LUND UNIVERSITY

which vs. *that*: a corpus study

Carl-Staffan Svenbro

ENGK01
Bachelor’s Degree Essay
Spring 2017
English Studies
Supervisor: Mats Johansson

Abstract

Based on corpora and earlier studies, this paper mainly attempts to answer the question how constructions of restrictive *which* and *that* have developed in comparison to one another in American and British English news until today. Corpus queries are designed to match patterns of particular object and subject gap constructions, such as *I like the ball that/which is green* with subject gap and *It is the toy that/which I prefer*, which has object gap. Each query generates a *query set*, which includes all search hits for that query. Rather than checking all entries in every query set, entries of randomized samples are verified manually. From each such sample, the proportion of relevant entries, *relevance index*, is calculated. Relevance index helps us to estimate the relevant frequencies of the query sets. These estimations are essential for calculation of *frequency indexes*, which compare how frequencies of *that* and *which* clauses have progressed over time. In British English, the results are mixed with opposite tendencies for different time periods and news categories. As for American English, all data consistently support a significant frequency increase in *that* with a corresponding decline in restrictive *which*.

Table of Contents

1. Introduction

A noun phrase may be postmodified by a *relative clause* (Biber et al., 2002, p. 257). In (1), the head of the noun phrase, *ball*, is post-modified by *,which is blue,* beginning with comma followed by the relativizer *which.* The comma indicates that the relative clause is *non-restrictive* (p. 280). It adds a descriptive piece of information to its antecedent *ball*, but without identifying it in a larger set.

(1) the **ball**, which is blue

There is, however, another type of relative clause that *does* identify a larger set from which the antecedent has been selected. This type of relative clause is called *restrictive*. By simply removing the comma from example (1) we get (2) with underscore marking the subject gap.

(2) the **ball** which _ is blue

While *which* may be used both restrictively and non-restrictively, there is one relativizer that is not supposed to be used non-restrictively, namely *that*. In other words, the *that*-relativizer could be used in a sentence such as *I like the ball that is yellow.* but not in **I like the ball, that is yellow*.

It should be noted that *that* is not always used as a relativizer. There are for example cases where *that* is followed by an independent clause, such as in *He said that he did not like the book*, referred to as complement clause. In such cases, *that* is said to be a complementizer (cf. Biber et al., 2002, p. 308).

The usages of *that* and *which* have been studied in different ways. For example, a study by Leech et al. claims that the usage of *that* in relative clauses (AmE) increased by 73,1% between 1961 and 1991/1992, while the usage of relative *which* declined during the same period (Leech et al., 2009, p. 229). According to Leech et al., a likely explanation for the big increase in American English of restrictive *that* is the influence by American prescriptivism (Leech et al., 2009, p.229-230).

This essay examines how *that* and *which* in restrictive relative clauses have developed in comparison to one another in modern British and American English. It also analyzes possible causes. Corpus data based on American and British newspaper-based news and web-based news are focused on. This particular choice of register will be motivated in section 3.

Data were collected from several corpora. Only restrictive constructions with either object gap, e.g. (3), or subject gap, such as (2), were included in the data. To find out more about underlying causes, a database of usage guides was studied.

(3) It is the toy that I prefer _.

As a starting point we formulate four initial hypotheses. Firstly, we hypothesize that the frequency of *that* as relativizer has increased until today (in comparison to *which*) in news (mostly in AmE and to a lesser degree in BrE). Secondly, we guess that an important explanation for this development could be that American usage guides generally favor a distinct separation between non-restrictive *which* and restrictive *that* more strongly than British counterparts. Thirdly, we expect that object gap constructions are overall more common than subject gap constructions in terms of raw frequency. Fourthly, we speculate that the typical reason why entries are excluded from the data (i.e. not relevant) is that they exhibit other gap types besides object and subject gap.

Except for verifying the hypotheses above, this essay has a methodological objective of developing a strategy for analysis of corpus sets[1] that are very large. Analyzing smaller subsets[2] allows us to make reasonably accurate predictions about these sets. The approach aims at facilitating manual processing of corpus entries with the final result still being statistically reliable. As partly addressed in section 5, further studies are necessary to fine-tune this methodology.

The following section serves as a background for the study. Some earlier studies and the concept of frequency index are focused on in that section. In section 3, we review the corpora used for this study. The usage of queries and other data related aspects (including calculation methods) are explained. Section 4 presents some general tendencies and contrasts our study with earlier studies.

[1] In this essay, such corpus sets are named *query sets*, which will be defined in section 3.2.
[2] We commonly refer to such smaller subsets as *samples* (see section 3.2).

2. Background

2.1 Reflections on earlier corpus studies

There seems to be a stronger inclination in American than in British English to adopt new language policies. The quote below by Leech and Smith (2009) serves as a background to the earlier studies by Leech et al. (2009) and Biber et al. (1999), and for our own corpus study later in this essay.

> The evidence provided by the Brown family of corpora […] often shows AmE to be in the lead or to show a more extreme tendency, and BrE to be following in its wake.
>
> (Leech and Smith, 2009, p. 176)

Statistics in Leech et al. (2009, p. 229; p. 309-310, table A10.10 & A10.11a,b), based on the Brown family of corpora, indicate that the frequency of relative *which* decreased by 9,4% in British English and by 34,4% in American English between 1961 and 1991/1992. Data based on the same corpora and time interval suggest that the frequency of *that* as relativizer increased by 73,1% in American English compared to just 15,3% in British English. It should here be noted that Leech et al. (2009, p. 309, table A10.10) do not explicitly state that the relative pronoun *which* is limited to restrictive cases, so we may reasonably assume that their data reflect both restrictive and non-restrictive constructions.

Similar tendencies as those found by Leech et al. but occurring somewhat later in time could be seen in a corpus study by Biber et al. (1999, p. 616), in which they compared American and British English news. As inferred from their data, restrictive *that* was about 50% more common in American news than British news and the frequency of restrictive *which* in American news was about 1/3 of the frequency in British news. For column charts based on these studies, see Appendix 3 (Chart A3.1, A3.2 and A3.3).

As for explanations of the stronger tendencies of restrictive *that* in American English, Leech et al. (2009, p. 229-230) focus on the American prescriptive tradition as a probable cause, indicating that many American usage guides disapprove of restrictive *which* (p. 5).

We will later supplement and compare our data with the studies conducted by Leech et al. (2009) and Biber et al. (1999). The question how these studies are relevant to our study and what the differences are will be further addressed in section 3.

2.2 Frequency index

In the two studies referred to in section 2.1, relative frequencies were measured, i.e. frequencies per million words. However, there is another indicator that measures the proportional relation between two constructions in terms of frequency, namely *frequency index*.

Frequency index is described by Mair (2006, p. 115), who uses it to compare constructions with *get*-passives and *be*-passives in contemporary English. The formula used by Mair (with different denotations) is essentially given in (4), where F_{get} represents the raw frequency of *get*-passive constructions + verb and F_{be} the frequency of *be*-passive constructions + verb. Put simply, (4) expresses the relation between the frequency of *get*-passives and the total frequency of *get*- and *be* -passives.

(4) $100 * \mathrm{F}_{\mathrm{get}} / (\mathrm{F}_{\mathrm{get}} + \mathrm{F}_{\mathrm{be}})$

In this paper, we use the same indicator for making comparison between frequencies of restrictive *which* and *that* during specific time periods, with the difference that we do not multiply the ratio expression by 100 as done by Mair (see section 3.2).

2.3 Some examples of recommendations given by usage guides

In sub-section 2.1, we referred to two earlier studies as background to this essay, which suggest that usage of restrictive *that* has been increasing from the early sixties and onwards and was much more common than restrictive *which* in American English according to the study published in 1999 (Biber et al., 1999, p. 616). In this sub-section, we will provide more background to our study by offering some examples from handbooks, style-guides, writing manuals etc. (in this study collectively referred to as *usage guides*, a term used by English Usage Guide Database (2017), from which we will later collect data (see section 3)).

Many American Usage guides seem to recommend binary distribution between *which* and *that* in restrictive and non-restrictive clauses. Thurman (2002), for example, declares "[f]or clauses that don't need commas (restrictive clauses), use *that*. For nonrestrictive clauses, which need commas, use *which"* (p. 16). Strunk (1959) suggests "*which*-hunting" and asserts that "it would be a convenience to all if these two pronouns were used with precision" (p. 47). Jordan (1976) states "*That* is preferred in restrictive clauses […]. In nonrestrictive clauses, *which* is mandatory […]" (p. 206). *The Chicago manual of style* (2003) provides similar recommendations for "polished American prose" (p. 230). Notably, none of the sources mentioned in this paragraph seem to motivate their position[3].

Two American usage guides that appear to be more tolerant towards restrictive *which* are Danesi (2006, p. 261) and Pinker (2015, p. 235). Examples of British grammars and usage guides that accept restrictive *which* are Collins *COBUILD English grammar* (2011, p. 381) and Aarts (2011, p. 199).

While opinions expressed by handful of sources such as the ones mentioned above certainly give us an idea about the opinions of *some* authors and usage guides, the scope is too limited to draw any general conclusions. To be able to explain more reliably why changes in language happen in a certain way, we would have to expand the scope, allowing us to study a larger number of entries related to a certain language feature. In the following section we describe our approach to accomplish this.

[3] The question *why* these American usage-guides appear to give recommendations without explaining their standpoint is interesting but has not been addressed by this paper. Hopefully, future studies will pay closer attention to this question.

3. Methods and Materials

3.1 Materials

Like the earlier studies described in the background section, our study is corpus based. Reliance on corpora has been helpful for several reasons. Firstly, it made it possible to extract a much greater number of relevant entries than would have been possible had we used physical resources (which was initially considered), making the data more statistically reliable. Secondly, the fact that the corpus data are grammatically annotated made it possible to formulate queries that specified the grammatical patterns that we wanted to compare. Thirdly, information about time, register and dialect enabled diachronic analysis of the data in specific registers and dialects.

The corpora used are the TIME Magazine corpus, the Now corpus and the Brown family of corpora. In this section, we will explain more about these corpora; what they are and what kind of data they contain in terms of register, regional variety, time scope etc. We will also describe what sub-corpora or subsets of corpus data we used for our study. In the next sub-section (3.2), we will provide more specific information how we interacted with the corpora and designed our queries to extract the data we needed. The main corpora used for this study (excluding sub-corpora) are listed in Table 1. For complete lists of queries and related information for each query such as corpus, retrieval date, variety, relevance index[4], sample[5] size etc., see Appendix 1.2 (table A1.2.1 and A1.2.2).

Table 1. Corpora used for our study.

the TIME Magazine corpus
the Now Corpus
the Brown family of corpora

[4] The term *relevance index* will be explained in section 3.2.

[5] The term *sample* will be defined in section 3.2.

As mentioned in the introduction, the corpus data are limited to news (news-paper based and web-based). This specific choice of register is inspired by Biber et al. (1999) who argue as follows.

> […] newspapers tend to be written for, and read in, a single region or nation, and thus they provide one of the best reflections of American English v. British English dialect differences in writing. (Biber et al., 1999, p. 16)

The TIME Magazine corpus is based on American English news from the TIME Magazine, and covers the time interval 1923 to 2010. It contains 275.000 articles or about 100 million words (TIME Magazine corpus, 2017).

The Now Corpus includes news that is web-based. It belongs to the same collection as the TIME Magazine corpus, but is much larger. It contains about 4.2 billion words of data or 10.000 news articles. It covers the time interval from 2010 to present and various regional varieties, such as Canadian, Australian, British and American English (Now Corpus, 2017). Of the regional varieties listed above, only British and American English data have been included in this study.

The version of the Brown family of corpora that we used belongs to a different collection than the two corpora described above. In this collection, it goes under the name *ICAME – Brown family*. It contains 4 sub-corpora relevant to this study. According to Leech et al. (2009, p. 9) these are Brown (AmE, 1961), LOB (BrE, 1961), Frown (AmE, 1991/92) and FLOB (BrE, 1991/92). On average, these corpora contain 1.150.000 words each (Corpuscle, 2017). The reader should be reminded that among the four Brown family text registers mentioned by Leech et al. (2009, p. 41) - *press, general prose, learned* and *fiction* – our study only relies on *press* (i.e. news).

Among the main corpora used (see Table 1), data have been collected from well-defined subsections, except for the TIME Magazine corpus, in which case data are obtained from the entire corpus. Each corpus has contributed with an equal number of data sets in American and British English, except for the TIME Magazine corpus, which is entirely based on American English news.

Apart from our own corpus data, we also rely on data from previous corpus studies as presented in section 2 conducted by Leech et al. (2009) and Biber et al. (1999). The data by Leech et al. are provided in raw format (see Leech et al., 2009, p. 309-310; table A10.10, A10.11a & A10.11b). The data from Biber et al. (1999, p. 616, table 8.9) were provided as a bar chart with markers, which could easily be counted, with each marker representing 200 per million words. The question how these studies are relevant as background and comparison material to this study will be addressed below.

As for Leech et al. (2009, p. 309, table A10.10, A10.11a & A10.11b), their study is relevant because our study is similar enough to allow us to compare our data with theirs. However, there is an important difference between our study and Leech et al.'s. Most likely, they do not restrict their data to restrictive *which* but even include non-restrictive cases. This may account for differences between our data and theirs, which will be further addressed in section 4.

The study by Biber et al. (1999, p. 616) is relevant as it, like our study, addresses frequencies of *which* and *that* in restrictive relative clauses in the news register. In section 4, our results will be contrasted with theirs and possible differences explained. The most important difference is that their study covers a period just before the turn of the millennium, when their grammar book was published, and earlier, while the coverage of our study continues after that. Unlike Biber et al.'s, our study distinguishes between newspaper-based news and news that is web-based.

While the corpora used for this study have been helpful for tracking diachronic developments of various constructions, they have not been of any assistance in explaining such developments. In the background section, we included a couple of references to some usage guides in order to at least give an indication why *that* seems to increase in popularity compared to restrictive *which* in American English. For a large-scale analysis, however, relying on book sources alone seems impractical, as there are so many of them. Instead of relying on such sources, this study has used a database available on the Internet named Hyper Usage Guide of English or H.U.G.E (English Usage Guide Database, 2017). The methods of extracting useful data from this database will be further addressed in section 3.2.

3.2 Methods

In this study, we employed queries to extract applicable information from the corpora. The query syntaxes differed between the Brown family of corpora and the other corpora, which essentially shared the same syntax. What they all had in common, however, was allowing inclusion of generic elements, such as part of speech. Whereas the Brown family of corpora required all information to be included in the query string, the other corpora allowed some information to be specified separately, for example in list boxes.

Apart from query syntax, it was necessary to make the queries more or less inclusive depending on corpus. Basic queries that corresponded exactly to the patterns that we needed (such as *noun* + *that* + *pronoun* + *verb* + *full stop*) were only possible in the TIME magazine corpus. When trying out these basic queries in the Now corpus and the Brown family of corpora, problems occurred. In the case of the Brown family of corpora, the main issue was that the query structure did not generate a data set large enough, so the structure needed to be changed to be more inclusive, for example by allowing optional elements between standard elements. For listings and explanations of the queries, please refer to Appendix 1.2 and 1.3.

The opposite problem occurred with the Now corpus. Because of the fact that the number of hits for each query position was too large, the queries were required to be redesigned to be *less* inclusive. Accordingly, instead of using a generic pronoun slot as in (6), separate queries with this slot replaced by some common pronouns (I, you, we, he she, they) were created, which resulted in a much larger number of queries in the Now corpus than in the TIME Magazine corpus. Subject gap constructions with the pattern *noun* + *that/which* + *verb* had to be modified as well for the same reason, with the verb slot replaced by *does* and *do,* which also allowed for negations to be included.

Essentially, two different constructions of restrictive relative clauses with *which* and *that* were targeted, namely constructions with either object gap or subject gap. Collecting data entries based on two different patterns instead of one seemed advantageous, as the two patterns would validate each other in terms of accuracy. To get constructions with subject gap, constructions with the patterns shown in (5) were searched for. To get constructions with object gap, we targeted the patterns in (6).

(5) noun + that / which + verb

(6) noun + that/which + pronoun + verb + full stop (.)

A hypothetical example of a construction generated by (5) would be *a book that pleased me* and for (6) *a book which he likes* + full stop. Please note that (5) and (6) are not the exact queries used but rather illustrate the general principles behind the queries (see Appendix 1.3 for specific information on query syntax). For a complete reference list of all queries used in this study, please refer to Appendix 1.2. The reader should be reminded that the queries generalized in (5) and (6) needed to be modified for the Now corpus and the Brown family of corpora as previously described in this sub-section.

Each query executed generated a random subset of entries automatically, with the exception for the Brown family of corpora were the randomizations were done manually. Typically, 100 or fewer entries were generated. In cases where the number of entries was below 100, the total set (sometimes referred to as *query set*) most often coincided with the generated subset. In this essay, we refer to such a list of entries as a *sample*. Each entry in every sample was analyzed manually to see whether it was relevant, i.e. contained a restrictive relative clause with either object or subject gap. We call the proportion between the frequency of relevant entries and sample size *relevance index*. The relevance index was used to estimate the total number of relevant entries in the query set. See Appendix 2.2 for a simple example on how it is calculated.

In this paper, frequency index is the most central key value for data comparison. As shown in section 2, it measures the frequency relation between two grammatical features, in which the frequency of one of the features is divided by the total frequency sum of both features. For example, when measuring frequency index for *that* constructions in relation to constructions with *which*, we calculate the ratio between *that* construction frequency and the total sum of *that* and *which* construction frequencies. In this paper, we often express the frequency index above as *that* (compared to *which*), *that* (vs. *which*) or *that* (in relation to *which*). Frequency index is calculated in (7) with I_{that} referring to frequency index for *that* and *F* indicating raw frequencies of *that* and *which*. Over time, frequency index measures how frequencies of constructions with *which* and *that* have developed in comparison to one another.

(7) $I_{that} = F_{that} / (F_{that} + F_{which})$

Based on the estimated frequencies for various points in time, frequency indexes were generated for each pair of constructions to compare the frequencies of *which* and *that* in restrictive clauses in the same type of construction. It often refers to a percentage value such as in a case described by Mair (2006, p. 115), which we referred to in section 2. Unlike Mair, however, we do not multiply the frequency index ratio by 100 to obtain a percentage value, as this is done automatically in Microsoft Excel when changing the number formatting to percentage. Appendix 2.1 contains an example how it is calculated.

As described earlier in this sub-section, the queries used with the Now corpus required modification to be less inclusive, which raised the question whether the samples were still representative. For example, would the query pattern (8) have given the same outfall as (9) in the Now corpus had it been possible to execute?

(8) noun + which / that + verb

(9) noun + which /that + do / does / don't / doesn't

To answer that question, the two queries (for *that* and *which*) were compared in a different corpus, the TIME Magazine corpus, to verify that the outfall would have been approximately the same. The result of this comparison is presented in section 4.1.

In section 2.1, we made a reflection on two previous studies made on relative frequencies of *which* and *that* as relativizers. To be able to compare our results with the results from these studies, the data were recalculated and adapted to conform to the format of our study, which is based on frequency index (described in Appendix 2.1 and above). After having recalculated the data in this manner, we obtained percentage values (frequency indexes) where each such value measured the relation between the frequency of *that* constructions in relation to the total frequency of restrictive *which* and *that* constructions.

We also had to make an assumption about the Biber et al. data, as the authors had not stated any exact time interval for these. We assumed that the fact that they presented the data as relevant, without stating an exact time frame, strongly suggested that it had been relevant at the time of publication of their book (i.e. 1999). We therefore opted to assign this point in time to these data rather than a time interval.

As described above, the queries were designed to match patterns of either object gap and subject constructions. As the queries were not completely accurate, it was necessary to check all entries manually in order to establish relevance index for each sample, which, in turn, was used to estimate the relevant frequencies of the query sets (see above).

In the process of checking the samples, not only were entries classified as included or excluded. Additionally, excluded entries were also classified by *exclusion factor* (sometimes also referred to as *exclusion category*), referring to specific causes *why* these entries were not included. Accordingly, each non-relevant entry was assigned one of six exclusion categories, listed in Table 2.

Exclusion ratios were then calculated for each of these exclusion categories for every sample and multiplied by the size of the query set. This approach gave us estimations how many entries were excluded for each query set and exclusion category. Using these results, the overall exclusion ratios were calculated[6] for samples from object gap queries and subject gap queries respectively with separate cases for *that* and *which*-queries (i.e. in total 4 cases for each corpus[7]). The same calculation was done for each of the six exclusion factors for every sample. A summary of the most frequent exclusion factors for different queries (with overall exclusion ratios specified) in various corpora and sub-corpora will be included in section 4.2.

Table 2. Exclusion factors and hypothetical examples, with grammatical features underlined and gaps marked with underscore. For additional examples from corpora, see page and table references[8] in the *Corpus examples* column.

Exclusion factors	**Hypothetical Examples**	**Corpus examples**
Complement clause	He told the shareholders that the venture had failed.	p. 22 (18) p. 22 (19)
Adverbial Gap	I saw her on the day that she was born _ .	p. 23 (22)
Subject Predicative Gap	We all helped change him into the person that he became _ .	table A1.1.2
Non-Restrictive clause	I like my new car, which is red.	table A1.1.2
PP complement gap	The ball which I played with _ was green.	table A1.1.2
Miscellaneous	I like my new toy, the talking teddy bear that is.	p. 21 (15) p. 21 (16)

[6] The overall exclusion ratios for a certain group of query sets were calculated by dividing the totals of the estimated exclusion frequencies by the totals of the query set sizes for that group.

[7] See Table 4 section 4.2

[8] See also table A1.1.1 in Appendix 1

Analysis of inclusion reasons was quite straightforward. As object gap and subject gap queries were separated by default, no subcategorization for included entries was needed. Instead, relevance index was used to estimate the relevant frequencies of the query sets. The total sums were then calculated for estimated frequencies[9] for all subject-gap and object gap queries respectively in each corpus. Finally, the ratio between these total frequencies in every corpus was established. The inclusion ratios for all corpora will be presented in section 4.

As described in section 3.1, the present study relies on an online database (H.U.G.E) of usage guides to explain the developments of *that* in comparison to restrictive *which* in British and American English. In the following, our methods of extracting data from this database will be addressed.

The database allows the user to enter different search criteria to search for relevant entries from usage guides (English Usage Guide Database, 2017). In order to separate American entries from British ones, we performed one search with author nationality set to United States and another search with the same parameter set to United Kingdom. For both these cases, the *problem term* parameter was set to *that/which*. About 20 entries by British authors and roughly the same number of entries by American authors were found. Each entry was then evaluated manually and coded depending on whether it seemed to be in favor or not of a clean distribution between restrictive *that* and non-restrictive *which* or if that information was not available. In section 4, the result of this analysis is presented in pie chart form.

All entries by American and British authors are listed in Appendix 1.4 (table A1.4.1 and A1.4.3 respectively). The same appendix contains general information about these data such as retrieval date for the entries and URL address. As the H.U.G.E entry listings lack specific publication information, our *year* information in table A1.4.1 and A1.4.3 is occasionally ambiguous in cases where one author is associated with several publications in a separate usage-guide list, which is not distinctly linked to the main entry list. In the ambiguous cases, the possible *year* options have been delimited with slash (/).

[9] To calculate estimated relevant frequencies, the total frequency was multiplied by relevance index for each query.

4. Result and discussion

The results presented in this section are based on data collected from six corpora (including sub corpora) described in section 3 and two previous studies (see section 2). 56 samples were extracted from the corpora by execution of an equal number of queries and more than 4000 entries were analyzed manually to determine how big proportion of the entries were relevant in each sample.

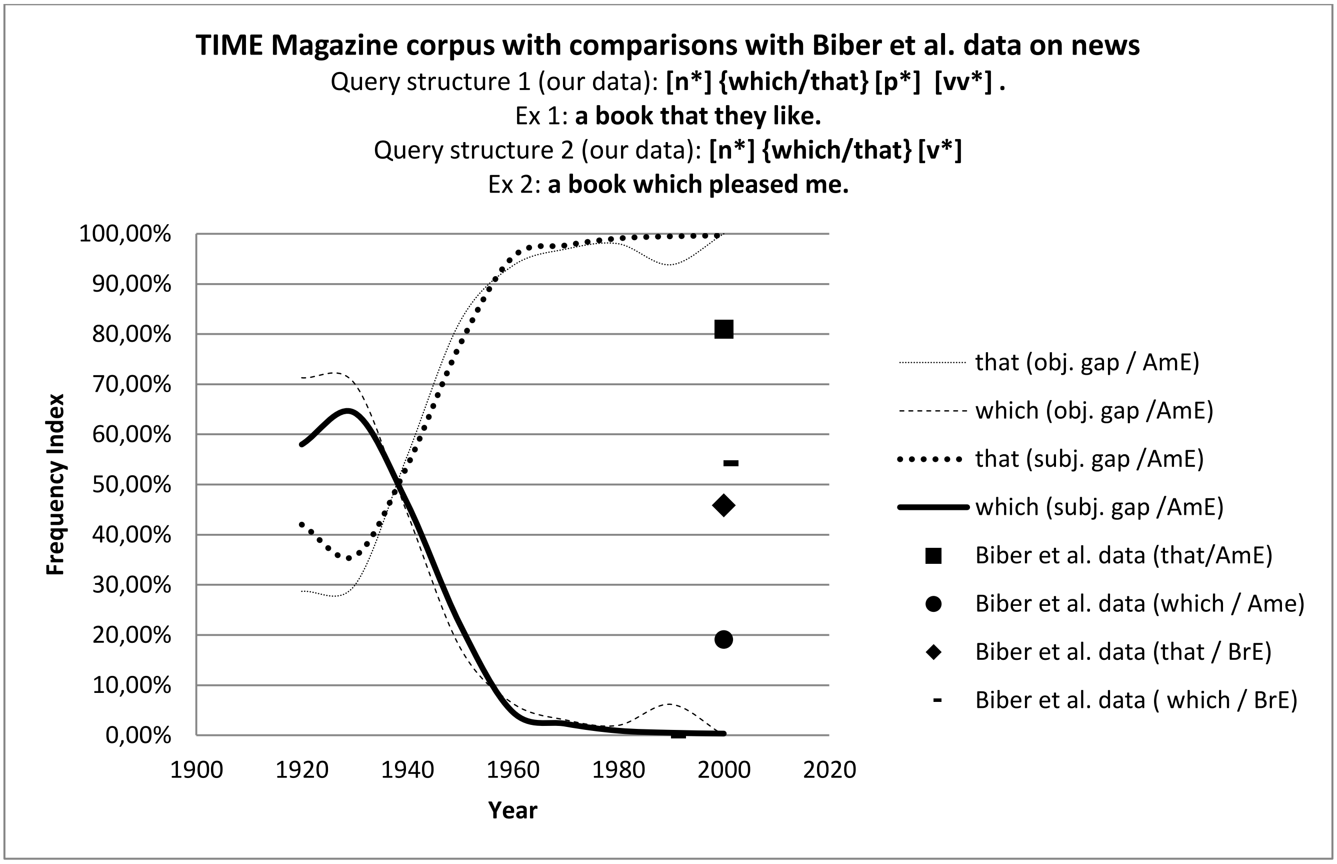
4.1 Comparison with earlier studies and general tendencies

In section 2.1, we reviewed two earlier studies done on relative frequencies of *which* and *that* (occurrences per million words). In this section, we will compare our data with these two studies. In the previous section, we described our method of adapting the data of the earlier studies to enable comparison.

Firstly, we will compare our data with Biber et al.'s study, which covers both American and British English (Biber et al., 1999, p.616). As our study lacks corresponding data for British English for this period, the comparison was done between our American English data and Biber et al.'s data based on both varieties. The results are shown in Chart 1. It should be noted that Biber et al. did not mention any exact time frame for their data. We have therefore chosen to assign the year of book publication to these data for reasons discussed in section 3.2.

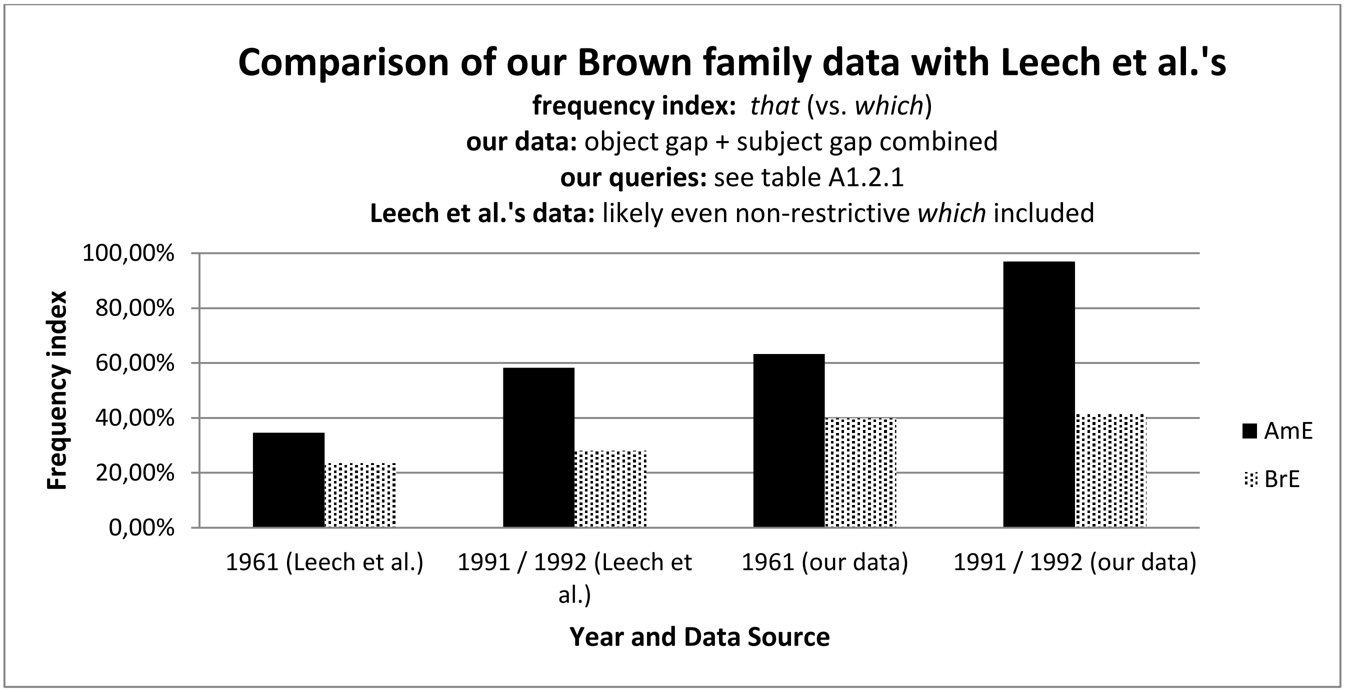
Possible explanations why our data slightly deviate from the American English data based on Biber et al.'s study could be that Biber et al.'s coverage of the news register is more general. We could guess that the less frequent usage of restrictive *which* in the TIME Magazine data during the period reflects stronger attitudes against restrictive *which* and in favor of *that* as relativizer among editors working for the TIME Magazine compared to other news-paper editors.

Chart 1. Comparison between our data collected from the TIME Magazine corpus (AmE) and the data contributed by Biber et al. (1999, p.616) for newspaper-based news in general in American English. British English data provided by Biber et al. are included as well. Each scatter plot that has not explicitly been specified as *Biber et al.* refers to our data. Ex 1 & 2 in the chart are hypothetical.



Next, we will compare our data extracted from the Brown family of corpora with the corresponding data provided by Leech et al. As described in section 3 and like in the case of Biber et al.'s data presented above, the format of the data was adjusted to enable comparison.

Chart 2. Comparison between the data presented by Leech et al. (2009, p. 309-310, table A10.10, A10.11a & A10.11b), with adaptations described in section 3.2, and our data. Unlike our data, non-restrictive *which* has likely not been excluded from Leech et al.'s data. In the chart, frequency index is calculated as F_{that} / (F_{that} + F_{which}) where F represents raw frequencies of *that* or *which*.



The result of the comparison between our data and the data by Leech et al. (2009, p. 309-310, table A10.10, A10.11a & A10.11b) is visualized in Chart 2. The tendencies are very similar. There are, however, differences that may be explained by the fact that Leech et al. most likely included non-restrictive *which* constructions in their data as well, which may account for the generally lower frequency indexes compared to our data. Another factor that may explain the differences is the fact that our data only cover the press/news register while the Leech et al. data comprise all written registers.

In the remaining part of this sub-section, we will first discuss the meaning of these data and then present and discuss data on news that is web-based, which have not been specifically covered by Leech et al. (2009) or Biber et al. (1999) to our knowledge.

As shown in Chart 1, there has been a steady increase of restrictive *that* (in relation to to *which*) from 1930s and onward in American English. A very rapid increase of restrictive *that* from the 1930s until the 1960s was seemingly replaced by a continuous (but slower) increase, which continued until around the year of 2000. The data from Biber et al. (1999, p. 616) as well as Brown corpus data from us and Leech et al. (see Chart 2) essentially support the increasing trend of restrictive *that* (compared to *which*) between 1961 and 1991/92 respectively until 1999, although there are differences, which have already been addressed. Interestingly enough, the data provided by Biber et al. (see Chart 1) also reflect usage patterns in British English news, which indicate that restrictive *which* actually seems to have been more common than restrictive *that* around 1999 in British English. In other words, according to these data, there seem to have been opposite tendencies in British English news compared to American English news.

The results of our Now corpus data have been presented in Chart 3 (American English) and Chart 4 (British English). The high frequencies in American English (Chart 3) were expected and rather consistent with the data presented in Chart 1. American web-based news seems to have a stronger preference for restrictive *that* compared to *which* than newspaper-based news as reflected by Biber et al.'s data in this chart.

What is more surprising is that the preference for restrictive *that* seems to be almost as strong in British English as American English web-based news with just 5-10 percentage points difference. We can only speculate what the reason might be. Could it be that British writers of web-based news have somehow been influenced by American prescriptivism? As Biber et al.'s data presented in Chart 1 indicate, restrictive *which* was even more common than restrictive *that* in British English news around the turn of the millennium. It is therefore most likely that the strong preference for restrictive *that* vs. *which* in British English web-based news does not apply to newspaper-based news in general but is limited to news on the Internet.

Chart 3. Diachronic development showing the usage of *that* (in relation to restrictive *which*) between 2010 until today in American English web-based news. In the chart, frequency index is calculated as F_{that} / (F_{that} + F_{which}) where F represents raw frequencies of *that* or *which*.

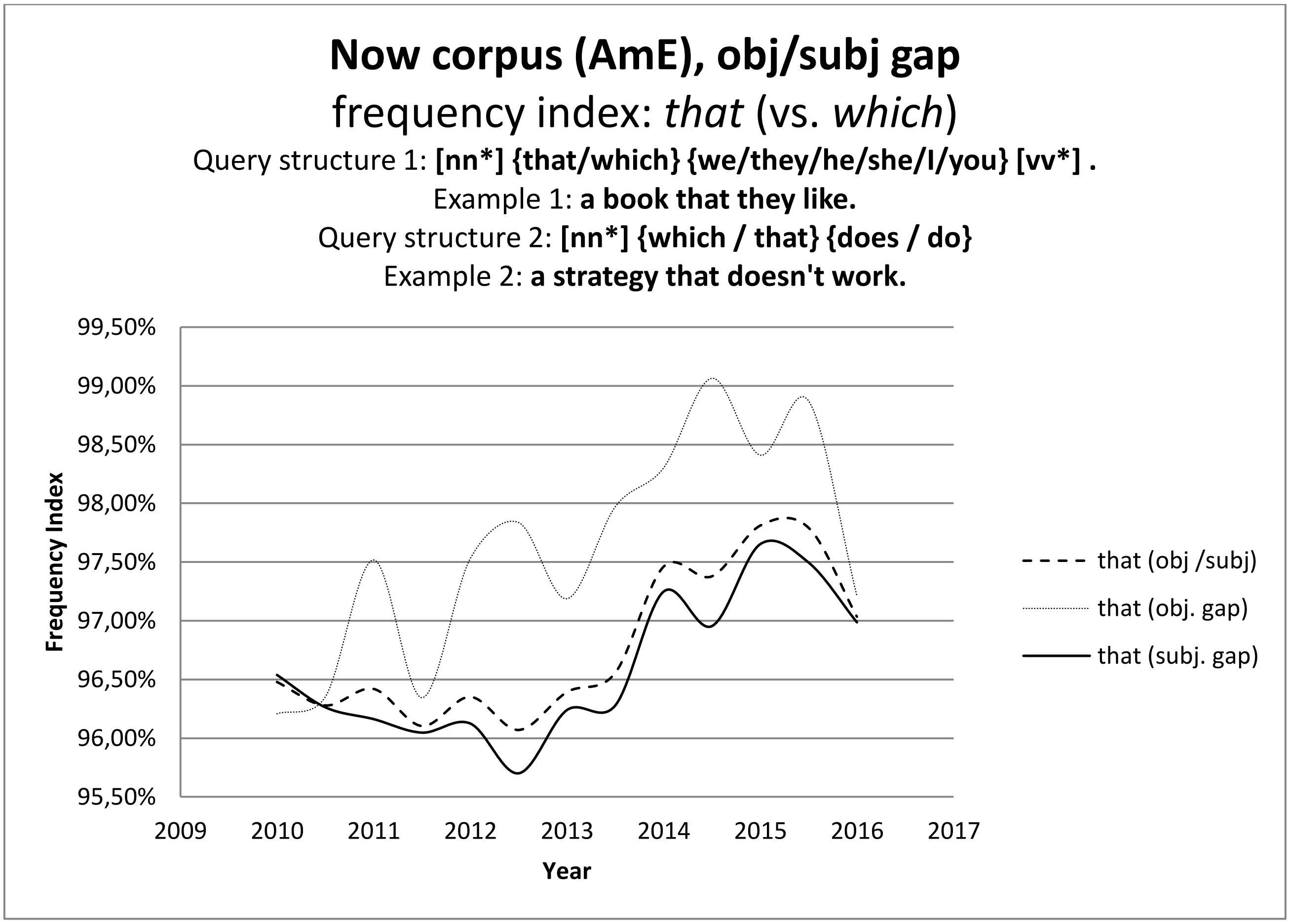
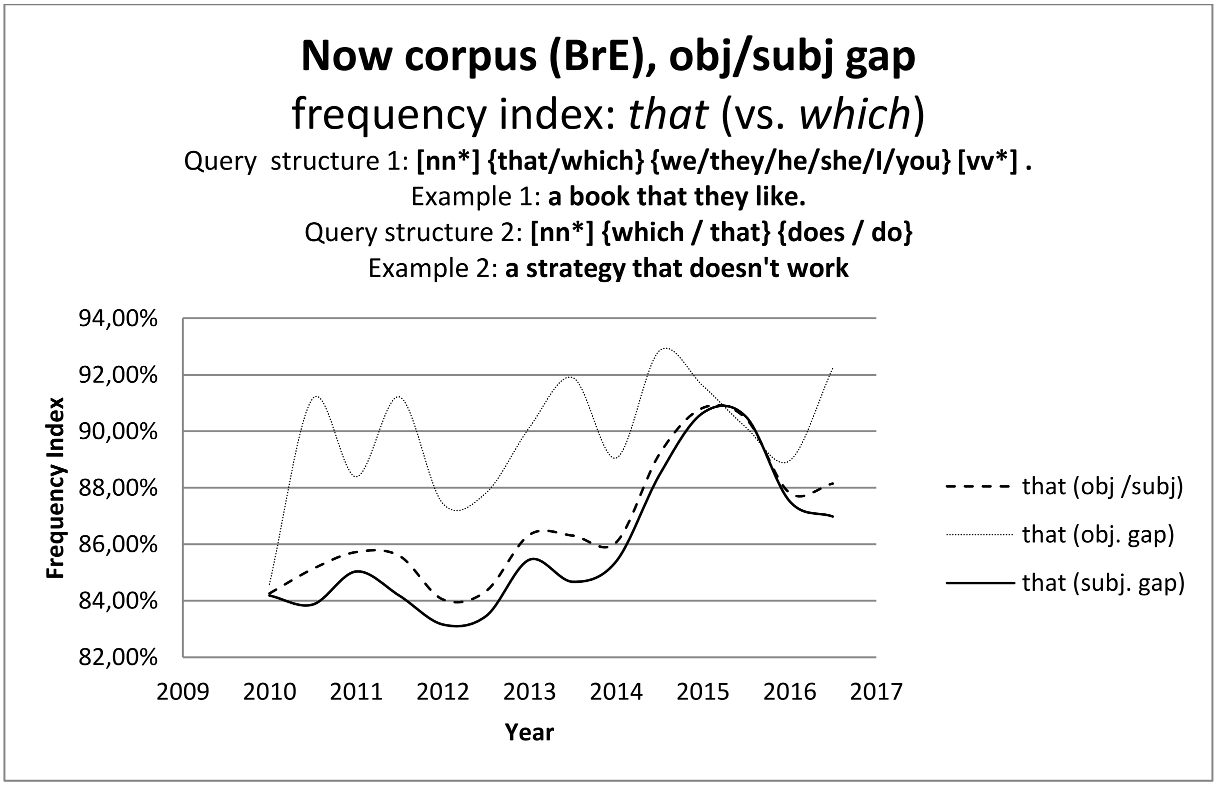


Chart 4. Diachronic development showing the usage of *that* (vs. restrictive *which*) between 2010 until today in British English web-based news. Restrictive *that* seems to be almost as frequent in British English as in American English (compare Chart 3).



It was found that the queries used for the TIME Magazine corpus had to be modified to work with the Now corpus, which has been described in section 3.2. The comparison with the original queries was done in the TIME Magazine corpus. The comparison showed that the differences in frequency index between the modified queries and the original queries are almost negligible (see Appendix 3, chart A3.4) and we may therefore conclude that the modified queries are adequate.

4.2 Inclusion and Exclusion

In section 4.1, many of the charts reflect results for both constructions with object and subject gaps. We observe in these charts that the differences in terms of frequency index between object gap and subject gap constructions are small. Studying frequencies of inclusion factors (object or subject gap) in different corpora shows that subject gap was by far most common (see Table 3).

Table 3. Inclusion factors for different corpora with the ratios between the estimated frequencies of the specified gap types and the total estimated frequencies stated as percentage values.

Inclusion factor	Brown Family[10]	Now (AmE)	Now (BrE)	Time Magazine
Obj. gap	6%	19%	20%	<1%
Subj. gap	94%	81%	80%	>99%
total	100%	100%	100%	100%

Examples of relevant entries from the TIME Magazine corpus are (10), with object gap and (11) and (12), which have subject gap. Please note that underscores have been added in all examples in this subsection where they appear. Also, note that all numbered corpus examples in this essay, such as (10), (11) and (12), have been listed in Appendix 1.1 (table A1.1.1), with information about corpus, retrieval date, etc. for each example. For the original queries, please refer to tables A1.2.1 and A1.2.2, which are linked to table A.1.1.1 by sample ID.

(10) […] the Government does not obtain the collaboration which it requests _ . (table A1.1.1)

(11) The nation that _ calls itself the West Indies is only ten months old […] (table A1.1.1)

(12) the state legislation is giving an extra push to experiments that _ were already successfully under way . (table A1.1.1)

As described in section 3.2, we had to design separate queries for different pronoun cases in the Now corpus. Example (13), with underscore added, was found with the query pattern for object gap *noun* + *that* + *we* + *verb* + *full stop* in the Now corpus for British English.

(13) Winkle Periwinkles are the smallest sea snails that we eat _. (table A1.1.1)

(14) is an example of an included entry with some optional elements, namely *can not* and *them*. Incidentally, (14) also exemplifies a ditransitive construction with indirect object (*them*) and direct object gap. Constructions like (14) were consistently included but only occurred with

[10] The proportions turned out to be almost equivalent for the American and British sub corpora of the Brown family; 93,96% (AmE) and 93,99% (BrE) for subject gap queries.

the Brown corpora queries, which had been made inclusive enough to match such constructions (see section 3.2; Appendix 1.3 (30)).

(14) […] equipment which others may have but which you can not give _ them. (table A1.1.1)

In the remaining part of this section, we will present and discuss our findings on exclusion factors. For subject gap queries, almost all entries were relevant in all corpora. Among the entries that were not, the exclusion categories were too diverse to pinpoint one specific factor as most common. All such exclusions were classified as *miscellaneous* (see Table 2, page 12). An example is (15) from the Brown corpus (AmE) where the pattern for subject gap (*noun* + *that* + *verb*) is satisfied. Still, there is no such gap as *that is* has a different meaning (same as the abbreviation *i.e.).* Another example from the same corpus of an excluded entry is (16) where there appear to be a subject gap (_) followed by verb omission (~~v~~). Similar constructions were consistently excluded.

(15) […] the government by force and violence; the British government that is. (table A1.1.1)

(16) The elements that _ did ~~v~~ were the introspective slow movement (table A1.1.1)

Table 4. The most usual exclusion factors in the different corpora (cf. Table 2). The overall exclusion ratio (see section 3.2) is given in percentage format for each exclusion factor and query set category listed in the table. Instances of high exclusion percentages for the Brown family corpora (30-70%) are most likely due to the fact that object gap queries had been made much more inclusive in those cases (see section 3.2).

Query Structure	**Brown Family (AmE)**	**Brown Family (BrE)**	**Now (AmE)**	**Now (BrE)**	**Time Magazine (AmE)**
Obj. gap (that)	Complement clause (67%)	Adverbial Gap (37%)	Adverbial Gap (2%)	Complement clause (4%)	Complement clause (3%)
Obj. gap (which)	Adverbial gap (62%)	Adverbial gap (57%)	N/A	N/A	N/A
Subj. gap (that)	Miscellaneous (<1%)	Miscellaneous (<1%)	Miscellaneous (2%)	N/A	N/A
Subj. gap (which)	N/A	N/A	Miscellaneous (<1%)	N/A	N/A

The most common exclusion factors for different corpora have been summarized Table 4. In the following, these will be exemplified and discussed. An example of complement clause from the Brown corpus is given in (17). The clause, *it must compete*, forms an independent clause with subject and predicate verb.

(17) The anti-trust laws inform a business that it must compete, […] (table A1.1.1)

Similar constructions from the Now corpus have been exemplified in (18) and (19). Like in (17), the *that*-complement clause constitutes a gapless independent clause in each one of these examples. *that* is a complementizer as defined in section 1.

(18) […] Nubia told the investigator that she fell . (table A1.1.1)

(19) […] it 's kind of a miracle that he graduated . (table A1.1.1)

Adverbial gap (and not complement clause) was found to be the most frequent exclusion factor in the American English section of the Now corpus for object gap queries with *that*. In (20), there is an adverbial gap referring to a particular occasion.

(20) they had the job of watching Slagle on the night that he died_. (table A1.1.1)

With two exceptions, complement clause was the typical exclusion factor for object gap queries with *that*. One exception was the American English section of the Now corpus, in which adverbial gap was most common. Some examples from such constructions are given in (21), (22) and (23). The constructions appear to be colloquial. The higher frequencies could therefore be explained by higher colloquial tendencies in American English and news that is web-based.

(21) […] in the way that he thinks. (table A1.1.1)

(22) […] the day that he died. (table A1.1.1)

(23) […] the way that he practices. (table A1.1.1)

Exclusion for *which* queries (object gap) were only found in the Brown family of corpora, most commonly due to adverbial gap. It may be explained by the fact that these queries had been made more inclusive (see section 3.2). Constructions with adverbial gap often matched the pattern *preposition* + *which* + *subject* + *verb* + *adverbial gap*, such as (24). The queries in the other corpora do not match this pattern, as they require a noun to precede *which*.

(24) […] the period in which they lived. (table A1.1.1)

Among the exclusion factors listed in Table 2 (page 12), only the most usual ones have been addressed in this sub-section. For examples of less common exclusion factors, see Appendix 1.1 (table A1.1.2).

4.3 Explanations: analyses of entries from the H.U.G.E. database

In section 3.2 we described our methods to collect entries from an online database of usage guides named the H.U.G.E database (English Usage Guide Database, 2017) and how these were analyzed. In this sub-section, the results of these analyses will be presented and discussed.

Chart 5. Entries by American authors in favor of using *that* only rather than *which* as relativizer in restrictive relative clauses (compare Chart 6).

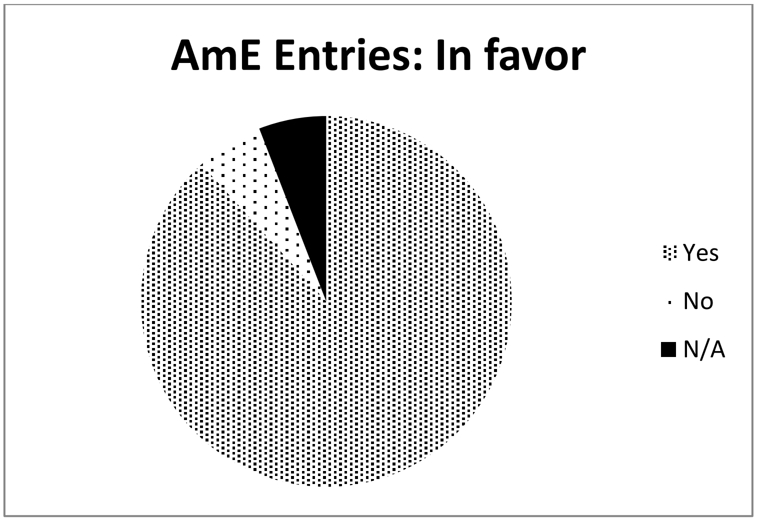


Chart 6. Entries by British authors in favor of only using *that* as relativizer (compare Chart 5).

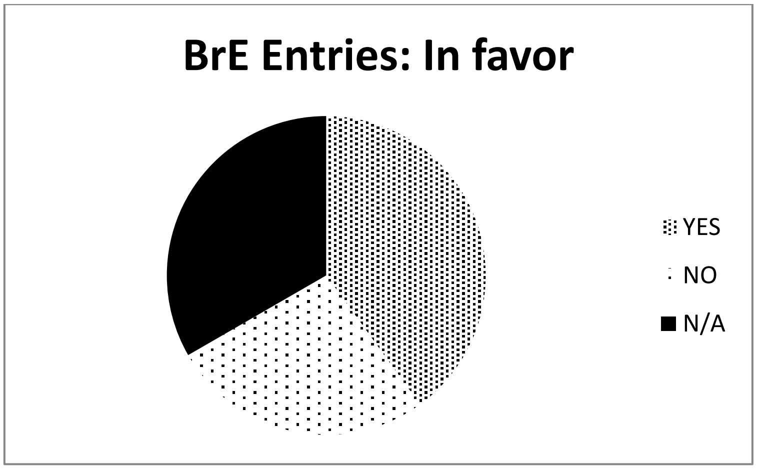


Chart 5 and Chart 6 show what proportion of the entries were in favor of using *that* (but not *which*) as restrictive relativizer in British English and American English. As we may see from these charts, entries from American usage guides tend to be much more strongly in favor of such usage compared to British counterparts.

These data indicate that such sources essentially agree about not using *which* as restrictive relativizer. It could therefore be reasonably assumed that usage guides such as these have played an important role in forming the opinion among educators and editors that only *that* and not *which* are legitimate as relativizers in restrictive clauses.

Undeniably, a limitation of our H.U.G.E. database study is its subjective nature, i.e. its reliance on the essay author's interpretations of different entries. The possibility of personal bias or misinterpretation having influenced the outcome cannot be dismissed. The influence of such factors could likely have been reduced with several judges evaluating the data independently. Such extended evaluation, however, was considered outside the scope of this study.

5. Conclusion

In this paper, we have explored diachronic developments of restrictive *which* and *that* in American and British English news. Some earlier studies along with several corpora have provided the data. Potential explanations have also been studied, mainly by consulting a database of usage guide. As our study only considers entries with either subject or object gap and excluded other cases, we also found it relevant to examine reasons for exclusion of entries.

Four initial hypotheses were formulated in section 1. We were right about the first hypothesis. *That* as restrictive relativizer in news appears to have been increasing in modern times, especially in American English. Based on our TIME Magazine corpus data, the frequency index for *that* (compared to restrictive *which*) increased by about 70 percentage points between 1930 and 2000 in American English. Studying web-based news from the Now corpus suggests that the increasing trend continued between 2010 to 2016 and today virtually all (about 97%) of the constructions in American English with either restrictive *which* or *that* use *that*. It is here worth mentioning that the final data from TIME Magazine corpus reflect even higher frequency indexes for *that* (>99%) around the turn or the millennium.

As for British English, there has not been much change before the turn of the millennium. Based on our Brown family corpora study, a slight increase in frequency index of restrictive *that* took place between 1961 and 1991/1992 (from 40 to 41 percentage points). Based on Biber et al.'s data (1999, p. 616, table 8.9), the number increased by about 4 percentage points until the turn of the millennium to approximately 45 percentage points. It is worth noting that restrictive *that* appears to have been an underdog in British English ever since 1961. By the year of 1999, restrictive *which* was still more frequently used than *that* in British newspaper-based news, with *that* only being used in 46% of the restrictive clauses using either *that* or *which* (Biber et al., 1999, p. 616).

What is even more surprising is the data for British English news from the Now corpus, which give a completely different perspective in comparison to the data from the Brown family of corpora and the data from Biber et al. (1999, p. 616). According to the Now corpus data, *that* had, in fact, a much stronger position in web-based news than restrictive *which* in 2010, with a frequency index of about 85%. By 2010, this value had increased to 92%. While the corresponding tendencies for American English were found to be even stronger, the numbers are still surprising. We can only speculate what the reasons may be. One factor could perhaps be stronger colloquial associations of *that* interfering with presumably colloquial tendencies in news that is web-based. Another explanation might be that writers and editors of British web-based news may have been influenced by American prescriptivism to a much higher degree than editors of newspaper-based news.

Our second hypothesis turned out to be accurate (with reservation for shortcomings discussed in section 4.3). Our study from the H.U.G.E database showed that book entries in usage guides by American authors generally favored a binary distribution between *that* in restrictive and *which* in non-restrictive clauses. In fact, it was found that 88% of relevant entries in American usage guides recommended this distribution. For British counterparts, the corresponding figure was 29%. Accordingly, it seems as though American prescriptivism indeed has contributed to the developments described above.

The third hypothesis, which stated that object gap would be more common than subject gap among relevant data entries, was incorrect. Contrariwise, subject gap was found to be a much more common reason for inclusion in all corpora studied. The largest difference was found in the TIME Magazine corpus with more than 99% of the relevant entries having subject gaps (and less than 1% object gaps).

The fourth hypothesis, according to which other gap functions would be the most common exclusion factor, was accurate only in part. For restrictive *which* queries with object gap, it turned out to be true, typically due to presence of adverbial gap. For object-gap queries with *that*, however, complement clause was found to be the predominant exclusion category except for two cases. The exceptions were American English web news and the British English sub corpora of the Brown family. In both these cases, adverbial gap was demonstrated be most common. For subject gap queries, the numbers of exclusions were small and the exclusion factors diverse. No single factor could be established as the most important one for these queries.

As pointed out in the introduction, this essay has a methodological aim for analyzing sizeable query sets. Examination of random subsets (samples) from these sets has made it possible to make estimations of frequencies of relevant and non-relevant entries. It has also helped us to estimate ratios of different exclusion factors. Indeed, our methodology has limitations, such as the absence of analysis of statistical deviations. It would for example be valuable to know how much the estimated relevance index based on a random sample could be expected to deviate from the actual relevance index for the query set. Devising a method for obtaining such information and possibly further improve other aspects of our methodology could be an area for future studies to explore.

References

Aarts, B. (2011). *Oxford modern English grammar.* Oxford : Oxford University Press, 2011.

Biber D., Conrad S. and Leech G (2002). *Longman student grammar of spoken and written English*. Harlow : Longman

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow : Longman, 1999.

Corpuscle (2017) ICAME – BROWN Family, Retrieved from http://clarino.uib.no, 2017

Collins *COBUILD English grammar*. (2011). Glasgow : HarperCollins Publishers, 2011.

Danesi M. (2006). *Basic American Grammar and Usage : An ESL / EFL Handbook*. New York : Barron’s Educational Series, Inc.

English Usage Guide Database (2017). Retrieved from http://huge.ullet.net/, 2017

Jordan, L. (1976). *The New York Times Manual of Style and Usage*. New York: The New York Times Company, 1976

Leech, G., Hundt M., Mair C. and Smith N., (eds). 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: CUP

Leech, G. and Smith N. *Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991* In Renoulf A., Kehoe A. (eds). 2009. Corpus Linguistics: Refinements and Reassessments. Amsterdam : Editions Rodopi B.V., 2009

Mair, Christian. 2006. *Twentieth-Century English: History, Variation, and Standardization*. Cambridge: CUP

Now corpus (2017), Retrieved from http://corpus.byu.edu/now/, 2017

Pinker, S. (2015). *The sense of style : the thinking person's guide to writing in the 21st century*. New York : Hudson Street, 2015

Strunk, W. (1959). *The Elements of Style*: New York: The Macmillan Company

The Chicago manual of style. (2003). Chicago, Ill. : University of Chicago Press, cop. 2003.

Thurman, S. (2002). *The Only Grammar Book You'll Ever Need* : Avon: F+W Media, Inc., 2002.

TIME Magazine corpus (2017), Retrieved from http://corpus.byu.edu/time/, 2017

Appendix 1: Corpora, samples, queries, databases

Appendix 1.1: examples from corpora

Table A1.1.1 in this appendix contains a list of all corpus examples referred to in the text. For URL addresses to the corpora, please refer to table A1.1.3 in this appendix. Please note that *date* refers to date of retrieval. “Sample ID” is a unique ID number identifying the sample from which the entry has been taken.

Table A1.1.1. All numbered corpus examples referred to in the running text. See table A1.1.3 for retrieval URL addresses for the corpora specified in the *Corpus / (Sample ID)* column. Underscores have been added by us to mark gaps.

Example number + example + query type[11]	**Corpus / (Sample ID)**	**Retrieved Date**	**Inclusion and exclusion factors (exclusion factors *italicized*).**
(10) […] the Government does not obtain the collaboration which it requests _ . ^{O}	TIME / (31)	2-Mar-17	Object gap
(11) The nation that _ calls itself the West Indies is only ten months old […] ^{S}	TIME / (37)	2-Mar-17	Subject gap
(12) the state legislation is giving an extra push to experiments that _ were already successfully under way ^{S} .	TIME / (37)	2-Mar-17	Subject gap
(13) Winkle Periwinkles are the smallest sea snails that we eat _ . ^{O}	Now (BrE) / (1)	2-Mar-17	Object gap
(14) […] equipment which others may have but which you can not give _ them. ^{O}	Brown family / (20003)	12-Mar-17	Object gap
(15) […] the government by force and violence; the British government that is. ^{S}	Brown family / (20009)	11-Mar-17	*Miscellaneous*
(16) The elements that _ did ~~v~~ were the introspective slow movement ^{S}	Brown family / (20009)	11-Mar-17	*Miscellaneous*
(17) The anti-trust laws inform a business that it must compete, […] ^{O}	Brown family / (20001)	11-Mar-17	*Complement clause*
(18) […] Nubia told the investigator that she fell . ^{O}	Now (BrE) / (15)	2-Mar-17	*Complement clause*
(19) […] it 's kind of a miracle that he graduated . ^{O}	Now (AmE) / (67)	12-Mar-17	*Complement clause*
(20) they had the job of watching Slagle on the night that he died_. ^{O}	Now (AmE) / (67)	12-Mar-17	*Adverbial gap*
(21) […] in the way that he thinks. ^{O}	Now (AmE) / (67)	12-Mar-17	*Adverbial gap*
(22) […] the day that he died. ^{O}	Now (AmE) / (67)	12-Mar-17	*Adverbial gap*
(23) […] the way that he practices. ^{O}	Now (AmE) / (67)	12-Mar-17	*Adverbial gap*
(24) […] the period in which they lived. ^{O}	Brown family / (20002)	12-Mar-17	*Adverbial gap*

[11] Coding for query type is superscripted O ('object gap query) and S ('subject gap query').

Table A1.1.2. Corpus examples for less common exclusion categories (cf. Table 2 on page 12). Please see table Table A1.1.3 for URL addresses of corpora given in the *Corpus / Retrieved Date* column. Underscores have been added by us, marking the positions of the gaps. Further note that all entries in this table are related to object gap queries.

Corpus Example + query type	Exclusion Factors	Corpus / Retrieved Date	Sample ID
I have never mentioned a new artist that Thompson didn't know about _ .	PP-Complement gap	Brown family / Mar-11-17	20001
[…] it includes many measures that Bush has called for _ […]	PP-Complement gap	Brown family/ Mar-11-17	20003
The President had set for himself the task, which he believed vital, […]	Non-Restrictive clause	Brown family/ Mar-12-17	20002
[…] helped turn her into the powerhouse that she became _ .	Subject Predicative Gap	Now (BrE) / Mar-02-17	15
[…] global superstar , into this gorgeous bundle of trouble that she became _ .	Subject Predicative Gap	Now (AmE) / Mar-12-17	71

Table A1.1.3. URL addresses for different corpora.

Full corpus name	URL
TIME Magazine corpus	http://corpus.byu.edu/time/
Brown family of corpora	http://clarino.uib.no/korpuskel/clarino-metadata?session-id=242536023026472&corpus=brown&default-corpus=brown&resource=brown
Now corpus	http://corpus.byu.edu/now/

Appendix 1.2: sample data

Table A1.2.1 in this appendix lists all relevant samples in the Now corpus and TIME Magazine corpus. Table A1.2.2 lists all relevant samples from the Brown family of corpora. It should be noted that Brown, Frown, Lob and Flob refer to sub-corpora of the ICAME Brown family corpus.

Wherever total size (i.e. size of the query set) is greater than sample size, the frequencies of relevant entries are estimations, calculated by multiplying total frequencies with relevance index. In these cases, the samples were randomly selected from the total set of entries for a query. Also, note that total size occasionally may be larger than sample size depending on the fact that the current year (2017) was excluded from the Now corpus. The queries have been presented as they were input in the corpora. Please refer to the corpora websites (see table A1.1.3) for detailed descriptions related to query syntax. Some explanations of query syntax have also been included in Appendix 1.3.

Table A1.2.1. Samples from the Brown family of corpora.

Corpus	variety	Retrieved Date	Rel. index	To get the exact query, replace {R} and {C} (if they exist) in (30) on page 34 with values indicated here	Query type[12]	Sample size (total)	Sample ID
Brown	AmE	11-Mar-17	8%	{R}=that; {C}=brown	O	52 (52)	20001
Brown	AmE	12-Mar-17	33%	{R}=which; {C}=brown	O	40 (40)	20002
Frown	AmE	11-Mar-17	34%	{R}=that; {C}=frown	O	91 (91)	20003
Frown	AmE	12-Mar-17	14%	{R}=which; {C}=frown	O	29 (29)	20004
Lob	BrE	12-Mar-17	15%	{R}=that; {C}=lob	O	78 (78)	20005
Lob	BrE	12-Mar-17	30%	{R}=which; {C}=lob	O	44 (44)	20006
Flob	BrE	12-Mar-17	17%	{R}=that; {C}=flob	O	48 (48)	20007
Flob	BrE	12-Mar-17	29%	{R}=which; {C}=flob	O	38 (38)	20008
Brown	AmE	12-Mar-17	98%	[pos = "SUBST"] "that" [pos = "VERB"] :: subcorpus = "brown" & genre = "press"	S	100 (277)	20009
Brown	AmE	12-Mar-17	100%	[pos = "SUBST"] "which" [pos = "VERB"] :: subcorpus = "brown" & genre = "press"	S	100 (147)	20010
Frown	AmE	11-Mar-17	100%	[pos = "SUBST"] "that" [pos = "VERB"] :: subcorpus = "frown" & genre = "press"	S	100 (382)	20011
Frown	AmE	12-Mar-17	100%	[pos = "SUBST"] "which" [pos = "VERB"] :: subcorpus = "frown" & genre = "press"	S	9 (9)	20012
Flob	BrE	11-Mar-17	100%	[pos = "SUBST"] "that" [pos = "VERB"] :: subcorpus = "flob" & genre = "press"	S	100 (143)	20013
Flob	BrE	12-Mar-17	100%	[pos = "SUBST"] "which" [pos = "VERB"] :: subcorpus = "flob" & genre = "press"	S	100 (203)	20014
Lob	BrE	12-Mar-17	99%	[pos = "SUBST"] "that" [pos = "VERB"] :: subcorpus = "lob" & genre = "press"	S	100 (136)	20015
Lob	BrE	12-Mar-17	100%	[pos = "SUBST"] "which" [pos = "VERB"] :: subcorpus = "lob" & genre = "press"	S	100 (208)	20016

Table A1.2.2. Samples from the Now corpus and TIME Magazine corpus.

Corpus	variety	Retrieved Date	Relevance index	Query + type[13]	Sample Size	Total Size	Sample ID
Now	BrE	2-Mar-17	97%	[nn*] that we [vv*] . ^{O}	100	804	1
Now	BrE	2-Mar-17	100%	[nn*] which we [vv*] . ^{O}	45	45	3

[12] Coding for query type is O ('object gap query) and S ('subject gap query').

[13] Coding for query type is superscripted O ('object gap query) and S ('subject gap query').

Now	BrE	2-Mar-17	94%	[nn*] that I [vv*] . ^{O}	99	653	5
Now	BrE	2-Mar-17	100%	[nn*] which I [vv*] . ^{O}	62	62	7
Now	BrE	2-Mar-17	87%	[nn*] that he [vv*] . ^{O}	100	565	11
Now	BrE	2-Mar-17	100%	[nn*] which he [vv*] . ^{O}	100	120	13
Now	BrE	2-Mar-17	83%	[nn*] that she [vv*] . ^{O}	100	233	15
Now	BrE	2-Mar-17	100%	[nn*] which she [vv*] . ^{O}	52	47	17
Now	BrE	2-Mar-17	91%	[nn*] that they [vv*] . ^{O}	100	972	19
Now	BrE	2-Mar-17	100%	[nn*] which they [vv*] . ^{O}	92	89	21
Now	BrE	2-Mar-17	94%	[nn*] that you [vv*] . ^{O}	100	387	23
Now	BrE	2-Mar-17	100%	[nn*] which you [vv*] . ^{O}	7	7	25
TIME	AmE	2-Mar-17	100%	[n*] which [p*] [vv*] . ^{O}	100	174	31
TIME	AmE	2-Mar-17	97%	[n*] that [p*] [vv*] . ^{O}	100	475	33
TIME	AmE	2-Mar-17	100%	[n*] which [v*] ^{S}	100	35469	35
TIME	AmE	2-Mar-17	100%	[n*] that [v*] ^{S}	100	161967	37
TIME	AmE	3-Mar-17	100%	[n*] which does ^{S}	100	158	41
TIME	AmE	3-Mar-17	100%	[n*] that does ^{S}	100	734	43
TIME	AmE	3-Mar-17	100%	[n*] which do ^{S}	100	120	45
TIME	AmE	3-Mar-17	100%	[n*] that do ^{S}	100	581	47
Now	BrE	3-Mar-17	100%	[nn*] which does ^{S}	97	1156	51
Now	BrE	3-Mar-17	100%	[nn*] that does ^{S}	83	7283	53
Now	BrE	3-Mar-17	100%	[nn*] which do ^{S}	98	926	55
Now	BrE	3-Mar-17	99%	[nn*] that do ^{S}	99	5865	57
Now	AmE	12-Mar-17	100%	[nn*] that we [vv*] . ^{O}	51	2238	59
Now	AmE	12-Mar-17	100%	[nn*] which we [vv*] . ^{O}	5	7	61
Now	AmE	12-Mar-17	98%	[nn*] that I [vv*] . ^{O}	100	1131	63
Now	AmE	12-Mar-17	100%	[nn*] which I [vv*] . ^{O}	34	37	65
Now	AmE	12-Mar-17	90%	[nn*] that he [vv*] . ^{O}	100	807	67
Now	AmE	12-Mar-17	100%	[nn*] which he [vv*] . ^{O}	44	44	69
Now	AmE	12-Mar-17	94%	[nn*] that she [vv*] . ^{O}	96	339	71
Now	AmE	12-Mar-17	100%	[nn*] which she [vv*] . ^{O}	14	13	73
Now	AmE	12-Mar-17	98%	[nn*] that they [vv*] . ^{O}	99	1469	75
Now	AmE	12-Mar-17	100%	[nn*] which they [vv*] . ^{O}	43	42	77
Now	AmE	12-Mar-17	96%	[nn*] that you [vv*] . ^{O}	89	1027	79
Now	AmE	12-Mar-17	100%	[nn*] which you [vv*] . ^{O}	14	13	81
Now	AmE	12-Mar-17	100%	[nn*] which does ^{S}	98	527	83
Now	AmE	12-Mar-17	99%	[nn*] that does ^{S}	100	12807	85
Now	AmE	12-Mar-17	99%	[nn*] which do ^{S}	96	448	87
Now	AmE	12-Mar-17	97%	[nn*] that do ^{S}	92	15937	89

Appendix 1.3: Some explanation on query syntax

All queries that have been used for this study are listed in Appendix 1.2. The current appendix contains some explanations of query syntax applied in selected examples, but it is by no means a complete reference guide. The reader who wants to learn more about the query syntax of the corpora in our study is recommended to consult the corpora websites, whose URL addresses may be found in table A1.1.3 (Appendix 1.1).

Query (25) (Now corpus, table A1.2.2, ID 1) contains two tags that need to be explained. *[nn*]* represents a slot for common nouns (not proper nouns). *[vv*]* refers to lexical verbs (excluding the auxiliary verbs *be*, *have* and *do*) (Now corpus, 2017). An example of a possible construction matching (25) would be *the book that we like*. Please note that while *[vv*]* was suitable for object gap queries, where transitive verbs were targeted, a similar form *[v*]*, which includes all verb forms, was convenient for subject gap queries, in which all verb types could be useful, as shown in (26).

(25) [nn*] that we [vv*] .

(26) [n*] that [v*]

Query (27) (TIME Magazine corpus, table A1.2.2, ID 31) contains the noun tag *[n*]*, which includes both common and proper nouns. *[p*]* refers to pronouns. Note that the query ends with full stop (.). A possible example would be *the book which I like.*

(27) [n*] which [p*] [vv*] .

In (28) (Now corpus, table A1.2.2, ID 83), negations are included as well. Possible constructions matching (28) could be *a strategy which doesn't work* or *a strategy which does not work.*

(28) [nn*] which does

Compared to the Now and TIME Magazine corpora, the query syntax for the ICAME Brown family of corpora (Corpuscle, 2017) was more complex, especially for object gap queries. In

(29), *[pos = "SUBST"]* represents both common nouns and proper nouns. *[pos = "VERB"]* comprises all verbs, including functional verbs such as auxiliary verbs and linking verbs.

(29) [pos = "SUBST"] "that" [pos = "VERB"] :: subcorpus = "brown" & genre = "press"

Our object gap query for the Brown family of corpora turned out to be rather complex for reasons explained in section 3.2. Its raw format, which is also included in table A1.2.1, is shown in (30) with **{R}** replaced by either *that* or *which*, and **{C}** by sub-corpus (*brown*, *frown*, *lob* or *flob*). Some, but not all, of the syntactical elements in (30) will be explained below. For a complete reference, refer to documentation (Corpuscle, 2017)**.**

(30) "**{R}**" ([morph= "NP1"] | [pos = "PRON"]|("a" [pos = "SUBST"]|"the" [pos = "SUBST"])) ([morph= "VV."] | ".*" [morph= "VV."] | ".*" ".*" [morph= "VV."]) ([pos = "STOP"] | "and" | "or" |".*" [pos = "STOP"]|".*" ".*" [pos = "STOP"]) :: subcorpus = "**{C}**" & genre = "press"

Due to its complexity, a simplification of (30) may be needed for the reader to understand the essential meaning of it. In table A1.3.1, a schematic structure of (30) is shown as well as some hypothetical examples. The rows in the query structure section of (30) should be understood as different options of elements that may occupy a certain position. In the table, the symbol [] represents any word. Such optional words have been italicized in table A1.3.1 under *hypothetical examples*.

Table A1.3.1. Schematic structure of the object gap query in (30) with some hypothetical examples given. In the *query structure* part, [] represents optional elements and full stop(.) may also represent comma (,). Optional words corresponding to the symbol [] have been italicized.

query structure	that	proper noun	verb	.
	which	pronoun	[] verb	and
		a + noun	[] [] verb	or
		the + noun		[] .
				[] [] .
hypothetical examples	that	he	sleeps	*too much*.
	that	John	*shall* remain	*suspended*,
	(by) which	the students	*must* be	*guided*.

In (30), *[morph= "NP1"]* represents a slot for a proper noun, but the same slot may also be taken by a pronoun (*[pos = "PRON"]*) or a noun phrase with either *a* or *the* as the determiner. The symbol | is used as delimiter between different options that may occupy the same slot, and all such elements must be enclosed within brackets. The expression ".*" represents an optional element. *[morph= "VV."]* represents a verb, including functional verbs. *[pos = "STOP"]* stands for a punctuation including comma (,) or full stop (.).

Appendix 1.4: H.U.G.E database data

This appendix contains specific information on the data from the H.U.G.E database that served as basis for Chart 5 and Chart 6 on page 24. All entries were retrieved from the interface available from the H.U.G.E database (English Usage Guide Database, 2017). Table A1.4.1 lists all the American entries (with table A1.4.2 containing general information such as retrieval date). Table A1.4.3 lists all the British entries (with table A1.4.4 containing the same type of general information as in A1.4.2 mentioned above). Please note that the column *year* is sometimes ambiguous in tables A1.4.1 and A1.4.3, for example in entry with ID 20017 where it may be either 1999 or 2000. An explanation of this ambiguity has been included in section 3.2.

Table A1.4.1. List of entries from American English usage guide entries.

Entry ID	Year	Author(s)	nationality of author	in favor of binary distribution	General info (e.g. retrieval date and URL)
10000	1998	O'Conner, Patricia	US	yes	See table A1.4.2
10001	1993	Wilson, Kenneth G.	US	no	See table A1.4.2
10002	2008	Fogarty, Mignon	US	N/A	See table A1.4.2
10003	2004	Batko, Ann	US	yes	See table A1.4.2
10004	2004	Batko, Ann	US	yes	See table A1.4.2
10005	1991	De Vries, Mary Ann	US	yes	See table A1.4.2
10006	1981	Vermes, Jean C.	US	yes	See table A1.4.2
10007	1984	Bryson, Bill	US	yes	See table A1.4.2
10008	1938	Turck Baker, Josephine	US	no	See table A1.4.2
10009	1992	Booher, Dianna	US	N/A	See table A1.4.2
10010	2003	Brians, Paul	US	yes	See table A1.4.2
10011	1920	Vizetelly, Frank H.	US	yes	See table A1.4.2
10012	1088	Randall, Bernice	US	N/A	See table A1.4.2
10013	1911	Ayres, Alfred	US	yes	See table A1.4.2
10014	1966	Follett, Wilson	US	N/A	See table A1.4.2
10015	1993	Mager, Nathan H.; Mager, Sylvia K.; Domini, John	US	yes	See table A1.4.2
10016	1993	Wilson, Kenneth G.	US	yes	See table A1.4.2
10017	1957	Evans, Bergen; Evans, Cornelia	US	yes	See table A1.4.2
10018	1975	Morris, William; Morris, Mary	US	yes	See table A1.4.2
10019	1998	Garner, Bryan A.	US	yes	See table A1.4.2
10020	1978	Ebbitt, Wilma R.; Ebbitt, David R.	US	yes	See table A1.4.2

Table A1.4.2. General information about the entries listed in table A1.4.1. Please note that retrieval date and URL are the same for all entries in that table.

Author nationality:	**United States**
Retrieval date:	18-Apr-17
Problem term:	that/which
URL:	http://huge.ullet.net/?content=search_ug

Table A1.4.3. List of entries from British English usage guide entries. Usage of slash (/) in the *year*-column indicates ambiguous cases as explained in section 3.2.

Entry ID	**Year**	**Author(s)**	**nationality of author**	**in favor of binary distribution**	**General info (e.g. retrieval date and URL)**
20000	1994	Blamires, Harry	UK	no	See table A1.4.4
20001	2010	Taggart, Caroline	UK	no	See table A1.4.4
20002	1922	Fowler, Henry Watson; Fowler, Francis George	UK	yes	See table A1.4.4
20004	2010	Heffer, Simon	UK	yes	See table A1.4.4
20005	2010	Lamb, Bernard C.	UK	yes	See table A1.4.4
20006	1994	Weiner, Edmund; Delahunty, Andrew	UK	N/A	See table A1.4.4
20008	2002	Ayto, John	UK	N/A	See table A1.4.4
20009	1980	Swan, Michael	UK	N/A	See table A1.4.4
20010	1999/2000	Fowler, Henry Watson; Burchfield, Robert W.	UK	no	See table A1.4.4
20011	1999/2000	Fowler, Henry Watson; Burchfield, Robert W.	UK	N/A	See table A1.4.4
20012	1926/1965	Fowler, Henry Watson	UK	yes	See table A1.4.4
20014	1926/1965	Fowler, Henry Watson	UK	yes	See table A1.4.4

20015	1926/1965	Fowler, Henry Watson	UK	no	See table A1.4.4
20016	1926/1965	Fowler, Henry Watson	UK	yes	See table A1.4.4
20017	1999/2000	Fowler, Henry Watson; Burchfield, Robert W.	UK	no	See table A1.4.4

Table A1.4.4. General information about the entries listed in table A1.4.3. Please note that retrieval date and URL are the same for all entries in that table.

Author nationality:	**United Kingdom**
Retrieval date:	2-Mar-17
Problem term:	that/which
URL	http://huge.ullet.net/?content=search_ug

Appendix 2: Definitions and calculations

Appendix 2.1: Calculation of frequency index

A definition of frequency index may be found in section 2. This sub-section contains a basic example how to calculate it. Assume we have a text with occurrences of *green* and *yellow* and that *green* occurs 2 times and *yellow* 6 times in this text. The frequency index for *green* (compared to *yellow*) would then be 2/(2+6)=0,25 (25%). The frequency index for yellow would be 6/(2+6)=0,75 (75%).

It should be noted that the total sum of the frequency indexes for the elements we compare, in this case *green* and *yellow*, always must be 1 (100%). Applied on the relations above, we have 2/(2+6) + 6/(2+6) = 0,25 + 0,75 = 1.

Appendix 2.2: Definition and calculation of relevance index

In this section, we briefly define relevance index and how to calculate it. Assume we would like to find occurrences of constructions that match the pattern verb + *water* and where *water* is a verb, which we must check manually. Suppose the query gives the hits listed in table A2.2.1.

Table A2.2.1. Constructions in a text matching the pattern verb + *water*.

no	Construction
1	I will water the flowers.
2	She poured water into the glass.
3	He must water the lawn.
4	We should water the plants.

We find that all constructions in table A2.2.1 are relevant, except for (2) *She poured water into the glass,* in which *water* occurs as a noun and not a verb. Thus, we have 3 entries out of four that are relevant. Relevance index is then calculated as 3 / 4 = 0,75 (75%).

Appendix 3: Additional Charts

Chart A3.1. Relative frequencies of *which* as relative pronoun between 1961 and 1991/1992 per million words, based on data from Leech et al. (2009, p. 309; table A10.10).

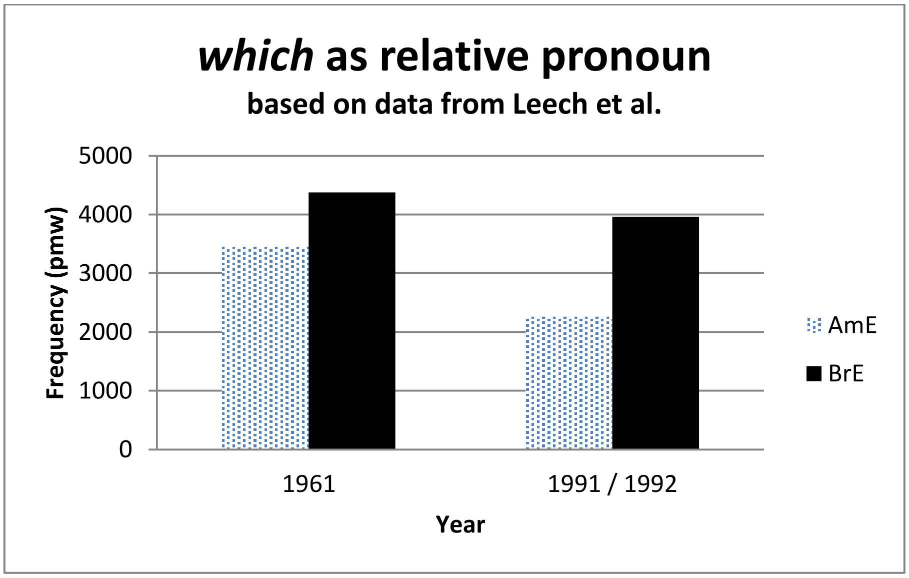


Chart A3.2. Relative frequencies of *that* as relative pronoun between 1961 and 1991/1992 per million words based on data by Leech et al. (2009, p. 309-310; table A10.11a & A10.11b).

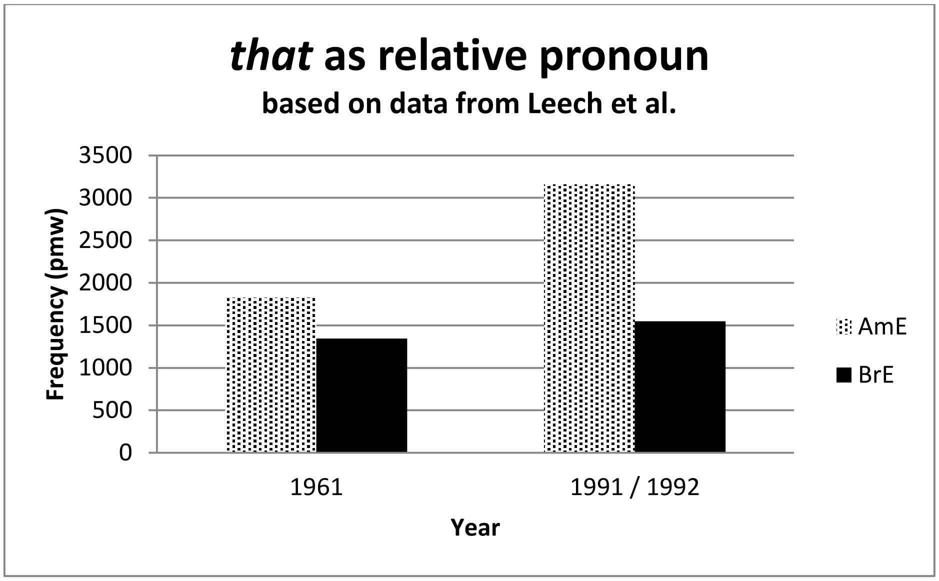


Chart A3.3. Chart, based on study conducted by Biber et al. (1999, p. 616), which reflects relative frequencies of restrictive *that* and *which* in British English and American English news.

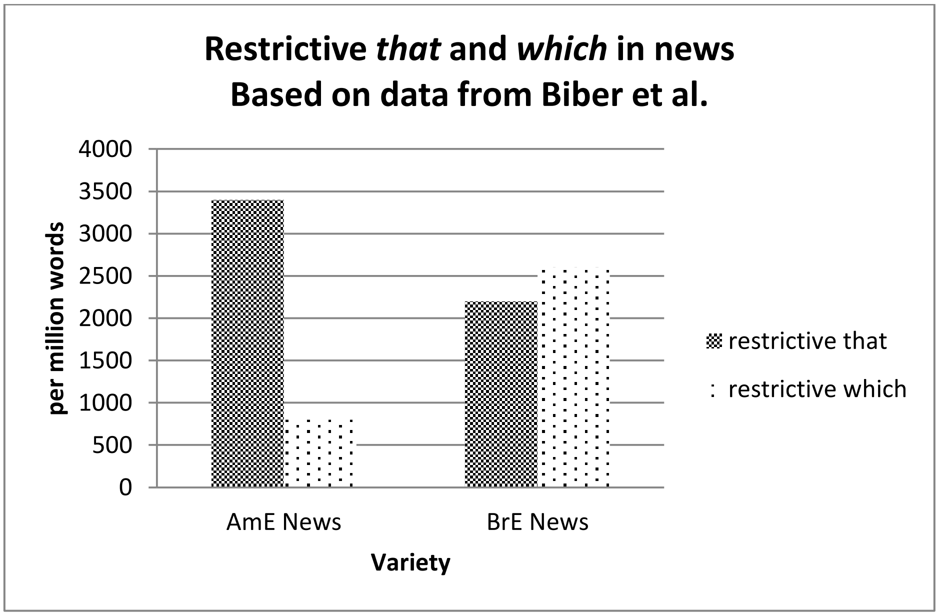
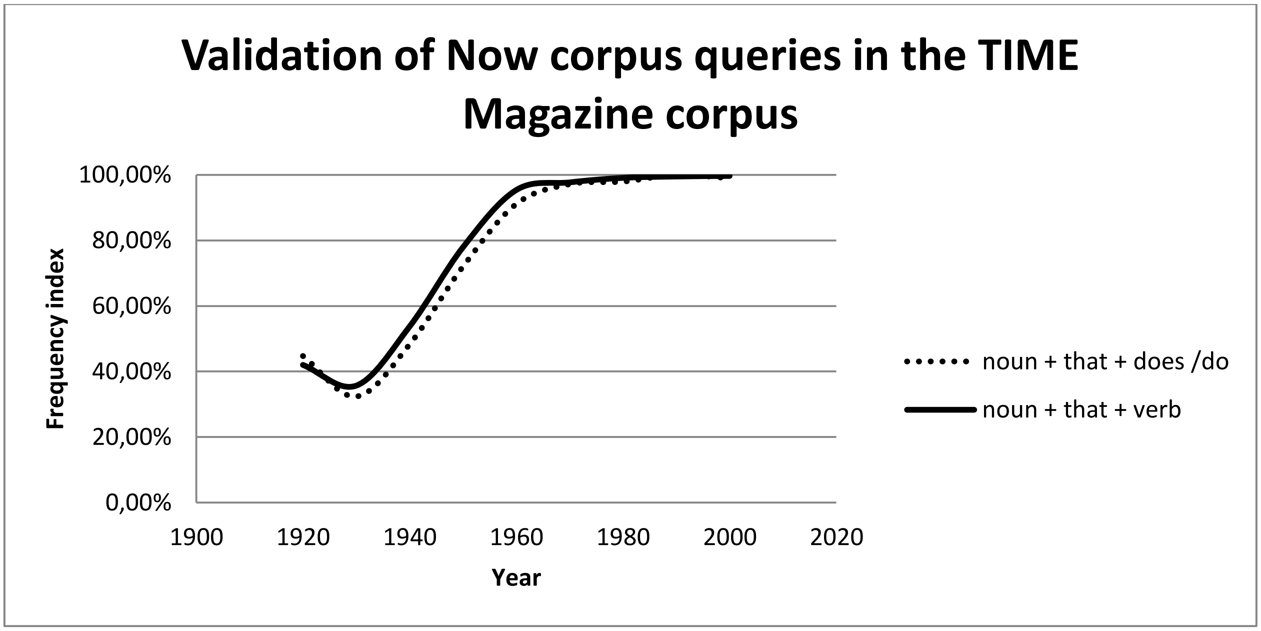


Chart A3.4. Chart comparing original queries (TIME Magazine corpus) and modified queries (Now corpus). Frequency index was measured for frequencies of *that* constructions compared to corresponding *which* constructions. Note the forms of the original queries (*noun + that/which + verb*) and the modified queries (*noun + that/which + does / do / doesn't / don't*). As shown by the chart, there are almost no differences between the modified queries and the original ones in terms of frequency index, so there seem to be virtually no interfering factors present.


Validation of Now corpus queries in the TIME Magazine corpus
Frequency index
Year
noun + that + does /do
noun + that + verb