Biographia Literaria* which may be regarded as not fully representing the author's own original output at the time of its composition, it is worth taking a brief look into the history, contents, and the overall character of the work itself. In a similar fashion, the inclusion of the text in the third version of CLMET warrants a glance at the rationale of the makeup of the corpus as well. As will be observed, the concerns seen in connection with *Biographia Literaria* are ones that the compilers of the corpus have generally taken notice of.

2.1 Coleridge's *Biographia Literaria*

The full title of Coleridge's book is *Biographia Literaria or Biographical Sketches of My Literary Life and Opinions*. Although the title suggests that the book is a literary autobiography, the book may justifiably be regarded as a peculiar mixture of some autobiographical elements, letters, and discussions on poetry and philosophy. Nevertheless, it is considered to be an important work on literary criticism. As Engell and Bate note in the introduction to their edition of Coleridge (1983), the book stems from Coleridge's original idea in 1815 to put together a collected volume of his poems, for which he began to write a preface "to discuss his general principles of poetry" (p. xlvi). However, over time the manuscript began to expand to include autobiographical as well as philosophical and metaphysical elements (Coleridge 1983: I, lii-liv). Partly because of illnesses and a decision to produce a two-volume book, in 1816 Coleridge found himself under pressure to finish the work and to make the second part match the first one in length. As a result, the book published in 1817 is generally considered to be incoherent and loosely structured (Modiano 2009: 206). The book consists of a so-called "philosophical part", in which Coleridge in several chapters discusses the ideas expressed by, for example, Spinoza, Kant, and Schelling, and the part analyzing the nature and character of poetry, in which Coleridge's analyses on Wordsworth's poems play a central role.

Among the literary critics even today, *Biographia Literaria* divides opinions. As noted by Modiano (2009: 204-205), it has been and is regarded as both challenging and rewarding to read, but perhaps the most critical comments adding to the controversial reputation of the book have had to do with Coleridge's plagiarism of Schelling, which was noticed rather soon after the publication of the work. Coleridge translated several passages from Schelling's works originally written in German, and although he often refers to Schelling in his text, there are so many unattributed sections lifted from Schelling that accusations of plagiarism were made.

Foreign languages are prominent on the pages of *Biographia Literaria*, and Coleridge placed great importance on the knowledge of them, as becomes explicitly evident in one of the several footnotes of the book (Coleridge 1983: 239 fn.):

> Were I asked what I deemed the greatest and most unmixed benefit, which a wealthy individual, or an association of wealthy individuals could bestow on their country and on mankind, I should not hesitate to answer, "a philosophical English dictionary; with the Greek, Latin, German, French, Spanish, and Italian synonymes, and with correspondent indexes."

Coleridge presents several quotations from works originally written in foreign languages, but instances of codeswitching by Coleridge can be attested as well. One example is found in Chapter 23 of the book (Coleridge 1983: II, 215; italics original):

> A similar and more powerful objection he would feel towards a set of figures which were *mere* abstractions, like those of Cipriani, and what have been called Greek forms and faces, i.e. outlines drawn according to a recipe. *These* again are not *ideal*; because in these the *other* element is in excess. "*Forma formans per formam formatam translucens*," is the definition and perfection of *ideal* art.

On the above sentence in Latin, presented as if it is a quote, Engell and Bate note that the words are very likely Coleridge's own, and that he "probably uses Latin to give the remark aphoristic strength or authority" (Coleridge 1983: II, 215, fn. 3).

Coleridge also presents many quotations from the works of other English authors, ranging from the Middle English author Chaucer to Coleridge's contemporaries. The quotes are of varying length, sometimes consisting of only a few words, while the longest continuous passage of another author's text is found in Chapter 19, where Coleridge quotes three stanzas by George Herbert, 766 words in total. As regards Coleridge's opinions on and references to other poets' works, however, *Biographia Literaria* is mostly known for its extensive critique of Wordsworth (Modiano 2009: 211-216), and the examination of Wordsworth's poetry is accompanied by a number of illustrations from his works. Engell and Bale even see a partial connection between the length and extensiveness of the critique of Wordsworth and the pressure felt by Coleridge to complete the second volume, saying that "[t]he discussion shows signs of 'padding', especially in the liberal use of quotation" (Coleridge 1983: I, lxii).

Coleridge's book contains some passages which have been translated from other languages, usually by Coleridge himself. In some cases the original foreign-language text is quoted and followed by the translation, either in the body text or in a footnote. From a corpus linguistic perspective, these parts of the book are also of interest for the purposes of the present paper. Typically translated texts are not included in general corpora representing a language, as they have been observed to differ from text written originally in a particular language in terms of simplicity, explicitness, and the collocations (see, e.g., Baker 1993, 1996; Mauranen 2000). Instead, in corpus linguistics translated texts are often examined separately and are used, for example, in parallel corpora.

In addition to foreign-language items, quotations, and translated passages, there are also extracts of texts that Coleridge himself had written before he had begun to write *Biographia Literaria*. As a basis to discuss his views on poetic diction, Coleridge presents examples of his own work across his literary career. Coleridge also included three revised letters (the “Satyrane’s Letters” between Chapters 22 and 23) which he had written about his trip to Germany and which had been previously published in his journal *The Friend* in 1809; according to Engell and Gale, this inclusion was also made for the purpose of completing the second volume of the book (Coleridge 1983: I, lxii-lxiii). It can thus be said that even as regards Coleridge’s own text in *Biographia Literaria*, there are elements of various types of text – poetry, letters, and essayistic expository writing – which, rather than representing his writings in 1815-1817, include extracts of Coleridge’s work from a period of some twenty years.

2.2 *The Corpus of Late Modern English Texts* (CLMET)

As regards the diachronic corpora of the English language, the compilation of the *Corpus of Late Modern English Texts* filled in a gap of a large corpus of British English fiction and non-fiction texts of the eighteenth and nineteenth centuries. The corpus consists of three parts representing the subperiods of 1710-1780, 1780-1850, and 1850-1920. After it was first introduced in 2005, the corpus has been developed further in later versions by adding more texts in order to bring the overall word counts of the three parts closer together as well as to achieve greater balance in terms of the different genres represented in the corpus.

The rationale and principles in the design and compilation of the corpus is discussed in De Smet (2005) and Diller, De Smet and Tyrkkö (2011). What is noteworthy in these articles is that the compilers acknowledge the problems that may be seen in the corpus texts:

> the exact bibliographical history of the corpus texts is often highly unclear. Internet sources tend to provide no specification as to which version of a text lies at the basis of its electronic edition, who the intermediate editors have been, and what they might have done to the original text. It is clear from occasional editorial footnotes and modernised spellings that the texts scanned in for electronic publication are often late 19th or early 20th century editions of earlier prints or manuscripts. (De Smet 2005: 79)

The question of unclear bibliographical information also happens to apply in the case of *Biographia Literaria*. Closer inspection of the text shows that the electronic edition created for Project Gutenberg is evidently not based on the original from 1817, as some of Coleridge’s footnotes in the CLMET 3.0 file of the text include references such as “April, 1825” (e.g. footnote 47). The file does not appear to have footnotes by external editors; the examination of Engell and Bate's annotated edition points out that Coleridge's additional notes with these references appeared in an edition published by his daughter in 1847. Although it is not possible to verify the exact edition which served as the basis for the electronic version, it seems that it contains elements of the 1847 edition. Observations of this kind, as well as feedback on the overall quality of the texts is exactly what the compilers indeed welcome from the corpus users, as Diller, De Smet, and Tyrkkö point out in the concluding chapter of their paper describing the 3.0 edition of the corpus:

> There will be gaps, and there will be poor texts, for it must be remembered that the growth in the quantity of online texts which we have seen in recent years has in some respects led to a loss in quality. The results of the Optical Character Recognition methods used in the production of many digitized texts are uneven, and though we will try to pick the best, we will not always succeed. Feedback from users will be essential. (Diller, De Smet & Tyrkkö 2011: 34-35).

One way in which the designers of corpora can alert the corpus users of texts which may be problematic or anomalous is to "flag" such texts in the corpus manual or other supplementary data provided. The CLMET 3.0 (as well as the recently updated version 3.1) includes an Excel file with details on the texts, including information on the word count, genre and subgenre, the year of publication, and the author's gender and year of birth. In addition, when considered necessary, the compilers provide further information on the genre classification as well as notes of other special features of some texts. For example, there are additional comments such as "contains some letters of earlier date" (on *The Life of Col. James Gardiner* by Philip Doddridge), "contains some Italian verse" (on *Stories from the Italian Poets, with Lives of the Writers* by Leigh Hunt and James Henry), and "contains considerable amount of quoted verse" (on *Crabbe* by Alfred Ainger) – i.e., exactly the kinds of elements observed in Coleridge's *Biographia Literaria*. However, the additional information on *Biographia Literaria* only sheds further light on the genre categorization with "[a]utobiography but containing few actual autobiographical elements; mostly meditative and essayistic in style, on philosophy and literary criticism". It is possible that compared to Coleridge's book, the other texts have proportionally more foreign language elements and quotations, so it is unclear whether *Biographia Literaria* would need to be flagged in the same way or not as it would require examining the other texts as well. The main purpose here is to observe the amount of such elements *Biographia Literaria* and then to contemplate their potential implications on search results.

3. The analysis of different elements in *Biographia Literaria*

To investigate the occurrence of passages of foreign language items, quotations, and translations from other languages into English in *Biographia Literaria*, the plain text file of the book included in CLMET 3.0 was examined manually, and relevant items were then copy-pasted into separate MS Word files according to the type of passage found. Identifying relevant passages is not an altogether straightforward process, and a number of difficulties were encountered. First, what exactly counts as a foreign language element is a problematic issue. A clear example would be the 82-word long German quotation of Goethe on the very first page of the book, but shorter items are trickier. The uses of single words were left out, especially as they were often used with English determiners, which is suggestive of words being incorporated into the system of English grammar and thus being used as loan words. For example, the use of the word *intermundium* in the following passage was not included as a foreign language item in the present study:

> While the former rest content between thought and reality, as it were in an intermundium of which their own living spirit supplies the *substance*, and their imagination the ever-varying *form*; the latter must impress their preconceptions on the world without, in order to present them back to their own view with the satisfying degree of clearness, distinctness, and individuality. (Coleridge 1983: I, 32)

Phrases can also be problematic, as in some instances one could argue that particularly some learned phrases of foreign origin have become regularly used expressions in English texts. A good case in point in this regard is the phrase *a priori*, which is also listed in the *Oxford English Dictionary* with a separate entry. Instances of this type were left out of the analysis, as well as metalinguistic references to foreign words and phrases[2], foreign language titles of treatises, poems, or plays.

[2] A typical example of a metalinguistic reference to a foreign language expression by Coleridge is “By ‘persuasa prudentia’, Grynaeus means self-complacent *common sense* as opposed to science and philosophic reason” (Coleridge 1983: I, 166; original italics)

Another problem faced involved the identification of quotations, as items that have the appearance of quotations are not necessarily actual quotations. This applies both to the use of quotation marks as well as passages separated from the body text and indented. Sometimes quotations marks are used when presenting fictional spoken utterances, and indentations can be used, for example, to give emphasis to key ideas. Furthermore, there are also sections in *Biographia Literaria* where the original source is not clearly specified in the text. In this respect, consulting the annotations in Engell and Bate's 1983 edition was very useful, as well as the identification of unacknowledged translated passages from the works of Schelling and others.

The word counts of different types of passages which can be regarded as not representing Coleridge's own original written English, are based on the occurrence of these items in the CLMET 3.0 version of the text. This is important to note, as a comparison of the 1983 edition (based on the original 1817 edition) and the CLMET version shows that the latter is somewhat shorter, with some paragraphs left out from the 1817 edition, while some footnotes have also been added. The word count of the entire book, which will be referred to in later sections of this paper, is 138,354, based on the plain text file of the book in CLMET 3.0[3].

3.1 The uses of languages other than English in *Biographia Literaria*

With the exception of the occurrence of foreign language items in loan words, established learned phrases, proper names, and metalinguistic references, the number of words of foreign language passages in Coleridge's *Biographia Literaria* amounts to 2,903, including elements in French, German, Greek, Italian, and Latin. The more detailed breakdown of the numbers of words of different languages, and the division of the passages into quotations of other people's works and instances of codeswitching by Coleridge, are presented in Table 1 below.

	French	German	Greek	Italian	Latin
quotations	0	226	454	542	1375

[3] The word count given for *Biographia Literaria* in the accompanying Excel file of CLMET 3.0 is slightly higher (140,784), and the difference may be due to the varying principles used by the different computer programs in the identification of items as words. Since the word counts of foreign language passages etc. in the present paper are based on those provided by the MS Word program, for the purposes of counting percentages of such passages in the whole of *Biographia Literaria*, the word count of the entire text will be based on the MS Word information.

codeswitching	25	14	6	0	261
total	25	240	460	542	1636

Table 1. The numbers of words in languages other than English in *Biographia Literaria.*

Latin is by far the most frequent of the foreign languages used in *Biographia Literaria*, constituting more than half of all the foreign language elements in the book. The individual quotations and instances of codeswitching in Latin are also the most numerous. The longest single Latin quotation is 263 words in length. The quotations and instances of codeswitching in the other languages are generally rather short, except for a quotation of Giovambatista Strozzi's madrigals in Italian (461 words). Being an enthusiast in reading texts in their original languages himself, Coleridge seems to trust the reader to appreciate foreign language extracts without an English gloss. In only about a dozen instances the foreign language passages were given an English translation, either immediately following the quotation, or in a footnote.

The instances of codeswitching from English into another language altogether add up to 306 words, and the words in foreign language quotations total as many as 2,597 words. Foreign language items thus represent 2.1 per cent of the total word count of the book. Alone these figures may admittedly appear rather inconsequential. However, to give a clearer idea of the potential significance of these numbers, it could be noted that if similar proportions of foreign language passages appeared across the corpus by extrapolation, the 35-million-word CLMET would have approximately 700,000 words of text in languages other than English. From the viewpoint of representing the actual language choices by the author in a corpus, the role of such items becomes important especially if we consider them in conjunction with quotations of other author's texts, as will be observed the sections below.

3.2 Quotations of other authors' works in English

Considering Coleridge's aims to discuss in detail his views on philosophy and poetic diction, it is understandable that quotations of other authors' texts are numerous in *Biographia Literaria*. They feature particularly heavily in the second volume (Chapters 14-24), and some chapters are rich with quotations of Shakespeare (Chapter 15), Wordsworth (Chapters 17, 20, and 22), and Maturin (Chapter 23). The book contains quotations of 38 British authors (11685 words), and some quotations of English texts whose authors are unknown (139 words).

Altogether quotations of these sources add up to 11,824 words, i.e., 8.5 per cent of the word count of the entire book. The full list of the authors quoted together with numbers of quoted words is presented in Appendix A.

Engell and Bate observe that "[t]he effective cause of the *Biographia* is Wordsworth's 1815 Preface [to *Poems*]" (Coleridge 1983: I, cxxxv), and it gave impetus to much of the critique that Coleridge expressed of Wordsworth's work. It is therefore hardly surprising that not only does Coleridge discuss Wordsworth's poetry extensively, but Wordsworth is also the most frequently quoted author, as becomes evident in Table 2 below, listing the most frequently quoted authors in *Biographia Literaria*:

author	words
Wordsworth, William (1770-1850)	5684
Herbert, George (1593-1633)	766
Shakespeare, William (1564-1616)	722
Maturin, Charles (1782-1824)	719
Shadwell, Thomas (1642-1692)	617
Milton, John (1608-1674)	486
Chaucer, Geoffrey (1343-1400)	358
Taylor, Jeremy (1613-1667)	284
Gray, Thomas (1716-1771)	186
Davenant, William (1606-1668)	185

Table 2. The ten most quoted authors (in English) in *Biographia Literaria.*

The prominence of the quotations of Wordsworth is striking – his words alone constitute 4.1 per cent of the entire book. The earliest of Wordsworth's texts that Coleridge quotes dates from 1789, and about half of the words (2,717) appear in texts written in or before the year 1800.

Of all the 38 British authors quoted, in addition to Wordsworth and other of his contemporaries, Coleridge also presents and discusses extracts from the works of older British classics. As many as 26 out of the 38 quoted writers were born before 1700, and the quotations of these writers include 4,849 words, i.e., 3.5 per cent of the total word count. It can therefore be noted that considering the proportion of foreign language elements and quotations, from a corpus linguistic perspective there may indeed be some concerns as regards *Biographia Literaria* which the corpus user should take into account, depending of course to some extent on the linguistic items which are examined. Not only are there several passages in the book which are not in English, many of the English passages are quotations of other authors' works, some of which date back centuries before the publication of Coleridge's book. Added up, the number of foreign language words and quoted items is 14,727, constituting 10.6 per cent of the whole book.

3.3 Coleridge's translations and his own earlier writings

As far as the variety of different elements in *Biographia Literaria* is concerned, it may also be worth observing the occurrence of passages translated from other languages into English by Coleridge himself as well as instances where Coleridge presents some of his earlier writings as bases of his discussions. One could argue that compared to foreign language passages and quotations of other authors, translations and Coleridge's own texts are not as problematic if we consider the relevancy and representativeness of possible search hits from *Biographia Literaria* in a corpus. These items clearly reflect Coleridge's own choices of using English words and phrases. However, as noted earlier, source languages may influence the use of constructions which would be rarer in non-translated English texts. Similarly, instead of trusting all search hits from *Biographia Literaria* to represent Coleridge's use of English expressions around 1815-1817, a corpus user may benefit from the knowledge that some extracts are of earlier date.

The total number of words in the passages translated by Coleridge into English is 1,466[4]. The numbers of words translated from different languages are presented in Table 3 below.

	German	Greek	Italian	Latin	**total**
words in translations	1095	296	24	51	**1466**

Table 3. The numbers of words in passages translated from foreign languages in *Biographia Literaria*.

Mostly the translations are from German, which largely also reflects the great role that the discussion on German philosophers plays in the book. As noted earlier, often the foreign

[4] In addition to these translated passages, *Biographia Literaria* also contains altogether 401 words of quotations from the Bible, and two quotations from the works of Bacon and Seward, and these two quotes are actually translations from Latin (19 words) and Greek (22 words), respectively.

language passages were not followed by an English translation, and this also becomes evident in the lack of correlation of the figures in Tables 1 and 3. For example, the book contains 240 words of German, but over one thousand words of English text translated from German. It must be pointed out that the figures in Table 3 include those cases where the translation is explicitly mentioned by Coleridge. According to Engell and Bate (Coleridge 1983: II, 253-254), there are also passages where Coleridge presents almost direct translations of German authors without acknowledgement or where the attribution is unclear, i.e., sections which gave rise to accusations of plagiarism. The word count of these passages according to Engell and Bate's calculation is 3,700[5], which brings the number of words in translated passages over 5,000.

Extracts of Coleridge's own earlier writings also feature prominently in the two volumes. The earliest of Coleridge's poems included in *Biographia Literaria* date from 1796, and the three "letters", sandwiched between Chapters 22 and 23, were ultimately based on personal letters written in 1798 and published in 1809. There are also some extracts from works (e.g. his verse play *Zapolya*) that Coleridge had written around the same time as *Biographia Literaria*. The earlier writings nevertheless represent distinctly different kinds of style of writing from the narrative in the main body of the text. Altogether, the number of words from Coleridge's previously published writings in *Biographia Literaria* is 18,253.

4. Implications of foreign language passages and quotations to corpus searches

Based on the observations made so far on the variety of different kinds of textual elements in *Biographia Literaria*, it is worth considering the implications that such items may have on corpus linguistic studies. It may be the case that the inclusion of search hits in analyses of linguistic items without due caution to the special features that individual texts may have can potentially lead to skewed results. In some corpora, attention has been paid to the occurrence of foreign language items or quotations in the corpus texts. For example, in the *Helsinki Corpus of English Texts*, lengthy passages written in languages other than English were either omitted from the texts altogether, or shorter instances of foreign languages were assigned with a special code. The codes allow the corpus user to perform searches which exclude quoted

[5] Engell and Bate also observe instances which can be regarded as intellectual plagiarism, but instead of involving direct translation, they are characterized by close or loose paraphrases or summarizing the ideas of German authors (Coleridge 1983: II, 253-254).

items from appearing in the query results (Kytö 1996: 30). Similar codes and tags are also used to identify quoted passages, for example, in the *British National Corpus.*

One practical example regarding the quotations in *Biographia Literaria* will be examined here. The possibility of quotations generally manifesting different kinds of linguistic features than the main text has been observed earlier; for example, in connection to his examination on the use of personal pronouns in different text types in Old, Middle and Early Modern English, Rissanen (1992) makes an interesting observation on their occurrence in sermons and homilies:

> For more accurate comparison [of the use of first and second person pronouns in homilies and sermons], the occurrences of the pronominal forms in quotations and in the preacher's own text should be separated. Naturally, second person pronouns occur much more frequently in quotations than first person pronouns. For one thing, Christ's advice and exhortations to his disciples and others are often quoted. (Rissanen 1992: 204, fn. 17)

A study of this kind can also be made on the most frequently occurring personal and possessive pronouns in *Biographia Literaria*, with the quotations and Coleridge's own text separated. The pronouns were selected on the basis of frequency lists created on the list of quotations and the entire texts with the AntConc concordancing program. The results are presented in Table 4 below, including both the absolute frequencies and the normalized frequencies (occurrences per 1,000 words) of the pronouns. The quotations here include those of other English authors, Biblical quotations, and the two quoted English translations from other languages made by other authors, with a total of 12,266 words. To calculate the normalized frequencies of the pronouns used by Coleridge, the word counts of the quotations and foreign language items were excluded from the word count of the entire book (i.e., Coleridge's own translations remained in the word count). The number of words in the passages considered to represent Coleridge's text is then 123,185.

pronoun	freq. in quotations	freq. in quotations/1,000 words	freq. in Coleridge's text	freq. in Coleridge's text/1,000 words
I	177	14.43	1463	11.88
me	49	4.00	289	2.35
my	78	6.36	598	4.85
you	48	3.99	216	1.36
your	23	1.88	74	0.60
he	129	10.52	606	4.92
him	52	4.24	175	1.42
his	100	8.15	880	7.14
she	31	2.53	68	0.55
her	65	5.30	71	0.57
they	58	4.73	318	2.58
them	22	1.79	222	1.80
their	55	4.48	402	3.26

Table 4. The frequencies of most commonly occurring personal and possessive pronouns in the quotations and Coleridge's own text in *Biographia Literaria*.

As can be seen in Table 4, pronouns usually occur with higher normalised frequencies in the quotations of other authors than in Coleridge's own text[6]. With some pronouns the difference between the normalized frequencies is fairly slight, but greater differences can be observed in connection with the pronouns *he* and *her*. In the case of the absolute frequencies of *her* in particular, it is interesting to note that the pronoun occurs in the quotations alone almost as many times as in the rest of the entire book. Similarly, almost every third instances of the pronoun *she* in the book appears in a quotation. Reasons for the differences in *Biographia Literaria* are just as logical as they are in the case observed by Rissanen above; in the largely theoretical discussions on philosophy, religion, and poetry, references to individual women (or people in general) are not very prominent, whereas they – hardly surprisingly – crop up fairly often in the poems and extracts of plays quoted by Coleridge.

If the normalised frequencies of the same pronouns were examined in the book with no adjustments with regard to the influence of quotations and foreign language items to the results, in the majority of cases, the differences would be minor (see Appendix B). However,

[6] It deserves to be noted that the frequencies of the pronouns in Coleridge's text also includes different types of writing, as it includes some of his own poems as well. However, the effect of the occurrences of pronouns in Coleridge's poems on the normalized frequencies are probably marginal.

with the pronouns *she* and *her* the normalised frequencies would appear more notably higher if their occurrences in quotations are included in the calculations.

The observation of these high-frequency items shows that it may be worth paying close attention to the internal variety of individual books included in large corpora. It is, of course, very likely that this kind of variety is not only characteristic to Coleridge's *Biographia Literaria*, but heavy uses of quotations and/or codeswitching are generally seen in texts on literary criticism. However, even with this genre one could expect differences as regards what kinds of authors are quoted and how much use is made of foreign language passages.

5. Concluding remarks

As has been observed, there are a number of characteristics in Coleridge's *Biographia Literaria* which may pose problems to a corpus user. Multilingualism, the prominence of quotations, translated passages as well as extracts of text of sometimes considerably earlier date may give rise to treating search hits from the book with special concern. Although the heterogeneous nature of the book may cause difficulties in terms of evaluating how well its contents represent the language uses of its time, this does not by any means disqualify the text from being included in a linguistic corpus. On the contrary, considering texts such as *Biographia Literaria* entirely undesirable for inclusion in a corpus and selecting only less problematic texts would be a drastic choice, or even a false one, from the part of the corpus compiler. Heterogeneity and structural complexity are also features of texts that deserve to be truthfully represented in a corpus, and such features can also be the focus of attention, for example, in the study of the evolving nature of genres such as literary criticism. The effects of "irrelevant items" resulting from multilingualism and quotations in corpus texts on the calculations of normalized frequencies could likewise be examined on a larger scale, along the lines of the case study presented here.

It is clear that potential outliers in corpora and the special characteristics of individual texts need to be identified and brought to the attention of corpus users. An important question then is how to efficiently alarm or caution the users of such problems. The role of corpus compilers is obviously a central one. On the other hand, interpreting corpus results is ultimately the responsibility of the corpus user. It is quite conceivable that even though corpus files are coded and tagged, and supplementary materials contain disclaimers on the texts (and the characteristics of Coleridge's text might be assigned with a disclaimer), corpus users do not necessarily remember to make use of them. To echo the warnings made by Rissanen (1989) of the different "fallacies" involved with corpus use, one can see that the developments in corpus design may lead to a paradoxical situation. The increased efforts by the compiler of a corpus in taking different kinds of practical problems into account may only lull the corpus user into a false sense of security, eventually resulting in hastily drawn conclusions on the data. Conducting more in-depth studies on the less-than-elegant characteristics of corpora might be helpful in the long run.

References

Primary sources

Coleridge, Samuel Taylor. 1983. *Biographia Literaria or Biographical Sketches of My Literary Life and Opinions* (in two volumes), ed. by James Engell and W. Jackson Bate. London: Routledge & Kegan Paul.

De Smet, Hendrik, Jukka Tyrkkö & Hans-Jürgen Diller. 2011. *Corpus of Late Modern English Texts (CLMET), version 3.0*. Available from https://perswww.kuleuven.be/~u0044428/clmet3_0.htm.

Secondary sources

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies: Implications and Applications". *Text and Technology: In Honour of John Sinclair*, ed. by Mona Baker, Gill Francis and Elena Tognini-Bonelli, 233–250. Amsterdam and Philadelphia: John Benjamins.

Baker, Mona. 1996. "Corpus-based Translation Studies: The Challenges that Lie Ahead". *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, ed. by Harold Somers, 175–186. Amsterdam and Philadelphia: John Benjamins.

De Smet, Hendrik. 2005. “A corpus of Late Modern English texts”. *ICAME Journal* 29: 69–82.

Diller, Hans-Jürgen, Hendrik De Smet & Jukka Tyrkkö. 2011. “A European database of descriptors of English electronic texts”. *The European English Messenger* 19, 21–35.

Kohnen, Thomas. 201?.

Kytö, Merja. 1996. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts.* Third edition. Helsinki: Department of English, University of Helsinki.

Mauranen, Anna. 2000. “Strange Strings in Translated Language: A Study on Corpora”. *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*, ed. by M. Olohan, 119–141. Manchester: St. Jerome Publishing.

Modiano, Raimonda. 2009. “Coleridge as Literary Critic: Biographia Literaria and Essays on the Principles of Genial Criticism”. *The Oxford Handbook of Samuel Taylor Coleridge*, ed. by Frederick Burwick, 204–234. Oxford/New York: Oxford University Press.

Mukherjee, Joybrato. 2004. “The state of the art in corpus linguistics: three book-length perspectives”. *English Language and Linguistics* 8(1): 103–119.

Rissanen, Matti. 1989. “Three problems connected with the use of diachronic corpora”. *ICAME Journal* 13: 16–19.

Rissanen, Matti. 1992. “The diachronic corpus as a window to the history of English”. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, ed. by Jan Svartvik, 185–205. Berlin and New York: Mouton de Gruyter.

Tyrkkö, Jukka, et al. 201?

Appendices

Appendix A. The authors and numbers of words quoted by Coleridge, arranged in alphabetical order.

Author	**words**
Akenside, Mark (1721-1770)	6
Bartram, William (1739-1823)	56
Brown, John (1715-1766)	7
Butler, Samuel (1612-1680)	68
Cartwright, William (1611-1643)	34
Chaucer, Geoffrey (1343-1400)	358
Congreve, William (1670-1729)	12
Cowley, Abraham (1618-1687)	181
Daniel, Samuel (1562-1619)	178
Davenant, William (1606-1668)	185
Davies, John (1569-1626)	63
Donne, John (1572-1631)	143
Drayton, Michael (1563-1631)	131
Dryden, John (1631-1700)	10
Fletcher, John (1579-1625)	38
Goldsmith, Oliver (1728-1774)	12
Gray, Thomas (1716-1771)	186
Harvey, Christopher (1597-1663)	68
Herbert, George (1593-1633)	766
Hooker, Richard (1554-1600)	179
Johnson, Samuel (1709-1784)	24
Lee, Nathaniel (c1653-1692)	21
Lipscomb, William (1754-1842)	4
Marvell, Andrew (1621-1678)	2
Maturin, Charles (1782-1824)	719
Milton, John (1608-1674)	486
More, Henry (1614-1687)	40
Otway, Thomas (1652-1685)	9
Pope, Alexander (1688-1744)	72
Shadwell, Thomas (1642-1692)	617
Shakespeare, William (1564-1616)	722
Smart, Christopher (1722-1771)	16
Southey, Robert (1774-1843)	118
Spenser, Edmund (1552-1599)	174
Taylor, Jeremy (1613-1667)	284
Tomkis, Thomas (c1580-1634)	8
Warton, Thomas (1728-1790)	4
Wordsworth, William (1770-1850)	5684

Appendix B. The frequencies of most commonly occurring personal and possessive pronouns in *Biographia Literaria* (the total word count of 138,354 words also including foreign language passages).

pronoun	freq. in whole book	freq. in whole book/1,000 words
I	1640	11.85
me	338	2.44
my	676	4.89
you	216	1.56
your	97	0.70
he	735	5.31
him	227	1.64
his	980	7.08
she	99	0.72
her	136	0.98
they	376	2.72
them	244	1.76
their	457	3.30