THOU SHALT NOT SPLIT…?
A CORPUS-BASED STUDY ON SPLIT INFINITIVES IN AMERICAN ENGLISH

Simon Johansson

Projekt i engelska (15 hp)
Engelska 61-90 poäng
Fristående kurs, hösten 2014

Handledare: Mattias Jacobsson
Examinator: Jens Allwood
ABSTRACT

Simon Johansson

Thou Shalt Not Split…? A Corpus-Based Study on Split Infinitives in American English

Fall 2014

This essay aims to shed light on the prevalence of the to + adverb + verb and to not + verb split infinitives in American English, both in a historical perspective and in present day usage, and how it varies in different contexts where different levels of formality are expected. Although students are taught to avoid splitting constructions, numerous grammarians and linguists question this prescriptive viewpoint. Two extensive corpora, the Corpus of Historical American English (COHA) and the Corpus of Contemporary American English (COCA), were used to gather data. The results revealed how the frequency of the split infinitive was, and still is, rising rapidly, and becoming more and more a standard and accepted feature in American English. The most common context in which to find a split infinitive was that of informal spoken language. However, it was in the most formal of settings, that of academic texts, where the largest increase in prevalence of the split infinitive was seen.

Key words: split infinitive, corpus linguistics, COHA, COCA, American English
# Table of Contents

1. INTRODUCTION....................................................................................................................... 1  
   1.1 Aim........................................................................................................................................ 2  
   1.2 Material and Method........................................................................................................... 2  

2. BACKGROUND........................................................................................................................... 5  
   2.1 Split Infinitives – Standard or Non-Standard?................................................................. 5  
   2.2 Previous Research.............................................................................................................. 6  

3. RESULTS..................................................................................................................................... 7  
   3.1 Historical Perspective.......................................................................................................... 7  
      3.1.1 1810s-1890s.................................................................................................................. 8  
      3.1.2 1900s-1980s.................................................................................................................. 10  
   3.2 Variation in Present Day American English........................................................................ 12  
      3.2.1 Overview..................................................................................................................... 12  
      3.2.2 Year-By-Year Genre Analyses.................................................................................... 15  
      3.2.3 An Analysis of the Academic Genre and a Direct Comparison to the Spoken Genre...19  

4. DISCUSSION............................................................................................................................. 22  

REFERENCES............................................................................................................................... 27
1. INTRODUCTION

Speakers of English, native and non-native alike, are told never to split an infinitive. Yet, finding a split infinitive is not a rare occurrence, but is it actually incorrect language use? Can something with a high frequency of occurrence even be considered “incorrect”? Is the split infinitive more common in informal language where rules, legitimate or not, are constantly bent, or does it also have a strong presence in formal language?

A split infinitive, also known as a cleft infinitive, occurs when a word or phrase comes between the infinitive marker *to* and a verb in its infinitive form. The most common item to interfere and create this construction is an adverb (McArthur, 1998). A famous and common example of a split infinitive is a quote from the “Captain's Oath” in the TV-series *Star Trek*, as the final line reads: *to boldly go*. In this example, the adverb *boldly* interferes with the infinitive marker *to* and the verb *go*.

As a student of English as a second language, I have been taught to avoid this construction as it is regarded as improper or non-standard use in the English language. If an instance of a split infinitive is found in student work, the student typically receives a mark and is told to revise. However, the legitimacy of the split infinitive as being ungrammatical has been, and still is, up for debate.

The split infinitive is not a recent addition to the language or one that has made its way into the language through any modern devices of communication. Recorded evidence reveals that split infinitives occurred as early as the 13th century (Visser, 1972: 1049-55), which pre-dates the Great Vowel Shift and the start of the modern English period. However, for several hundred years it was not regarded as a faulty construction and no mentions of it were to be found in any grammatical guides (Crystal, 1985: 16). In 1762, the Bishop of London Robert Lowth opposed split infinitives on the basis of it not existing in classical languages, such as Latin and Greek (Lederer, 2013: 43), and in the early 19th century the debate began to take shape as more complaints were starting to get heard (Crystal, 1985: 16).
1.1 Aim

The aim of this study is to examine split infinitives in present day American English in order to find out in which contexts the constructions generally occur. Additionally, how the frequency of split infinitives has either increased, decreased or stayed stagnant will also be analysed, with the time period ranging from the start of the 19th century up until the summer of 2012. While the split infinitive has a range of different appearances, the focus of this study will be on the \( to + adverb + verb \) construction, as well as a comparison between the usage of the splitting negative construction \( to \ not + verb \), and its non-splitting counterpart \( not \ to + verb \).

When studying the presence of split infinitives in present day American English, focus will be on how these constructions differs between various text-types in which various levels of formality are expected. The genres of interest in this study are: “spoken”; “fiction”; “magazine”; “newspaper”; and “academic”. Furthermore, a comparison will be made between the “spoken” and “academic” text-types, which are arguably the two most informal and formal genres, respectively.

1.2 Material and Method

All data gathered in this essay was extracted from corpora. A corpus is a collection of text that has been assembled according to set criteria that vary depending on the purpose of the particular corpus. The source of the text can be written, spoken, or a combination of both, and can be from one or several different genres. Corpora allow researchers to analyse real-life authentic examples of language use in various true contexts (Adolphs and Lin, 2011: 597-98). Additionally, corpora allow for studies of change in language and structures over long and short periods of time, which is highly beneficial as language change is ordinarily a gradual event (Nevalainen, 2006: 23). Furthermore, because the data of corpora are the exact words once spoken or written, it is completely free from prescriptivism and interpretation (Maguire and McMahon, 2011: 50).
Corpora can vary considerably in size, ranging from only a couple of thousand tokens to more than a billion. In corpus linguistics, of course, a large corpus is desirable to get the most accurate results and allow generalisations (Ellis, O'Donnell and Römer, 2013: 35).

Two large corpora were used in this essay, both compiled by Professor Mark Davies of Brigham Young University. All of Davies' corpora run on an interface that allows users to search for and extract data without the need of any third party software. The main corpus of use was the Corpus of Contemporary American English (COCA). The number of tokens in COCA exceeds 450 million, which are equally divided over a time period of 22 years. The earliest texts originate from 1990, while the most recent texts are from the summer of 2012. Besides allowing comparison of data on a year-by-year basis, the data is also equally divided amongst five genres, which are: “spoken” (extracts from unscripted TV and radio programs); “fiction” (fictitious material from books, short stories, plays, and movie scripts); “magazine” (texts from nearly 100 different magazines); “newspaper” (texts from ten popular US newspapers); and “academic” (texts from nearly 100 different peer-reviewed journals). These genres can then be examined even deeper as each contains five-to-eleven sub-categories (e.g. “academic: Education” and “academic: Humanities”). It is worth noting, though, that fictitious material is not limited only to the “fiction” category. One may find, for instance, quotations or passages from various sources of fiction in the “newspaper” texts. Furthermore, one of the eleven sub-categories of “magazine” is “entertainment”, in which fictitious material is present. By and large, however, all five genres individually represent a different source of American English.

In addition to COCA, the Corpus of Historical American English (COHA) was used to examine the increase or decrease in frequency of split infinitives in American English history. COHA is a corpus that encompasses more than 400 million words from three centuries, stretching from the year 1810 to 2009. This makes it the largest structured corpus of historical English available for researchers according to Davies. The texts are extracts from four genres: “fiction”; “magazine”; “newspaper”; and “non-fiction books”.
Both corpora allow for extensive research into specific grammatical constructions. Searches for exact words or phrases will provide results in context of year and origin. More advanced queries are also available, such as searching for nouns followed by verbs in the gerund form, for instance. This is possible since both COCA and COHA are tagged corpora, which means that every word has been grammatically tagged. Because of this, the user can make queries based on these tags to include all possible words that share the same grammatical function (e.g. \([r^*] = \text{adverbs}\)). The results can then be sorted in either alphabetical order, by relevance, or by frequency.

This essay is corpus-based, i.e. the usage of the split infinitive was analysed from the data I collected via corpora. While either an adverb or an adverbial phrase (or very rarely, a pronoun) can cause a split infinitive, I decided to focus my research solely on constructions with an adverb as they make up the vast majority of all split infinitives (approximately 91 per cent in COCA and COHA). Likewise, I opted to study the negative to not + verb construction instead of the to never + verb construction for the same reason (approximately 84 per cent).

To begin the study, I analysed the presence of the split infinitive in a historical perspective. I used COHA to analyse the 1800s through the 1980s on a decade-by-decade basis to see how the usage of the construction varied in written language throughout the time period. To do this, I searched for all instances where to was immediately followed by an adverb, which in turn was immediately followed by a verb. The next step after analysing any trends was to look at the words that made up these constructions and examine which particular split infinitives with adverbs that were the most common, and if there were any clear patterns to be found.

After my initial research I moved on to the core of my study, which concerns the modern presence of split infinitives. In this section, I used COCA and began with a similar approach as I did with COHA. The focus was again on the variation in frequency on a year-by-year basis, but with the different genres as reference points. I compared all five genres and analysed any potential differences in not only the frequency of split infinitives, but more precisely the constructions and in what contexts they generally occurred. Additionally, the most common to + adverb + verb
constructions in each category were examined in depth by expanding the search to include collocates. Furthermore, extra attention was given to what could be considered as the most formal of the five text-types, which was “academic”. Its nine sub-categories were compared to study any potential variations between the academic fields. These results were subsequently analysed side-by-side with the arguably most informal genre, which was “spoken”. In addition to everything above, separate searches for and comparisons between the splitting negative construction to not + verb, and its non-splitting counterpart not to + verb, were made in order to strengthen the analysis. For convenience, the results have been normalised to frequencies per million words (pmw).

2. BACKGROUND

2.1 Split Infinitives – Standard or Non-Standard?

Language usage guides or manuals serve to teach and explain how a language should be used in a specific context. They can be in the form of a grammar book or a dictionary. On the subject of the split infinitive, the instructions from various sources will vary. Arguments can be found both in favour of and against the use of the construction, as well as no mention of it at all.

The root of the argument against the split infinitive lies in the grammar of Latin, on which certain portions of English prescriptive grammar were modelled. In Latin you can never split an infinitive, as infinitives are single indivisible words (i.e. no infinitive marker). Thus, some grammarians believe English should still abide by the Latin rule and keep the infinitive as a closed unit, despite the differences in structure (Aarts, Chalker and Weiner, 2014; Stevenson and Lindberg, 2010; Walker 2005: 18). It was on this basis that the Bishop of London, Robert Lowth, voiced his opinion against the construction in the mid-18th century which sparked the debate, much like the British 17th century poet John Dryden opposed ending a sentence with a preposition due to the Latin heritage (Lederer, 2013: 40-43).
Generally, English grammar books, which are aimed at students, will advise against the construction. To avoid it is the safer option, as it is never “wrong” not to split an infinitive. As an example, Estling Vannestål (2007: 276) notes how the split infinitive is a frequent item in informal language, but often avoided in writing where a more formal language is to be expected. Extra emphasis is put on avoiding the *to + adverb* combination where the adverb creates a negative construction, such as *to not* and *to never*.

However, if one scans linguistic journals and other linguistic works it should not be surprising to find more linguists that question the stand against split infinitives than those who support it. Crystal (1985: 17) states that people “who harangue the press about split infinitives are doing the language no service. They are, rather, promoting a spirit of uncertainty which will ultimately do far more harm.” He also refers to these people as having an “obsession”. Additionally, Crystal (1985: 16) questions why split infinitives would be an issue when one can split nouns (e.g. “the big man”) without any worries. Other arguments in favour of the debated construction include the following remarks: splitting of infinitives can be beneficial or even crucial when emphasis or clarity is needed (Allen, 2008a; Stevenson and Lindberg, 2010; Walker, 2003: 8); conscious avoidance can lead to awkward and ambiguous expressions (Quirk et al., 1985: 496-97); the usage of split infinitives is simply a “matter of taste” (McCloskey, 1985: 193); the sentiment against the construction is a superstition (Allen, 2008b); nothing is actually split, since the formative *to* is not a part of the infinitive in the English language (Trask, 1993).

### 2.2 Previous Research

While corpus-based research is a rather common method in linguistic studies, the collection of corpus-based studies centred on split infinitives is not particularly extensive. There are, however, a few in which I found comparative data. In Moisés D. Perales-Escudero's “To Split or to Not Split: The Split Infinitive Past and Present” (2010), the author examined the presence of multiword lexical items in COCA, more specifically certain split infinitives. Six of the most common *to + adverb*
constructions were selected, and the ratio between the use of the adverbs in both splitting and non-splitting constructions were studied. Furthermore, Perales-Escudero investigated collocational patterns in the split infinitives he found in his searches, as well as using a second corpus, the TIME Magazine Corpus (also compiled by Mark Davies), to compare the frequencies of *not to* and its splitting counterpart *to not*. My study analyses areas that Perales-Escudero's study did not cover, which includes examining the various sub-categories, as well as to take a look at the historical development, which dates much further back.

Another study that did take a historical look was carried out by Calle-Martín and Miranda-García (2009), where the authors used three much smaller corpora than COHA (a combined 40 million words compared to 400 million words), as well as using the larger 100 million-word British National Corpus to a smaller degree. Calle-Martín and Miranda-García investigated the use of split infinitives in the different corpora, as well as taking a deeper look at the adverbs involved.

Other studies include one focused on split infinitives in Asian varieties of English by Calle-Martín and Romero-Barranco (2014), which compared written and spoken English in Great Britain, India, Singapore, Hong Kong and The Philippines. Furthermore, there have been corpus-based studies on spoken and written academic language by Biber (2006) and adverbs in English by Simon-Vandenbergen and Aijmer (2007), amongst others.

3. RESULTS

3.1 Historical Perspective

In order to obtain the results found in tables 1 and 4 below, the search string in COHA consisted of the following tags: *to [r*] [v?i*]*, and included the decades 1810 through 1980. This query searched for all instances where *to* was immediately followed by an adverb, which in turn was immediately followed by a verb in its infinitive form. This was not without issues, however, as the results from this query also yielded constructions that were in fact not split infinitives. One such construction
was *to time + verb*, which exclusively derived from the idiomatic expression *from time to time*. While it is not a split infinitive, the corpus' interface recognised it as equal to the other *to + adverb + verb* constructions, and subsequently included it in the results. This had a noticeable impact on the numbers as the earliest decades of COHA all contain a much smaller number of tokens compared to the overall average, which is just above 20 million. To counter this problem, all lists were alphabetically ordered and thoroughly examined in an effort to spot these non-splitting constructions, as well as other inaccuracies, and exclude them from the results. The results were then manually re-calculated to provide the correct numbers (normalised pmw) portrayed in the tables.

### 3.1.1 1810s-1890s

<table>
<thead>
<tr>
<th>Decade</th>
<th>1810</th>
<th>1820</th>
<th>1830</th>
<th>1840</th>
<th>1850</th>
<th>1860</th>
<th>1870</th>
<th>1880</th>
<th>1890</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per Mil</td>
<td>9.31</td>
<td>4.33</td>
<td>3.85</td>
<td>6.67</td>
<td>12.43</td>
<td>15.77</td>
<td>32.46</td>
<td>30.95</td>
<td>26.62</td>
</tr>
</tbody>
</table>

As seen in table 1, the frequency of split infinitives throughout the 1800s varied quite drastically between the decades. The starting point of 9.31 per million words in the 1810s is not far off from the 12.43 frequency seen in the 1850s, but before it reaches that level, it plummets to a century-low mark of 3.85 in the 1830s. However, following this drop, a steady increase can be seen over the next 40 years. In just two decades, from 1830 to 1850, a 223 per cent increase brings the frequency above the starting point, to which it never again returns. Thus far, the three most common adverbs to interfere with an infinitive are *so, thus*, and *again*, which together combine for 16 per cent of all split infinitives of this variant. Up to this point, the data has been extracted from the “fiction”, “non-fiction” and “popular magazines” genres of COHA, as extracts from the “newspaper” category are not available until the 1860s and onward.
With newspaper language included, the frequency of split infinitives continues its upward spiral to 15.77 in the 1860s, before more than doubling to its century-high of 32.46 in the 1870s. A slight decrease can be seen in the 1880s before falling even more, and rounds out the century by landing on frequency of 26.62 per million words. With the introduction of newspaper texts, a significant increase can be seen of split infinitives involving the verb *understand*. As a prime example, *to fully understand* becomes the most common instance of a split infinitive in the latter half of the 19th century. As shown in table 2, also found in the top five are two constructions that both look very similar, as well as having similar meanings: *to fully appreciate* as the joint second-most common, and *to fully realize* at number four.

**Table 2.** The top five expressions during the first upsurge; 1860s-1890s

<table>
<thead>
<tr>
<th>Rank</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>To Fully Understand</td>
</tr>
<tr>
<td>2.</td>
<td>To Fully Appreciate / To Even Think</td>
</tr>
<tr>
<td>4.</td>
<td>To Fully Realize</td>
</tr>
<tr>
<td>5.</td>
<td>To Really Know / To So Arrange</td>
</tr>
</tbody>
</table>

**Table 3.** *to not + verb* versus *not to + verb*; COHA; normalised to frequencies pmw

<table>
<thead>
<tr>
<th>Decade</th>
<th>1810</th>
<th>1820</th>
<th>1830</th>
<th>1840</th>
<th>1850</th>
<th>1860</th>
<th>1870</th>
<th>1880</th>
<th>1890</th>
</tr>
</thead>
<tbody>
<tr>
<td>To Not</td>
<td>0</td>
<td>0.58</td>
<td>0.07</td>
<td>0.12</td>
<td>0.18</td>
<td>0.12</td>
<td>0.22</td>
<td>0.39</td>
<td>0.29</td>
</tr>
<tr>
<td>Not To</td>
<td>288.69</td>
<td>300.41</td>
<td>241.02</td>
<td>231.24</td>
<td>243.15</td>
<td>230.14</td>
<td>233.7</td>
<td>225.64</td>
<td>205.96</td>
</tr>
</tbody>
</table>
Table 3 examines the adverb *not*, which creates a negative construction. When placed after the preposition *to*, it creates a splitting construction. During the 1800s, the splitting alternative is practically non-existent when compared to its non-splitting counterpart. Still, the splitting construction does become more prevalent in the final three decades of the 1800s, with the 1820s being the only exception.

The frequency of the *not to + verb* construction also fluctuates throughout, but after a big drop from the 1820s to the 1830s, it never recovers. The decrease from 300.41 in the 1820s to 205.96 in the 1890s, which is a drop of 34 per cent, is quite significant and only the start of a continuing downward trend.

3.1.2 1900s-1980s

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Per Mil</td>
<td>27.02</td>
<td>19.91</td>
<td>14.04</td>
<td>12.9</td>
<td>9.98</td>
<td>16.8</td>
<td>23.28</td>
<td>34.67</td>
<td>42.42</td>
</tr>
</tbody>
</table>

The following century starts off in a pattern that very much resembles its predecessor. Despite there being a large upsurge overall in the late 1800s, that momentum does not carry over as the frequency numbers, as displayed in table 4, rapidly descend back down to the same low numbers seen a century before. It starts off with a minimal increase, from 26.62 in the 1890s to 27.02 in the 1900s, before a constant decrease over a 40-year period where the average once again falls below 10, as it settles on 9.98 per million words in the 1940s. Furthermore, a shift can be seen in the type of adverbs used. The previously popular adverb *fully* loses its spot at the top and is instead replaced with the arguably more colloquial *really*, which is the culprit in 6 per cent of all split infinitives of this kind. Likewise, *to fully understand* is replaced by the arguably more colloquial expression *to really know*, whose meaning is essentially synonymous.
However, as we move closer and closer to present day, the tide turns once more as another substantial increase of split infinitives can be seen, and without any interruptions along the way. A steady and swift increase on a decade-by-decade basis, beginning from the lowest point of 9.98 in the 1940s, drives the frequency up by 325 per cent as we reach the 1980s. The frequency of 42.42 is the highest seen in all of these selected decades of COHA. In terms of the adverbs used during the last surge, the to + adverb + understand construction makes a noticeable comeback, as to better understand is the second-most common splitting expression, found spread out in all four different genres. The top result is to even think, the joint number two in the late 1800s, which is mainly found in fictitious texts.

### Table 5. to not + verb versus not to + verb; COHA; normalised to frequencies pmw

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>To Not</td>
<td>0.36</td>
<td>1.06</td>
<td>0.23</td>
<td>0.57</td>
<td>0.37</td>
<td>0.69</td>
<td>0.71</td>
<td>0.97</td>
<td>1.66</td>
</tr>
<tr>
<td>Not To</td>
<td>204.14</td>
<td>195.9</td>
<td>181.69</td>
<td>166.32</td>
<td>170.2</td>
<td>169.93</td>
<td>166.99</td>
<td>184.5</td>
<td>175.94</td>
</tr>
</tbody>
</table>

The negative to not + verb construction also sees a late distinct increase in usage that coincides with the to + adverb + verb split infinitive upswing. As seen in table 5, after rising to 1.06 in the 1910s, its frequency fluctuates through a couple of decades. However, from the 1950s and onward, no drops are seen as the average climbs to reach its highest point as of yet, which is 1.66 per million words in the 1980s. Just as the to + adverb + verb construction reached its highest point in the 1980, the negative split infinitive steadily increases to culminate at a frequency of 1.66 per million words in the same decade, with the most common expression being to not be. Its non-splitting counterpart, not to + verb, also sees another overall decrease. While it never sinks to a frequency below 200 during the 1800s (lowest being 205.96 in the 1890s), it only remains above that mark for the first decade. It reaches its lowest point of 166.32 in the 1930s, before eventually settling at
175.94 in the 1980s. It is still by far the more common construction, yet it has lost a marginal bit of ground on its splitting counterpart.

3.2 Variation in Present Day American English

The same search technique was used in COCA as in COHA, and since they run on identical software, the same issues applied. However, because of different circumstances, I could not manually exclude all irrelevant hits from the results. While the amount of tokens in each decade of COHA average around 20 million, only two individual years (2010 and 2012) in COCA are below that same number. Because of this, the software would “time out” due to too many tokens when attempting to display all the examples over a multi-year period. Furthermore, when delimiting the search to one individual year, COCA will not account for any selection made regarding the text-types. However, in the top 500 of the most common to + adverb + verb constructions found in COCA, not one single oddity or inaccuracy was found. While a very small amount did exist further down in the results, their impact on the statistics as a whole is negligible, and in no way destructive to the purpose of the study. For instance, there are only a total of 20 tokens of the to time + verb expression in all of COCA, and the most frequent one, to still be, is not found until spot 1183.

3.2.1 Overview

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Per Mil</td>
<td>106.39</td>
<td>124.98</td>
<td>140.01</td>
<td>155.4</td>
<td>177.58</td>
</tr>
</tbody>
</table>

As seen in table 6, the sharp upward trend seen in COHA continues into the 1990s and 2000s in COCA. While the frequency more than doubles from the 1980s to the 90s, the inclusion of spoken data, which is commonly a place for a more informal language, helps inflate the numbers. While
there is definitely a large escalation in the use of split infinitives, the jump may not have been as huge had a “spoken” genre been included in COHA, as it is reasonably to believe the frequencies would have been higher to begin with.

Nevertheless, the results clearly show how the split infinitive is becoming more and more a familiar and regular feature in the English language. After eclipsing the frequency of 100 per million words during the early 90s (100.89 in 1991 to be precise), that number is not far off from doubling once more, as the 177.58 frequency in the three-year period 2010-12 culminates with a 185.77 in the year 2012. On a year-by-year basis, a decrease can be seen only twice (1995-96 and 2007-08), but overall the growth is quite extensive with no signs of slowing down.

Table 7 displays the frequency of the to + adverb + verb construction in each of the five available genres. As portrayed, the “spoken” genre by far features the most instances of the split infinitive. In COCA, this category consists of transcripts of unscripted conversation from more than 150 different American TV and radio broadcasts, with the total amount of words exceeding 95 million. On its own, it makes up for 32 per cent of all split infinitives of this variant, found in the texts of COCA. Due to the more informal nature of spoken language, this result should not come as a surprise.

The “fiction” genre, however, which features the highest frequency of split infinitives in the data from COHA, is now the only sub-100 category and only contains a low 9 per cent of these splits found in COCA. The fictitious material is made up of more than 90 million words extracted from books, movie scripts, as well as short stories and plays form various magazines.

Following “fiction” are the “magazine” and “newspaper” genres, whose frequencies are close to identical at 116.99 and 119.06, respectively. The “magazine” genre is a collection of over 95 million
words from nearly 100 million different magazines, which cover all sorts of different subjects. The newspaper texts are extracts from ten popular American newspapers, including *USA Today* and *New York Times*, whose combined total is close to 92 million words. Likewise, different subjects are covered, such as “sports” and “finance”.

Perhaps the biggest surprise in table 7 is the high frequency number in the last remaining genre, “academic”, which is supposed to be the most formal of all the above. At a frequency of 169.78, it contains the second-most instances of the *to + adverb + verb* construction. The texts are a compilation of more than 91 million words from nearly 100 different peer-reviewed journals.

### Table 8. *to not + verb* versus *not to + verb*; COCA; normalised to frequencies pmw

<table>
<thead>
<tr>
<th>Genre</th>
<th>Spoken</th>
<th>Fiction</th>
<th>Mag.</th>
<th>News.</th>
<th>Acad.</th>
<th>Years</th>
<th>90-94</th>
<th>95-99</th>
<th>00-04</th>
<th>05-09</th>
<th>10-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>To Not</td>
<td>16.27</td>
<td>4.19</td>
<td>4.65</td>
<td>7.19</td>
<td>3.42</td>
<td>To Not</td>
<td>5.11</td>
<td>6.32</td>
<td>7.14</td>
<td>8.93</td>
<td>9.96</td>
</tr>
<tr>
<td>Not To</td>
<td>149.07</td>
<td>208.83</td>
<td>142.84</td>
<td>141.53</td>
<td>122.97</td>
<td>Not To</td>
<td>159.88</td>
<td>156.36</td>
<td>151.98</td>
<td>147</td>
<td>144.72</td>
</tr>
</tbody>
</table>

As the *to + adverb + verb* construction continues its increase, so does the negative *to not + verb* construction, as seen in table 8. Once again, it is the “spoken” genre that by far features the most, with a 16.27 per million word frequency. This time, though, the “academic” genre features the least instances, with a low average of 3.42. Overall, a fairly constant increase can be seen throughout the time periods, as the average rises from 5.11 in 1990-94, to 9.96 in 2010-12. *To not be*, which is the most common expression in COHA, remains as number one.

The non-splitting variant also continues its downward trend, as it slowly but steadily decreases from 159.88 in 1990-94, to 144.71 in 2010-12. There is still a large margin between the usage of the splitting and non-splitting constructions, but as time moves forward, it continues to shrink. Additionally, just as with its counterpart, it is the verb *be* that is the top collocate.
3.2.2 Year-By-Year Genre Analyses

Graph 1. to + adverb + verb; COCA; normalised to frequencies pmw

Graph 1 shows how the prevalence of the to + adverb + verb split infinitive changes throughout the 1990s, 2000s, as well as the years 2010 through the summer of 2012. To produce accurate representations, manual re-calculation were again needed due to an algorithm error in COCA, as noted by Jäger (2013). When delimiting the search to a specific genre and studying it on a year-by-year basis or over one of COCA’s pre-selected time intervals (e.g. 90-94), COCA incorrectly compares the number of tokens found in that specific genre and time period to the overall number of tokens in the same time period, without considering the delimited genre selection. Hence, manually excluding the tokens from irrelevant genres was needed to provide accurate results.

As evident by the graph, all five genres see a notable overall increase as we move to present day, though with varying degree. The steepest increase can be seen in academics, as a growth of 148 per cent takes it from a frequency of 104.21 per million words in 1990, to 257.98 in 2012. A more in-depth look at the “academic” text-type, and its sub-genres, will be the focus of section 3.2.3.
The “fiction” category stands out at the bottom of the graph, as it is the only genre never to eclipse a triple-digit frequency. However, that should not be far off as the frequency more than doubles over the 22-year period, when it in 2012 reaches an average of 93.26 instances per million words. The two most common split infinitives are *to even think* and *to just let*. In the former, *about* is the top collocate, while the verb *go* most often goes with the latter. In both cases, the constructions generally appear in either dialogue or inner monologues, which in fiction is subject to be written in a natural and talkative way, where the importance of formal language may not necessarily be at the forefront.

Just as the overall per million word frequencies are very even between the “magazine” and “newspaper” genres, the numbers are also incredibly similar on a year-by-year basis. Their lines mesh together, as they both see an almost identical growth pattern. In newspaper texts, the frequency rises from 97.24 in 1990 to 169.22 in 2012 (with a dip down to 79.31 along the way), while it in magazines rises from 70.42 to 145.66.

In both genres, the top two most common splitting constructions are identical, including the top collocates. The most popular expression is *to better understand*, with *how* most often attached. In the “magazine” category, the sources in which these expressions are found are mainly magazines on the subjects of “science” and “research”. In newspaper texts, they are found in an even mix of interview transcripts and regular in-text writing. The second-most common expression is *to just be*, with the verb *want* as the top collocate. In magazines, the subject is now no longer “science” and “research”, but rather “entertainment”. In newspapers, it is almost exclusively found in transcripts of various interviews.

The “spoken” genre, with the highest total tokens of split infinitives, sees the smallest increase of all categories in COCA, percentage wise. However, as it was already so far ahead of the bottom three genres to begin with, even with a smaller increase in percentage, the margin widens quite significantly in all three cases. With an increase of 66 per cent, from 157.4 to 266.59, it begins as the most common genre to feature split infinitives in 1990, while just barely remaining in the top
spot in 2012 with the “academic” genre as an extraordinarily close second. The expression to just be is for the third straight time the second-most common item. However, the top collocate is no longer the verb want, but rather the adjective able. In every instance, the construction is pulled from spoken language by interviewees, and never news anchors. However, several of the interviewees are educated people discussing very serious matters, which means that the splitting constructions appear in surroundings of very formal language. The top expression is to really get, which features the preposition into as its top collocate. The context is again mainly interviews, but on more relaxed subjects in the entertainment department.

**Graph 2. to not + verb; COCA; normalised to frequencies pmw**

![Graph 2](image)

Graph 2 shows the development of the splitting to not + verb construction throughout the genres. While table 7 displays how the total amount of to not + verb constructions steadily increases, this is not the case for all five genres. In academics, the fastest growing genre for the to + adverb + verb construction, the amount of tokens in each year from 2010 to 2012 is actually less than in 1990. The line spikes on two occasions, as it goes above an average of 6 per million words in 2003 and 2009, but overall it remains around the starting point throughout. This is the only genre in which no clear
increase can be seen. Interestingly enough, as seen in graph 3, the largest decrease of the non-splitting *not to + verb* construction is also found in academics, in which it more than halves, which suggests that the usage of the expression as a whole is diminishing.

Regarding the splitting version, the “fiction”, “magazine” and “newspaper” genres are very much alike in the sense that they all start around the same average in 1990, and also settle around the same average in 2012, although their paths are a bit varied. While the average stays below 5 throughout the 90s and into the early 00s in the “fiction” and “magazine” categories before eclipsing that mark and finally settling at around 8 in 2012, it reaches that mark as early as 1992 in newspaper texts, and after 1994 it never again returns. It even goes as high as above 10, before again joining the previous two genres at an average around 8 in 2012. Regarding the non-splitting construction, however, the “fiction” genre breaks free and is the only category in which an increase can be seen, as it rises from 201.83 in 1990, to 231.85 in 2012. In contrast to the “academic” genre, then, the expression as a whole is noticeably growing in popularity. The “magazine” and “newspaper” genres again share almost identical starting and finishing points, and the roads there are very similar.

The last remaining category is spoken language, where the splitting expression sees the largest increase as it more than doubles over the 22-year period. It has its ups and downs along the way, but the ups are more significant than the downs. Starting at 9.69, it eventually reaches 21.29, which is far above any of the other genres. Simultaneously, it also exhibits the second-biggest drop of the non-splitting expression, which suggests the splitting construction is gaining ground, and doing so relatively quickly. The top two expressions are *to not be* and *not to have*, which is also the case for all five genres.
3.2.3 An Analysis of the Academic Genre and a Direct Comparison to the Spoken Genre

As previously mentioned, the biggest rise in prevalence of the split infinitive construction with adverbs is seen in the field of academics. A 148 per cent increase brings the frequency from 104.21 in 1990 to 257.98 in 2012, and almost surpasses the “spoken” genre in the process. What makes this particularly interesting, and worth further analysis, is the fact that this genre is the one in which the most formal language is to be expected. One could assume, then, that the split infinitive would be avoided where possible due to its controversy.

In COCA, the genre of “academic” is divided into nine sub-genres that cover different areas and subjects. The sub-genres are: “education”; “history”; “geology/social science”; “law/political science”; “humanities”; “philosophy/religion”; “science/technology”; “medicine”; as well as a “miscellaneous” collection. Due to the undefined nature of the “miscellaneous” category, it will be present in the table but excluded from the discussion.
As shown in table 9, there is a notable difference between the sub-genres of academics. The highest average per million words is found in the subject of “education” at a frequency of 291.53, which is remarkably higher than any of the other sub-genres. Only one, “medicine”, is also in the 200s, at a frequency of 213.12. The lowest is found in the field of humanities, where the average just stays above the century line, at 105.98.

The overwhelmingly most popular construction is to better understand, as it is the top result in each of the sub-genres, and subsequently, of course, also the top result in the entire “academic” genre. More variety can be seen between the various sub-genres when looking beyond just the top result, but overall they all look fairly similar with no major discrepancies worthy of note. The formality of the language is very apparent, with phrases such as to just and to even very rarely appearing in the list of the top 50 expressions in any of the given sub-categories. Instead, the lists are riddled with expressions such as to better, to further, to adequately, to accurately, and so on and so forth, with the accompanying verbs often being understand, explore, investigate, or variants thereof. Furthermore, the rather high per million word frequency in each sub-genre, particularly the top two, shows the split infinitive is a fairly regular and accepted grammatical item in academic literary works.

When comparing the aforementioned expressions to the ones found in the spoken data, the differences are very noticeable. The long and more complex verbs are for the most part entirely gone from the top 100 list. Instead, basic verbs such as get and be are seen aplenty, and often preceded by the adverb just.

Table 9. to + adverb + verb; sub-genres of “academic” in COCA; normalised to frequencies pmw

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Per Mil</td>
<td>291.53</td>
<td>124.95</td>
<td>185.66</td>
<td>167.2</td>
<td>105.98</td>
<td>144.36</td>
<td>188.2</td>
<td>213.12</td>
<td>73.06</td>
</tr>
</tbody>
</table>


Table 10. Top 8 to + adverb + verb constructions; "academic" versus "spoken"

<table>
<thead>
<tr>
<th>Rank</th>
<th>Academic</th>
<th>Spoken</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>To Better Understand</td>
<td>To Really Get</td>
</tr>
<tr>
<td>2.</td>
<td>To Fully Understand</td>
<td>To Just Be</td>
</tr>
<tr>
<td>3.</td>
<td>To Further Examine</td>
<td>To Really Be</td>
</tr>
<tr>
<td>4.</td>
<td>To Further Explore</td>
<td>To Just Say</td>
</tr>
<tr>
<td>5.</td>
<td>To Strongly Disagree</td>
<td>To Really Make</td>
</tr>
<tr>
<td>6.</td>
<td>To Strongly Agree</td>
<td>To Just Get</td>
</tr>
<tr>
<td>7.</td>
<td>To Further Develop</td>
<td>To Just Go</td>
</tr>
<tr>
<td>8.</td>
<td>To Further Investigate</td>
<td>To Actually Get</td>
</tr>
</tbody>
</table>

Table 10 displays the top 8 splitting constructions in each genre, and the differences in both meaning and appearance are quite clear. None of the top 8 appears on both lists. Actually, most of them do not even feature in the top 100 in both genres. From the “academic” list, only the top two make it on both lists. To better understand sits at number 31 in the “spoken” genre, while to fully understand barely makes it in, as it sits at number 92. When looking the other way around, the top result to really get is the sole expression to be featured on the academic list, as it is found at number 89. In fact, this is the only time the verb get is featured at all in the top 100 of the “academic” genre. Furthermore, the verb be is only seen thrice, with to also be at spot 38, to still be at spot 72, and to always be at spot 88. To take it a bit further to show how vast the differences are, two of the top expressions in the academic texts, to further examine and to strongly agree, are not found even once in all of the spoken data in COCA.
4. DISCUSSION

This essay intended to tackle two subjects regarding to the to + adverb + verb and to not + verb split infinitives in American English. The first was how the frequencies have changed from the early 1800s to present day. The second was focused on how the usage and appearances of said split infinitives differ in context of different genres. It is more in-depth with regard to the particular text-types than the previous studies mentioned in section 2.2. However, as only American English is covered, similar studies on other dialects of English may be of future interest. Additionally, studying split infinitives with adverbial phrases or pronouns, as well as the to never + verb construction, may also be of interest.

The first part of the corpus study focused on the historical perspective. There were no signs in the data from COHA to suggest that the increasing negativity surrounding the construction in the early 1800s had a lasting negative effect on its usage. Although the frequency tended to fluctuate up and down on a decade-by-decade basis for a good 150 years, the increases were on average far heftier, and the end result was a frequency that had more than quadrupled over a time period of just about two centuries.

Of course, this was without any spoken data being accounted for. With the change from COHA to COCA, which includes a “spoken” genre of equal size to the other genres, the frequency per million words of the to + adverb + verb more than doubled instantly. That is not to say that the overall usage of the split infinitive doubled from the 1980s to the 90s, of course, but as the frequency just kept rising and rising in the decades covered by COCA, and in a very quick pace as well, it is safe to say the split infinitive is becoming more and more a standard feature for every new generation of English speakers.

With that in mind, is there even an argument against the split infinitive that can still be logically and convincingly supported? A high frequency of occurrence does not automatically make something “correct”, as, for instance, something may become a standard within a specific culture or
a secluded population while remaining questioned or unaccepted outside of it. However, as this study shows, along with other studies on the split infinitive, its usage is widespread and generally accepted across all genres and various levels of formality. To refer to it as “incorrect” or “ungrammatical”, then, seems irrational.

The second half of the corpus study focused mainly on the question regarding the differences between the various genres of COCA, especially between “spoken” and “academic”. The “spoken” genre proved to be the most common to feature the split infinitive between the years 1990 and 2012, but that should come as a surprise to no one due to the informal nature of spoken language. However, to see “academic” as the second-most common genre is quite peculiar. Being the most formal of all the genres, one would likely assume the split infinitive would be avoided if it was indeed incorrect language use. Instead, the frequency is far higher than in any of the remaining three genres, with “fiction” falling way behind despite exhibiting the majority of the most frequent splitting expressions in COHA. Not only that, but the “academic” genre is the fastest growing in regards to the usage of the $to + adverb + verb$ construction, and it is not even a close race. As graph 1 displays, in 2008 and 2011 it even featured more than the “spoken” category, despite the vast difference in expected formality. This frantic upward trend of splitting constructions in academic texts, with no signs of slowing down, seems to be a good indicator of the acceptance level of this grammatical construction, and whether or not grammar books aimed at students should continue to advise against its usage could perhaps be questioned.

Which particular expressions, i.e. what adverbs and verbs, that were the most common in each genre was also covered, along with the top collocates. Here, both differences and similarities between the genres were seen. The top expression in newspapers and magazines was the same as in academic texts, while the second-most common was akin more to the informal nature of the expressions found in fictitious texts and spoken transcripts.

A more in-depth analysis was given to the “academic” genre, as well as a comparison with the “spoken” genre. While the frequency between the sub-genres of academics did differ considerably,
the particular expressions remained the same throughout. The actual language also remained true to context, as the words used to produce the splitting expressions did not stray away from the formality level expected in the genre. The constructions remained complex with longer words being preferred, which is a standard in formal writing (Matthews, 2014). Furthermore, the purposes of the expressions in which the split infinitives occurred very much fit into context as well. In other words, they appeared naturally where the authors needed an adverb, and the splitting variants were simply preferred, and accepted, instead of non-splitting alternatives.

The reason behind this increase of split infinitives, especially in academic writing, may simply be a question of semantics. By placing the adverb directly in front of the verb it is supposed to influence, there should be zero chance of ambiguity, which is of outmost importance. In academic writing there is no room for inadvertent ambiguity. By changing this order, and leaving the infinitive intact, questions surrounding the adverb’s purpose may arise. For instance, the expression *to better understand cell development* (true example from COCA) causes no ambiguity, as the message is patently clear. Switch the order, and the adverb may be read as an adjective (*better cell development*). Separating the adverb from the verb may solve that particular issue (*to understand cell development better*), but may on occasion open the door for potential ambiguity as the sentence continues. To get around this, the entire sentence could require rewriting. Judging by the rising frequencies across the board, particularly in academic writing, it seems that absolute clarity is being preferred rather than adhering to the old sentiment against the split infinitive. Furthermore, if we can “split nouns” by placing a modifier directly in front of a noun, why should we not be able to do the same when dealing with verbs? It is only logical, which is a sentiment that may be catching on.

The differences between academic and spoken data were clearly on display in table 10. Very few expressions in the top 100 of each genre crossed over to the other, and to label which list was which was probably not even truly needed. Curiously enough, a couple of the top expressions were very similar, though not in appearance. In many contexts, *to better understand* and *to really get* could convey the exact same message, as could *to fully understand* and *to actually get.* These were the
exceptions to the rule, however, as the majority were not convincingly comparable. Furthermore, the difference in formality was also on display. As Leech and Svartvik (2002: 33) explain, “[m]uch of the vocabulary of formal English is of French, Latin, or Greek origin. In contrast, informal language is characterized by vocabulary of Anglo-Saxon origin.” In table 10, all but one (understand) of the verbs belonging to the “academic” genre have either Latin or French heritage, while all of the verbs in the “spoken” genre have Anglo-Saxon origin.

The inclusion of a few comparisons between the negative constructions to not + verb and not to + verb were made in effort to provide a look at a different kind of split infinitive. I had an extra bit of personal interest in this area as well, as this splitting construction was given extra attention to in a previously mentioned course book of mine aimed at students of English as a second language (Estling Vannestål, 2007).

In all the decades throughout the 1800s to 1980, which were covered by COHA in this study, the splitting version was practically non-existent. It is interesting to note, however, that the non-splitting version also decreased significantly during the same time period, which might suggest the lack of an increase of the splitting alternative could partly be due to an overall decrease in usage of the expression as a whole. It was not until the addition of spoken data in COCA that the to not + verb construction reached numbers that were actually worth considering. Furthermore, the frequency of the non-splitting counterpart continued to slide, and even with the inclusion of another genre the numbers did not receive an initial inflation. Now, the splitting variant was by far mainly found in spoken language, and barely found at all in academic literary works. In fact, in academics, both variants saw decreases, with the per million word frequency of the not to + verb construction dropping as low as to 64.39. In conclusion, the splitting expression is not the preferred version in any genre, nor is it even relatively close, although the margin is successively shrinking. However, if one is to find one of these negative split infinitives, the odds are it will be in the context of informal spoken language.
It is of course important to note that the “spoken” genre in COCA does not consist of 100 per cent natural everyday small talk, as all the data are from recordings where the participants were fully aware of the fact they were being recorded, and perhaps slightly altered their speech accordingly (consciously or not). Either way, to say that split infinitives are most commonly found in spoken language is a relatively safe statement. However, considering how accepted it is in academic language, and how rapidly the frequency is rising, maybe it is time to stop advising against it. I made sure to avoid splitting infinitives in this essay as that is what I have been taught, although the data would have supported me if I had. After all, to purposely split an infinitive seems to be quite alright.
REFERENCES

Primary sources


Secondary sources


