1 Introduction

This study was designed and developed to answer different questions based on the dogma that split infinitives should be avoided and are ungrammatical according to traditional grammars. Such a concept has had repercussions on both native and l2 speakers, since both are taught to avoid this construction.

Even though this topic seems to be quite controversial, it is undoubted that the frequency of split infinitives has raised since they were first mentioned. This increase has subsequently sparked a lot of interest on the nature of the phenomenon, its history and its presence among different languages, language forms and contexts.

This corpus-based study tried to test some hypothesis that could shed some light on the use of split infinitives among different kind of students in different fields. The questions that started this research were aimed at university graduate students in two different disciplines, namely: Arts and Humanities and Physical Sciences. These questions were:

- Is it possible that humanities students are less inclined to the use of split infinitives due to the literary studies involved in their career?
- Is it possible that either one of the native speaker or l2 speaker groups use more frequently split infinitives?
- Are the answers to these questions somehow related to the general adverb frequencies in each group?

n order to answer these questions, the British Academic Written English Corpus (BAWE) was inquired through the freeware AntConc. The data collected was then filtered and organized with python code and Libre Office Calc. As it will be explained, the data first collected by the corpus needed to be filtered due to the limits of AntConc and the irregularities which occur among adverbs.

In spite of the stereotypical expectations that were held before the data analysis, this study has showed that working with corpora can give surprising results and lead to more questions about the nature and dynamics of language which can push research forward.

2 On the split infinitives

The Cambridge grammar of the English language defines a split infinitive as "the construction with an adjunct in post-marker position" (Huddleston R., Pullum G., 2002; pg.581). The following examples were taken from the BAWE corpus from Humanities students, except the last one. These will show some occurrences involving different adverbs that can interfere between the preposition to and the main verb:

One has to truly believe that paper is valuable.
I aim to critically examine criteria.
Tom Eliot is able to symbolically enact the death of the patriarchy.
We thought to immediately call the police.

It is often taught to students to avoid as much as possible split infinitives, since they have been considered ungrammatical by many grammars throughout history. The first report on the negative view towards split infinitives comes from a letter to the editor of the New England Magazine written in 1834 (Bailey, 2006). Both native students and L2 students are often taught that this is a bad practice and the reason of this idea may lie in the fact that in Latin it is impossible to split an

infinitive form (as in any modern Romance language), so grammarians and scholars have tried to reinforce this rule on English as well, even though infinitives are made up by two words in English.

However, in Youngjun Jang and Sunjoo Choi (Chung-Ang University, 2014) it is showed how the frequency of split infinitives has raised constantly from the '20s to the '70s, then it decreased only to start raising again from the '90s until 2012. These data already shows that the English language behaves at times in opposition to the rule enforced against split infinitives. As a matter of fact, some cases of necessary or even required split infinitives can be easily inferred and are also reported in the Cambridge Grammar of the English Language. By way of example, our last example would have a very different meaning if the adverb was moved before the preposition:

We thought to immediately call the police .
vs.
We thought immediately to call the police.

An article on the Language Log on the University of Pennsylvania website (Zwicky, 2004) analyzes the obligatory use of some adverbs such as more than, over and up. As showed in the article, any structure other than this would be considered ungrammatical:

We expect it to more than double.

These words were not included in this study though, because as the author himself states, these are actually prepositional modifiers and as such they must be adjacent to the phrase they combine with.

3 Corpus Methodology

The BAWE corpus was realized through the project 'An Investigation of Genres of Assessed Writing in British Higher Education' from the Economic and Social Research Council (2004 – 2007). This massive corpus, which contains

2761 writings from university students, is divided in four major fields (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences) which are then also divided by genre, discipline, grade, students' gender, age, native language and education. These writings range in length from 500 words to 5000.

Moreover, the BAWE is made up only by writings from advanced academic students and this means that it can be considered a specialized corpus and a corpus of learner. This is a key element, since academic writing is one of the most formal writing styles and it requires excellent knowledge of the English language and the stylistic rules involved like avoiding split infinitives.

First, in order to collect data from fairly different sources, only the Humanities and Physical Sciences fields were chosen, since the writings from these fields are the most different from the point of view of text linguistics. Moreover, these two fields seem to provide the greatest number of writings. Next, the writings from both fields were divided between native and non- native speakers, with a total of 4 groups: Native Humanities (NH), Native Sciences(NS), L2 Humanities (L2H) and L2 Sciences (L2S).

Dividing the files from the BAWE is not an easy task. The corpus is very well organized and Coventry University provides a very well designed Excel sheet to order and filter the file list in many ways, nevertheless collecting the single files can be daunting and easily subjected to human error, especially in a case like this were many of the writings were screened out. To prevent this, a really short yet useful python script allowed to copy in a new directory the files needed and listed in a text file.

```
import os
import shutil

destination = "NEW_DIRECTORY_HERE"

### THIS SAYS WHICH FILES TO SEARCH######
with open('FILE_LIST') as fh:
    content = fh.readlines()
    content = [x.strip() for x in content]

### THIS FINDS AND COPIES THE FILES ### FULL BAWE CORPUS DIRECTORY ###
for root, dirs, files in os.walk("\\corpusbawe\\2539\\download\\CORPUS_TXT\\"):
    for _file in files:
        if _file in content:
            print('Found file in:   ' +_file)
            shutil.copy(os.path.abspath(root + '/' + _file), destination)
```

Figure 1

The result was a total of four directories (or new corpora), each dedicated to one of the groups. The NH directory contained 164 writings, the NS directory contained 172, the L2H 74 and the L2S 94. Now that the files were all organized it was easy to import the corpora with the freeware AntConc. The software was used to inquiry the four different corpora created from the main BAWE one. To find all occurrences of split infinitives the first search in the query was a simple combination of to *ly which provides a wildcard to find all words ending with

the morpheme -ly after the preposition to. There were obviously hits in each corpus but the data needed to be refined.

First of all, occurrences of verbs ending in -ly (to apply, to comply, to fly, etc...) needed to be removed, to achieve this the raw results in AntConc were copied on a text file and then the verbs were filtered through a simple piece of python code.

```
tokens = FILE_WITH_ANTCONC_RESULTS

stop_words = ['to apply', 'to comply', 'to apply', 'to fly', 'to rely']

filteredtokens = []
### THIS FINDS THE UNWANTED VERBS AND FILTERS THE TOKENS ###
for w in tokens:
    if w not in stop_words:
        filteredtokens.append(w)
```

Figure 2

Second, irregular adverbs needed to be accounted for (just, well, fast and late). Therefore a search of each followed on AntConc and the results, quite sparse in this case, were added to the regular adverbs. With the raw frequencies available the first thing to do to get a general idea of the situation was to normalize the data. As it can be seen in tables 1 and 2, NH and NS share an extremely close occurrence of split infinitive form, while the ratio between the two L2s more than doubles.

	Split Infinitive	Total Tokens
NH	128	494036
NS	117	453697
L2H	23	168112
L2S	69	213879

	Split Infinitive	Normalized Total
NH	259.09	1M
NS	257.88	1M
L2H	136.81	1M
L2S	322.61	1M

Table 1: Raw Frequency Normalized Frequency

Without using any further process or resorting to more math it is clear and safe to claim that both native students from the Humanities and Physical Sciences use the split infinitive in the same amount. This basically disproved the hypothesis that NH students are more conservative and careful in avoiding split infinitives than NS students. The numbers where calculated in python and then represented in the following bar chart by the same programming language.

Nonetheless, it is interesting to notice that the occurrence in the L2H is just slightly more than half of the NH and NS groups. This raised the question “Is

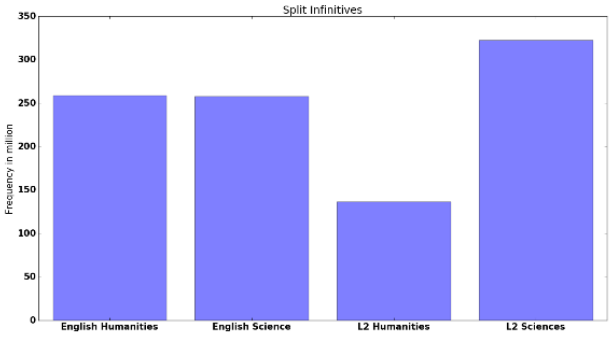

Figure 3

it a coincidence that L2H occurrences are so low, or is it because L2H students do indeed split the infinitive much less?". From the different ways available to test such hypothesis the Chi-squared test was chosen since it is fairly reliable with large data.

	NH	SH	L2H	L2S	
split	128	117	23	69	337
Total tokens	493908	453580	168089	213810	1329387
TOTAL	494036	453697	168112	213879	1329724

NH	SH	L2H	L2S
125.2	114.98	42.6	54.2
493910.79	453582,01	168069,39	213824,79

Figure 4

Once calculated the degree of freedom from the Chi-squared distribution chart, the result was 13.19, with the L2H and L2S split infinitives giving the highest contribution of respectively 9.01 and 4.04. According to the chart of the degrees of freedom this gives a degree lower than 0.01 which means that the result is significant.

Even though this proves an interesting fact, it would have been even more interesting to see the general adverb frequency in each group compared to the split infinitive frequency relative to it. The hypothesis was that there was a relation between the previous results and the general adverb frequencies and that probably the groups that used adverbs the most were the NH and L2 ones. By repeating the process of inquiring the corpora with AntConc (this

time searching only for *-ly and irregular occurrences) it was possible to find out each raw frequency and then normalize them by the million. Surprisingly, the starting hypothesis was disproved, since the groups that use adverbs the most are NH and L2Sci.

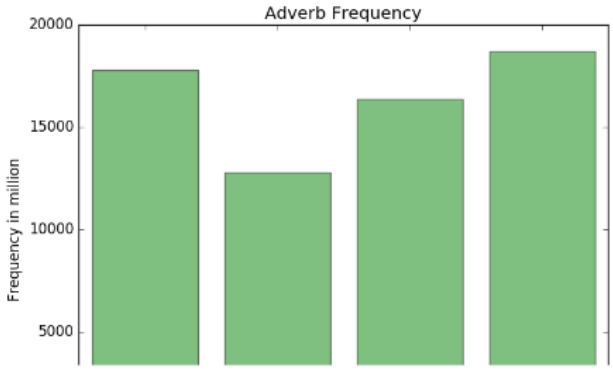

Figure 5

All things considered, it could have ended here but one more thing that could have given more information was to find out which group had the highest concentration of split infinitives in a ratio between split infinitives : general adverbs. In order to do this the numbers were normalized to 100 in order to give a percentage. The result was that the group NS, which uses the lowest amount of adverbs, was also the one with the highest concentration of split infinitives, while L2H was again the one which had the lowest rate of split infinitives.

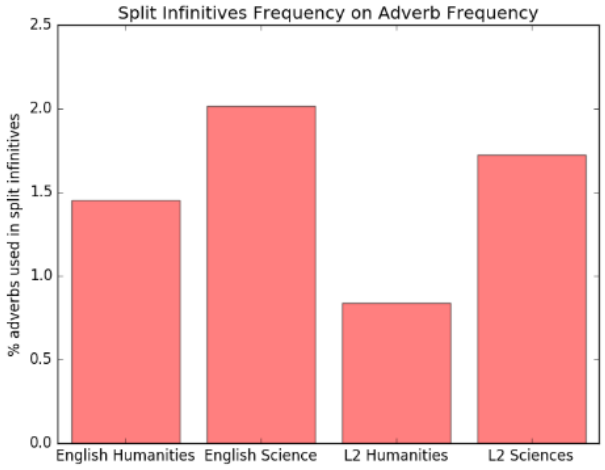

Figure 6

4 Conclusion

To summarize, this research has accomplished its aim of answering all of the preliminary hypotheses and questions asked.

Firstly, The results are that students from the NH group do not split infinitives less than students from the NS group, but the L2H group does have a substantial drop in the frequency of the form. This extra finding has been proven not to be due to coincidence but to have a meaningful relation compared to the other groups.

Secondly, it is clear that the answer to the preliminary second question is that L2 students from the Humanities field do indeed rely much less on split infinitives than L2 students from the Physical Sciences, who also are the major users of this structure.

Finally, going more in depth has rewarded the effort by showing how general adverb frequencies relate to the use of split infinitives among the groups. It is clear that even though NS students are the ones who use the lowest amount of adverbs, they are the ones who employ split infinitives the most while L2H are still the last ones to rely on them.

This study can then claim that in L2H students there can be found the least frequencies of split infinitives from any point of comparison. It can also be stated that L2S are the ones who apply split infinitives the most while no differences can be found between native students except in the percentage of adverbs employed in split infinitives.

References

Anthony, L. (2014). *AntConc*. Tokyo: Waseda University. URL: http://www.laurenceanthony.net/.

Freddi, M. (2014). *Linguistica dei Corpora*. Carocci Editore.

Huddleston, R. and Pullum G. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". In: *Computing In Science & Engineering* 9.3, pp. 90–95. DOI: 10.1109/MCSE.2007.55.

Rossum, Guido (1995). *Python Reference Manual*. Tech. rep. Amsterdam, The Netherlands, The Netherlands.

Wickens, P. (2004-2007). *British Academic Written English Corpus*. Tokyo.

Youngjun, J and C. Sunjoo (2014). *Split infinitives in English: A Corpus-Based Investigation*. Chung-Ang University.

Zwicky, A. (2004). *Obligatorily Split Infinitives*. Language Log. University of Pennsylvania. URL: http://itre.cis.upenn.edu/~myl/languagelog/archives/000901.html.