Using Very Large Corpora to Teach Modern English (1500–1945)

Erik Smitterberg (Uppsala University)

Abstract
This article describes the incorporation of a corpus-based research assignment in a 7.5-credit Master's-level module on Early and Late Modern English. The design of the module as a whole as well as of the research assignment is discussed, and it is shown how this design tallies with intended learning outcomes based on Bloom's revised taxonomy. I also suggest ways in which students with little previous experience of corpus-based research can be introduced to the use of very large corpora relatively quickly with the aid of, among other things, exercises and pre-recorded lectures. A key component of the research assignment concerns methodological desiderata such as ensuring recall and precision in corpus-based retrieval of historical features, operationalizing frequency appropriately, taking into account the influence of the genre parameter, being mindful of the limitations of the corpus used, and citing and evaluating secondary sources. Students learn about the value of these desiderata largely through data-driven learning before and during their work on the assignment; examples of how they have been addressed in individual papers are provided. Finally, the value of including empirical, corpus-based components in a historical course is discussed.

Keywords: corpus linguistics; English historical linguistics; Early Modern English; Late Modern English; data-driven learning

1. Introduction

Uppsala University's Master's programme in English offers a specialization in English Linguistics. This specialization currently contains two course modules of 7.5 credits (Sw. *högskolepoäng*)[1] each that

[1] In Swedish tertiary education, one year of full-time studies (which is divided into two terms) corresponds to 60 *högskolepoäng*, which means that *högskolepoäng* are comparable to ECTS credits in this regard. A module worth 7.5 *högskolepoäng* normally corresponds to c. five weeks of full-time work; however, the most common set-up in the Swedish system is probably that two

Smitterberg, Erik. 2023. 'Using Very Large Corpora to Teach Modern English (1500–1945).' *Nordic Journal of English Studies* 22(1): 93–112.

are devoted to the history of the English language: English in Transition I, which covers Old and Middle English, and English in Transition II, which focuses on Early and Late Modern English. As part of the latter module, students get an opportunity to carry out a small-scale research assignment resulting in a term paper that accounts for 45% of their total grade. In this article, I will discuss the use of very large corpora as the basis for these assignments, specifically the diachronic corpora available at english-corpora.org, such as Early English Books Online (EEBO) and the Corpus of Historical American English (COHA).[2] I aim to demonstrate how students can reach some of the learning outcomes for the module, which are based on Bloom's revised taxonomy, through analyses of these corpora.

The paper is structured as follows. In section 2, I describe the English in Transition II module, including learning outcomes, and how students are gradually prepared for the final research assignment.[3] Section 3 is

such modules run simultaneously for ten weeks during each half of term, so that students study each module at 50% of full time for ten weeks.

[2] The threshold for what counts as a 'very large' corpus of historical English is inevitably somewhat subjective. For the purposes of the present paper, I draw the line at c. 50 million words. Davies (n.d.), who compiled many of the corpora at english-corpora.org, compares the Corpus of Online Registers of English, which contains almost 53 million words, with 'other very large corpora', which indicates that 50 million words may be a suitable cut-off point. I return to the value of corpus size for student papers in sections 2.2 and 4. Drawing the line at 50 million words also sets very large corpora apart from other sizeable—and very valuable—historical corpora such as the Old Bailey Corpus (OBC) and the Corpus of Late Modern English Texts (CLMETEV), which contain 14 and 15 million words, respectively (the Corpus Resource Database, n.d.). In addition, the very large corpora available at english-corpora.org use very similar interfaces and search engines, which makes them suitable for discussion in a single paper. Uppsala University also has an institutional licence for these corpora, which means that students can log on as affiliated with the university and get unrestricted access to them.

[3] A module like English in Transition II may be taught by different teachers in different years and may also be co-taught by several teachers. The account given in this paper concerns mainly the years in which I have been the sole teacher on the module, i.e., 2017–2019 and 2021. The module has of course changed somewhat over time in response to, among other things, student evaluations; the account given here focuses on the most recent version of the module.

devoted to the topics that are taken up in teaching in conjunction with students' work on their research assignments, such as recall vs. precision in retrieval and how to operationalize frequency in historical linguistics. Section 4 provides a concluding discussion of the value of giving students at the advanced level experience of carrying out their own research on the history of English. In sections 3 and 4, I illustrate some of the points made with examples from students' assignments.

2. English in Transition II

2.1 General description and learning outcomes

The two historical modules English in Transition I and II (see section 1) are given in the second term of the Master's programme. They run consecutively; students thus work their way from Old to Late Modern English throughout the term. However, the modules differ in structure: while English in Transition I is a survey course on the grammar, pronunciation, and vocabulary of Old and Middle English that ends with a final written exam, English in Transition II includes more research-oriented activities and is examined through class presentations, the submission of short assignments, and the final research assignment that is in focus in this paper. The reason for the difference in structure is twofold. First, owing to time limitations, students cannot be expected to reach the stage where they are able to do independent research on Old or Middle English. Secondly, students who may wish to write their MA theses on historical topics are far more likely to do so on Early or Late Modern English, as these periods are where the Department of English's historical research expertise lies, and as Modern English is easier for students to master than Old and Middle English are.

One prerequisite for completing a written research assignment at the end of the module is of course that students have sufficient background knowledge of the language of the period when they start working on their topics. In addition, general knowledge of Early and Late Modern English is of course an end in itself as well as a means to an end in the form of a research project. These considerations are reflected in the learning outcomes in the syllabus. The formulation of these outcomes was informed by Anderson and Krathwol's (2001) description of the cognitive-process

dimension in Bloom's revised taxonomy.[4] Anderson and Krathwol (2001: 67–68) identify six categories of cognitive processes used to reach educational objectives: *remember*, *understand*, *apply*, *analyse*, *evaluate*, and *create*. These processes can be used as a guide when formulating objectives that students should reach in order to complete a module successfully. Suggestions for how to construct concrete learning outcomes that belong to each of these categories typically focus on what verbs are used to formulate the outcomes; for this module, a list made available by the University of Toronto (n.d.) was consulted.

Students who successfully complete the module should be able to:

1. describe important features of Modern English phonology, lexis, orthography, morphology, and syntax;
2. describe important linguistic features that characterize Early Modern English compared with Late Modern English;
3. describe important linguistic features that characterize Late Modern English compared with Present-day English;
4. apply [their] knowledge of changes in Modern English phonology to the pronunciation of individual words;
5. categorize lexical innovation in Modern English etymologically with the aid of dictionaries;
6. discuss features of Modern English in fluent and correct English, both orally and in writing;
7. compare and evaluate different descriptions of Modern English in secondary sources;
8. analyse Modern English text phonologically, morphologically, and syntactically; and
9. construct a small empirical study of Modern English phonology, morphology, or syntax using appropriate methodology.

As can be seen from the list, outcomes (1)–(3), with a verb from the *remember* category, concern core linguistic content on Modern English.

[4] Anderson and Krathwol (2001)'s model is two-dimensional: in addition to the six different cognitive processes on one dimension, four major types of knowledge—factual, conceptual, procedural, and metacognitive—are identified, and a learning objective can thus be placed in a two-dimensional grid comprising 6 × 4 cells (see Anderson and Krathwol 2001: 92). When the learning outcomes for English in Transition II were constructed, differentiating between the four types of knowledge was deemed less important; I thus focus on cognitive processes in this paper.

Outcomes (4)–(5), with verbs from the *apply* and *analyse* categories, respectively, also mainly target core content but are more demanding in that students need to make active use of the knowledge of phonology and lexis they acquired in meeting outcomes (1)–(3). Parts of outcomes (6) and (8), with verbs from the *create* and *analyse* categories, respectively, also concern core content, although the final research assignment draws directly on these outcomes. The core content is examined through exercises, assignments, and class presentations throughout the module.

The basis for students' work towards these core-content outcomes is one textbook for each period covered—currently Beal (2004) and Nevalainen (2006)—and additional material that is supplied electronically, such as PowerPoint slideshows, handouts, and pre-recorded lectures. The overall structure of the module is given in Table 1, in which 'EModE' stands for Early Modern English, and 'LModE' for Late Modern English.

Table 1. Class structure for English in Transition II

Class	Reading	Topic
1	Nevalainen (2006): chs. 1–2 Beal (2004): ch. 1	Course Introduction
2	Nevalainen (2006): chs. 6–8	EModE Grammar
3	Beal (2004): chs. 4–5	LModE Grammar Introduction to Research I
4	Nevalainen (2006): chs. 3–5	EModE Spelling, Lexis, and Word-Formation
5	Beal (2004): chs. 2–3	LModE Lexis and Word-Formation Introduction to Research II
6	Nevalainen (2006): ch. 9 Beal (2004): chs. 6–7	EModE and LModE Phonology Class Presentation I
7	—	Introduction to Research III
8	Nevalainen (2006): ch. 10 Beal (2004): ch. 8	Class Presentation II Course Wrap-Up

As Table 1 shows, in terms of linguistic fields, the module first focuses on grammar, followed by lexis and pronunciation. The order of presentation is designed to facilitate students' work on their final written assignments. Most assignment topics so far have concerned grammar, so it is important to provide students with background knowledge on Modern English morphology and syntax at an early stage in the module.

The final research assignment, on which I focus in this paper, of course draws on this background knowledge, but also specifically targets outcomes (6)–(9), with verbs from the *analyse* (*analyse*), *evaluate* (*compare*, *evaluate*), and *create* (*construct*, *discuss*) categories of the taxonomy. The three topic areas labelled 'Introduction to Research I–III' in Table 1 are group sessions that focus on providing students with the tools they need to complete the final research assignment; they are described in more detail in section 3.

2.2 Preparing students for a research component

In this paper, I focus on projects that make use of a very large corpus to investigate a topic on grammar, which make up the majority of all final research assignments completed within the framework of the module since its current version was launched in 2017.[5] Students are encouraged to come up with their own topics for investigation based on what has interested them in the background reading. Student-initiated topics that have been investigated include the choice between the prepositions *upon* and *on*, the *get*-passive and other constructions with the verb *get*, *his* vs. *its* as a possessive determiner with inanimate reference, the variation between *my* and *mine* and between *thy* and *thine* as determiners before vowel sounds, and the distribution of relative markers.

However, not all students are able to identify—or interested in identifying—a topic themselves. To reduce the risk that students fail to complete the module because they take too long deciding on a topic, I provide a list of possible topics that can be chosen, such as the decline of *thou* forms in the second person singular, *do*-support in affirmative and/or non-affirmative clauses, the rise of permissive meanings of *can*, the distribution of different expressions of futurity, nouns as premodifiers in noun phrases, and *not*-contraction. These topics cover both Early and Late Modern English, so that students still have a choice of period. Owing to

[5] Students are welcome to choose topics within historical phonology (and its orthographical representation) as well; it is also possible not to work with a corpus at all or to make use of a smaller corpus, such as the Corpus of English Dialogues (CED). Students who wish to focus on topic areas outside grammar and/or on varieties that are not covered by the available corpora are given support as regards reading up on the relevant research area in advance, locating possible sources of data, etc.

the size of the corpora (and, especially for Late Modern English, the fact that several different corpora are available), several students can write on what is more or less the same topic and still work on data that have not been analysed by other students, since different decades, genres, etc. can be selected for each project. This advantage of working with very large corpora can be of considerable help to students and teachers alike. Teachers can keep the number of separate topics manageable by suggesting similar topics to several students (which also facilitates supervision), and students working on similar topics can even draw on one another for help.

Given the contents of the module, it is important that the final research assignment be an opportunity for students to hone their skills not only as historical linguists, but also as empirical linguists, most of whom use corpora. One challenge in this regard is that the students who take the module come from a wide variety of backgrounds; while some of them have gone through Uppsala's undergraduate programme, Master's or doctoral students from other language departments, exchange students, and/or international Master's students are usually present as well. This diversity of the student body is very rewarding and increases the quality of classes, as discussions about the history of English can be greatly enriched by comparison with other languages—Indo-European as well as non-Indo-European—which students have acquired or studied or which they speak natively. However, it also means that students' degree of familiarity with corpora and corpus linguistics varies greatly. A few of the exercises and assignments that students complete while they engage with the two textbooks therefore introduce them to the use of historical corpora at an early stage in the module; this facilitates their reaching learning outcome (9) through the final research assignment. I will provide two examples here.

For one of the exercises on grammar, students are asked to analyse 100 randomly selected tokens of *if* from one of two decades in the Corpus of US Supreme Court Opinions (SCOTUS). Since some students are typically new to corpus linguistics, I supply them with randomly selected tokens in Excel worksheets for this exercise, but also demonstrate how the tokens were retrieved from the corpus and imported into Excel. The students' task is to determine whether there have been any changes in the distribution of modality marking over time (the two decades selected are separated by c. 100 years). To do so, they need to (i) remove irrelevant

instances from the data, e.g. tokens where *if* introduces a nominal clause or where the indicative and the subjunctive are not morphologically distinct, and (ii) classify the verb phrases in the relevant adverbial clauses into three categories: indicative verb phrases, subjunctive verb phrases, and verb phrases incorporating modal auxiliaries. They are referred to Grund and Walker (2006) for information on this variant field. Several students are typically given the same decade and concordance. After the exercise has been completed individually between two classes, they are divided into groups in class so that all members of one group have worked with the same concordance. They are then asked to compare their results and discuss tokens that they have analysed differently in order to reach consensus. After consensus has been reached, the two decades are compared, and students typically see a clear diachronic shift away from the subjunctive. In addition to illustrating the decline of the subjunctive in adverbial clauses, this exercise implicitly familiarizes students with concordances and teaches them about manual post-processing of concordancer output to improve precision and about how to calculate relative frequencies (percentages) in what is basically a variationist set-up (see Biber et al. 2016). This data-driven exercise thus uses inductive techniques to teach key methodological elements of corpus linguistics and targets outcomes (1), (3), and (8).

Students are given a further opportunity to use corpora in a small-scale assignment in historical semantics. This time, they need to access COHA themselves and retrieve 50 randomly selected tokens of a given word from a certain decade. The words that have been used for this assignment (e.g. *sophisticated*) must (i) be relatively frequent and (ii) have gone through some semantic change during the period 1820–1945, such as the development of new senses or quantitative shifts in the relative frequency of different senses. Students should then attempt to assign each of their 50 tokens to one of the senses listed for the word in the *Oxford English Dictionary* (*OED*). Each student has a unique combination of word and decade, which makes it possible to see whether the semantic change suggested by *OED* attestations and/or background literature is mirrored by the quantitative output when different decades are compared. The assignment also implicitly teaches students to consider genre differences (one of the prompts asks them to check whether one or several senses

occur predominantly in one of the genres in COHA)[6] and to resolve problems with indeterminate cases (since tokens are often ambiguous between two or more senses). In addition, students have to supply at least two numbered corpus examples illustrating the senses that they have identified as part of the text they submit; this requirement ensures that students know how to present, refer to, and discuss corpus material. Taken together, this assignment thus addresses outcomes (3), (6), and (9).

3. The final research assignment

The expected outcome of the final research assignment is a piece of independent, empirical linguistic research accounted for in an IMRaD-style paper with a length of c. 3,500 words (excluding references, appendices, etc.). Teaching intended specifically for the final research assignment typically consists of group sessions, individual supervision, and pre-recorded lectures.[7] I will account for the typical progression of teaching below.

In order to be able to make use of the corpora at english-corpora.org, students begin by watching a brief pre-recorded lecture that shows them how to access the website, where they can register and log in, and how to carry out a simple corpus search with and without part-of-speech tags. (This pre-recorded lecture is made available by way of preparation for the assignment on historical semantics discussed in section 2.2.) Students are encouraged to carry out a few searches of their own to familiarize themselves with the corpus interface. Because the interface for these corpora is based on the same architecture, and because the texts have been tagged and lemmatized in similar ways (Davies 2019: 321–322), skills acquired through exploring one corpus can typically be applied to other corpora as well.

In the first group session, which takes place at an early stage in the course (before students have chosen a topic), I go through matters such as

[6] The next time the module is offered, in the spring of 2023, I plan to ask each student to analyse output from one particular genre/decade subsample. This set-up will make genre differences even clearer when different students' results are compared.

[7] English in Transition II is typically offered as a campus course. During the Covid-19 pandemic, however, group sessions as well as individual supervision were conducted over Zoom.

what a corpus is, what counts as material and data in corpus linguistics, how to formulate a research question and a hypothesis, how an IMRaD paper in linguistics is typically structured, and how to format in-text citations and reference-list entries. Students are also given access to electronic material designed for the module as well as references to works such as McEnery et al. (2006), Smitterberg (2016), and *The Chicago Manual of Style*, where several of these matters are discussed in more detail. This session thus targets outcome (9) while also providing the bibliographical tools necessary for reaching outcome (7).

When students contact me with a topic suggestion or to indicate that they need help selecting a topic, they receive individual supervision in person or over e-mail. As up to roughly a dozen students may take the module simultaneously, time for individual supervision of projects is limited. However, each student receives assistance with selecting a topic, narrowing that topic down to a suitable scope, and choosing time periods to focus on. Some students also need additional support as regards managing the corpus interface, especially if they wish to make use of available corpus annotation such as part-of-speech tagging. Finally, when necessary, I help students to locate relevant secondary material. Students are required to use at least two secondary sources in addition to the textbooks, to ensure that they can reach outcome (7).

The second group session is devoted to corpus-linguistic method: it addresses the concept of *validity*, operationalized as high *recall* (no false negatives) and high *precision* (no false positives). I demonstrate how, in corpus linguistics, the corpus searches themselves are typically designed to maximize recall, while manual post-editing of the output is often necessary to ensure satisfactory precision. These concepts are linked to the corpus-based exercise on modality in adverbial clauses that the students will already have completed (see section 2.2). Searching for the form *if* ensures high recall for Late Modern English (as long as the scope of the analysis is limited to adverbial clauses introduced by *if*); however, as the students will already have noticed themselves, the resulting concordance must be post-edited manually to remove irrelevant tokens and thus increase precision. If time allows, I also use passive voice as an example. For this feature, searching for a passive (semi-)auxiliary—*be* or *get*—followed by a past participle (allowing for intervening words) will typically yield high recall for a reliably tagged corpus; however, manual post-editing of the output is required to remove potential tokens where the

participle is adjectival rather than verbal. This example also illustrates how the researcher's decisions can be informed by previous work, e.g. Quirk et al.'s (1985: §§3.74–78) passive gradient and its practical application in Schwarz (2017).[8] Finally, the group session addresses the concept of *reproducibility* as a cornerstone of scientific method and the importance of writing the Method section of an IMRaD paper so that reproducibility is attained. Outcome (9) is thus in focus throughout the session.

Depending on topic choice, some supervision in person or over e-mail may be necessary to adapt the general information given in the second group session to an individual student's project; for example, a student working on the decline of *which* as a relative marker with animate antecedents may need advice on how to distinguish interrogative from relative *which*, while a student who focuses on *on* vs. *upon* may need criteria for eliminating adverbial tokens of *on*. Yet other projects may not require much manual post-processing at all; however, improving precision manually is such an important general—and partly transferable—skill in corpus linguistics that all students should be exposed to the problem to ensure that, in the future, they can carry out independent research on topics where addressing it is crucial.

Once students have begun working on their projects, they are given access to a detailed pre-recorded lecture devoted to the concept of *frequency* in corpus linguistics. The lecture goes through the two main ways of operationalizing frequency in texts, namely the *variationist* perspective, where the frequencies of *variants* are compared with one another and proportions of occurrence calculated, and what Biber et al. (2016) refer to as the *text-linguistic* perspective, where the frequency of a linguistic feature is *normalized* to make it independent of text length (e.g. tokens per 1,000 words). Important constraints on studies using the two approaches are also treated, e.g., the requirement that variants should be ways of saying 'the same thing' (see, for instance, Tagliamonte 2012: 2). The lecture also covers what an independent variable is and how to present frequency results in tables (exemplified with variationist and text-linguistic calculations). The lecture is pre-recorded so that students can

[8] The next time this course is offered, in the spring of 2023, I plan to incorporate a separate exercise on this topic so that students get an opportunity to work with corpus output: they will analyse historical language data comprising passive as well as non-passive combinations of *be*/*get* and a past participle in order to improve precision by removing non-passive combinations.

watch it as many times as they like and, to some extent, copy the steps of the different calculations when they apply a framework to their own data-driven project.

The pre-recorded lecture also addresses an important methodological aspect of working with very large corpora. As the scope of these projects is necessarily very limited, the number of tokens returned by the corpus searches is frequently so large that students cannot analyse all of them within the framework of a 7.5-credit module (for instance, a search for the form *which* in COHA yielded 95,251 hits from the 1910s alone in a search carried out in July 2021). To make projects manageable, students learn how to work with *random samples* of corpus output. The total raw frequency of a feature in part of a corpus is then estimated based on a random sample of, say, 500 tokens (students may need to work with even smaller random samples and hedge the reliability of their results accordingly). To take the (invented) example from the lecture, if there are 1,345 tokens of *get* + past participle in a decade sample, and if 164 out of 500 tokens randomly selected from those 1,345 are genuine *get*-passives, the estimated raw frequency of *get*-passives in the decade sample is (164 / 500) × 1,345, i.e., c. 441 (see Schwarz 2017 for genuine examples of this calculation). Depending on the nature of the project, this estimated raw frequency can then be used as input to variationist or text-linguistic frequency measures.

One student made use of this method when analysing the variation between *his* and *its* as possessive pronouns with inanimate antecedents in Early Modern English, using the EEBO corpus. After criteria for when these two forms were interchangeable had been established, random samples of 100 tokens of *his*, *its*, and *it's*[9] from each decade in focus were analysed to estimate total raw frequencies. For instance, the student found that, as there were 3,190 tokens of *its* in one decade sample, and 89 per cent of the tokens in their random sample of *its* from that decade were valid, the estimated total frequency of valid tokens was 2,839 (0.89 × 3,190). Again, limited individual supervision over e-mail or in person is often required in order for students to be able to apply these concepts and techniques to their own work.

[9] The use of an apostrophe for the contraction meaning 'it is' or 'it has' but not for the possessive pronoun was not fully established in the seventeenth-century texts included in the student's analysis.

Some topics lend themselves to neither variationist nor text-linguistic set-ups, and in such cases it is necessary to make students aware of the special characteristics of their analysis. For instance, one student who analysed the rise of permissive *can* in Late Modern English compared the frequency of permissive meanings of *can* with the frequency of other meanings (e.g., ability and possibility) of the same modal auxiliary in a set-up that is reminiscent of variationist calculations from a purely statistical perspective. However, their framework was non-variationist by default: if semantic distinctions are used to separate categories, these categories are by definition not ways of saying 'the same thing'. In this respect, the student's approach was more akin to *form-to-function mapping* within historical pragmatics (see Jacobs and Jucker 1995).

Students present their preliminary work to the teacher and to their peers during the last class for the module, typically between one and two weeks before the deadline for submitting the final research assignment. For this class, they prepare a brief presentation on their project, which should contain an account of the linguistic background, research questions (and possible hypotheses regarding those questions), the primary material, and methodological aspects such as how data were selected and frequencies calculated. In addition to providing an opportunity for feedback from the teacher as well as their peers, these presentations give students an opportunity to learn from other corpus-based projects that may entail different challenges compared with their own work. It thus broadens their experience of empirical methodology.

4. Critical discussion

In this section, I will take up some advantages of working with a corpus-based final research assignment in order to reach some of the outcomes for English in Transition II. I will also address some challenges that have become evident over the past few years.

The most important advantage of giving students data-driven experience of working with very large corpora is clearly the obvious connection between teaching and research that is established. Students get first-hand experience of tools that are used regularly by researchers, which makes their connection to the world of research far more tangible than if they only read about other scholars' work. Reading advanced secondary sources also takes on a new relevance, as they are now doing so for the purpose of furthering their own research, not merely to master the content

for its own sake. As regards the history of English, corpus-based projects are very much a form of 'learning by doing' while facilitating reaching outcomes (6)–(9).

The use of very large corpora for these projects has several advantages. To begin with, getting access to sufficient data is rarely a problem; even low-frequency features can be analysed with a high degree of reliability (Davies 2012: 162). Secondly, several students can work on the same linguistic feature(s) during the same term (or consecutive terms). As mentioned in section 2.2, different corpora and decade samples can be used by each student. Moreover, random samples culled from these corpora often yield sufficient data for individual projects, and even if students should start out from the same list of thousands of tokens, they will in practice analyse different sets of data once these tokens have been individually narrowed down to, say, 200 randomly selected ones. There is thus little risk that students will simply be repeating analyses which have already been carried out (and which could then potentially be plagiarized).

One extralinguistic parameter that is often in focus in student projects is *genre*. This focus tallies with clear trends in published research; as Nevalainen and Raumolin-Brunberg (2017: 2) note, genre differences have been the predominant topic of studies of historical variation in English for the last 30 years. Moreover, linguistic genre differentiation can be used to shed light on past speech (see, for example, Culpeper and Kytö 2010). For instance, a term paper on the incidence of *not*-contraction in 1940s samples of fiction from COHA and the Movie Corpus revealed higher frequencies as well as proportions of contraction in the latter, indicating that contraction ratios were most likely even higher in contemporary informal, unscripted speech (see Rissanen 1986: 98 for such extrapolations).

Some corpora that are not stratified according to genre can instead raise awareness of the limitations of the results reached, and it is important that students learn to hedge their results in this regard, when necessary. If a corpus such as EEBO, which conflates several genres, is used, the lack of control of the genre parameter is a limitation that must be acknowledged. In one term paper, the student noted that the decrease in the normalized frequency of *does*/*doth* in affirmative questions between their EEBO samples from the early and late seventeenth century, which was unexpected against the background of Ellegård's (1953) study, may be affected by the genres covered by the corpus during the period

investigated. Results based on a single genre can instead raise questions about generalizability to other text categories; for instance, another student pointed out that it would have been valuable to complement their results on *not*-contraction in Late Modern English general fiction from COHA with other genres, such as scientific writing and letters, but also with specific types of fiction texts (e.g., romance novels). As student projects are limited in scope, the aim is to promote awareness of the limitations of a study in this fashion rather than requiring more comprehensive genre coverage.

Having students reach outcomes (7) and (9), as well as parts of outcomes (6) and (8), through a final research assignment is also advantageous for reasons external to the module itself. From the perspective of English historical linguistics, it is hoped that including an empirical research component in the module makes it more likely that students who continue their studies towards a Master's degree in English linguistics will choose a historical topic for their thesis. In the long run, this is also likely to encourage some students to specialize in the history of the English language at the doctoral level. At present, international scholarly interest in English historical linguistics is clearly increasing, but the time allocated to the subject in Swedish undergraduate and advanced-level curricula is limited. Given this limitation, the history of English is frequently introduced mainly as facts to be learnt, and there is rarely time for students to use the knowledge they have gained from such introductions in analyses of their own; this restriction complicates addressing some of the more complex processes—*analyse*, *evaluate*, and *create*—in Bloom's revised taxonomy. However, it is important not to ignore the processes that go beyond retention of content if students are to be able to become future researchers, because these processes focus on *transfer*: outcomes related to these processes give students the future-oriented ability to use what has been learnt in new contexts (Anderson and Krathwol 2001: 63–64). It is thus hoped that modules such as English in Transition II can contribute to sustained research interest in historical perspectives on English linguistics in Sweden.

Finally, several of the methodological questions facing scholars become apparent to students only when they engage in their own research. Some of these challenges—e.g., selecting suitable primary material, ensuring high recall and precision, and using the most appropriate frequency measure—are common to many corpus-based projects.

However, there are also difficulties that are especially characteristic of historical corpus linguistics, such as the lack of a standardized spelling, the decreased reliability of taggers when applied to historical texts, and the question of *representativity*; for instance, what parts of a historical English-speaking population does a corpus represent when only a minority of speakers were literate and literacy was stratified according to socio-economic status as well as gender? Historical corpora can stimulate students to engage with these questions in a way that fosters critical thinking about their own and others' work.

Nevertheless, including data-driven learning in the form of a corpus-based research component in a course on Early and Late Modern English also introduces some pedagogical challenges, the most obvious of which is the unavoidable trade-off between time spent acquiring general knowledge about Early and Late Modern English and time devoted to research on a highly specialized topic. I am currently attempting to address this problem by including a number of pre-recorded lectures in the module, which students can access through their online learning platform. This type of complementary teaching frees up class time for discussion of exercises, looking at genuine Early and Late Modern English texts, etc., in a type of flipped-classroom set-up. However, there is of course also a limit to how much time teachers can invest in creating such resources. Increased collaboration among teachers, with electronic materials being shared among learning platforms, will hopefully contribute to providing students with the assistance they need while keeping teachers' workload manageable.

Another problem concerns ensuring that all students have an opportunity to reach the intended learning outcomes. When almost half of a student's grade depends on a specialized essay, each student's trajectory through the module necessarily becomes more individualized—as is appropriate for the advanced level of this course. Each final research assignment must thus be adapted to the learning outcomes so that students are exposed to comparable challenges overall (which is also important in terms of assigning fair grades). The final class, where each student reports on their research to the teacher and to the other students, fills an important function in this regard, as it provides an opportunity to raise awareness of difficulties that are characteristic of some topics more than others.

Finally, one recurring problem concerns completion. Despite reminders, students frequently postpone selecting a topic until a fairly late

stage in the module. Common consequences are that the completion of their final research assignments is in turn postponed until the following term or that the quality of work that is submitted on time suffers somewhat; the oral presentations during the final class also become less educational when students are further away from completing their analyses than anticipated. Moreover, in a few cases, students who have not previously carried out much independent research experience great difficulty in completing their final assignments at all (although, at least in the case of students on the Master's programme, it is arguably better to discover—and attempt to remedy—that problem during this module than when they are working on their Master's theses).

These challenges notwithstanding, I remain convinced that independent, corpus-based research has an obvious place in modules on the history of English. The advantages are noticeable both in terms of how students who complete the module engage with their Master's theses a year later and regarding their interest in further work on the history of the English language. Of the three problems that have been discussed above, the first can be solved through co-operation among teachers and gradually remedied through more supporting materials being made available to students in addition to the contact hours with their teacher(s); the second is mainly a matter of syllabus and project design. As regards completion rates, the ability to carry out and complete research-related work independently is a transferable skill in itself, and it is hoped that deadlines as well as supporting activities and materials will help to reduce the proportion of students who postpone completing the final research assignment; by extension, the proportion of students who do not finish their subsequent Master's theses on time may also decrease, as previous experience of research design is likely to stand them in good stead when they engage with research projects that are larger in scope.

In sum, the familiarity with theoretical and methodological perspectives on corpus linguistics as well as the recent history of English that students get from working with Modern English corpus data contributes not only to their expertise as regards the English language between 1500 and 1945, but also to their general skills as empirical linguists. The more complex processes in Bloom's revised taxonomy become available to students whose previous experience of the history of English may have concerned mainly description and summary. Even more generally, carrying out one's own research and thus making an actual

contribution to our knowledge of an academic field—while also becoming aware of the limitations of that contribution—is one of the best ways of fostering critical thinking, which is a key goal of tertiary education in any discipline. The value of historical linguistics thus increases exponentially when students do their own research, and very large historical and/or diachronic corpora greatly facilitate their transition from learners to researchers.

References

Anderson, Lorin W., and David R. Krathwol (eds.). 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives*. New York: Addison Wesley Longman.

Beal, Joan C. 2004. *English in modern times: 1700–1945*. London: Arnold.

Biber, Douglas, with Jesse Egbert, Bethany Gray, Rahel Oppliger, and Benedikt Szmrecsanyi. 2016. Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In *The Cambridge handbook of English historical linguistics*, edited by Merja Kytö and Päivi Pahta, 351–375. Cambridge: Cambridge University Press.

CED = A Corpus of English Dialogues 1560–1760. 2006. Compiled under the supervision of Merja Kytö and Jonathan Culpeper.

The Chicago Manual of Style Online. 2017. 17th ed. Chicago: University of Chicago Press.

CLMETEV = The Corpus of Late Modern English Texts (Extended Version). 2006. Compiled by Hendrik De Smet.

COHA = The Corpus of Historical American English: 400 Million Words, 1810–2009. 2010–. Compiled by Mark Davies.

The Corpus Resource Database (CoRD). n.d. Accessed on 13 June 2022. https://varieng.helsinki.fi/CoRD/index.html/.

Culpeper, Jonathan, and Merja Kytö. 2010. *Early Modern English dialogues: Spoken interaction as writing*. Cambridge: Cambridge University Press.

Davies, Mark. 2012. Some methodological issues related to corpus-based investigations of recent syntactic changes in English. In *The Oxford handbook of the history of English*, edited by Terttu Nevalainen and Elizabeth Closs Traugott, 157–174. Oxford and New York: Oxford University Press.

Davies, Mark. 2019. Using (and useful) corpora for the study of HEL. In *Teaching the history of the English language*, edited by Colette Moore and Chris C. Palmer, 320–323. New York: The Modern Language Association of America.

Davies, Mark. n.d. Accessed on 13 June 2022. https://www.mark-davies.org/.

EEBO = Early English Books Online. 2017. Compiled by Mark Davies; created as part of the SAMUELS project.

Ellegård, Alvar. 1953. *The auxiliary 'do': The establishment and regulation of its use in English*. Stockholm: Almqvist & Wiksell.

Grund, Peter, and Terry Walker. 2006. The subjunctive in adverbial clauses in nineteenth-century English. In *Nineteenth-century English: Stability and change*, edited by Merja Kytö, Mats Rydén, and Erik Smitterberg, 89–109. Cambridge: Cambridge University Press.

Jacobs, Andreas, and Andreas H. Jucker. 1995. The historical perspective in pragmatics. In *Historical pragmatics: Pragmatic developments in the history of English*, edited by Andreas H. Jucker, 3–33. Amsterdam and Philadelphia: John Benjamins.

McEnery, Tony, Richard Xiao, and Yukio Tono. 2006. *Corpus-based language studies: An advanced resource book*. Abingdon: Routledge.

The Movie Corpus. 2019. Compiled by Mark Davies.

Nevalainen, Terttu. 2006. *An introduction to Early Modern English*. Edinburgh: Edinburgh University Press.

Nevalainen, Terttu, and Helena Raumolin-Brunberg. 2017. *Historical sociolinguistics: Language change in Tudor and Stuart England*. 2nd ed. Abingdon and New York: Routledge.

OBC = The Old Bailey Corpus. Spoken English in the 18th and 19th centuries. 2012. Compiled by Magnus Huber, Magnus Nissel, Patrick Maiwald and Bianca Widlitzki.

OED = Oxford English Dictionary online. https://www.oed.com/.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London and New York: Longman.

Rissanen, Matti. 1986. Variation and the study of English historical syntax. In *Diversity and diachrony*, edited by David Sankoff, 97–109. Amsterdam and Philadelphia: John Benjamins.

Schwarz, Sarah. 2017. 'Like getting nibbled to death by a duck': Grammaticalization of the GET-passive in the TIME Magazine Corpus. *English World-Wide* 38(3): 305–335.

SCOTUS = the Corpus of US Supreme Court Opinions. 2017. Compiled by Mark Davies.

Smitterberg, Erik. 2016. Extracting data from historical material. In *The Cambridge handbook of English historical linguistics*, edited by Merja Kytö and Päivi Pahta, 181–199. Cambridge: Cambridge University Press.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Chichester: John Wiley & Sons.

University of Toronto. n.d. Active verbs for Bloom's revised taxonomy. Accessed on 18 June 2022. https://teaching.utoronto.ca/teaching-support/working-w-grads/ci-ta-relationship/active-verbs/.