Quantitative Comparison of the Usage of Clefts between General Population and British MPs

Dimitris Boumparis
dimitrios.boumparis@student.uantwerpen.be

Abstract

The study at hand compares the use of *it*-clefts between the general population and ten Members of the British Parliament. The comparison in primarily made on the different properties of the *it*-cleft, following automatic extraction from the two corpora and manual annotation thereof. We explore how concrete social factors, such as the gender and the political affiliation of the MPs, affect how they use *it*-clefts. In a nutshell, politicians tend to use the cleft much more often than the general demographic in spoken discourse. In addition, in both corpora used, although the number of occurrences of the most common categories of cleft properties are very similar, the occurrences of other categories tend to differ significantly when used in political speeches or in everyday life. The most striking observation is the one relevant to the active givenness, namely whether the information contained in the cleft structure is directly relevant to the previous context. The fact that speeches delivered in the Parliament are almost always prepared ones and real aloud lets MPs have a more structure train of thought when talking compared to the general public in spoken discourse. Another interesting finding is the property of the clefts that has to do with the factuality of what is being said. A very low *p*-value in this case indicates that unclear factuality is present in both corpora but for different reasons.

Keywords: *it-clefts*, *corpus linguistics*, *Hansard corpus*, *British National Corpus*.

1. Introduction

One of a politician's most valuable assets is their ability to persuade others. In Ancient Greece, this technique was known as "rhetoric". The term is first mentioned in Plato's dialogue *Gorgias* where Socrates describes it as a pseudo-art, specifically a branch of Flattery, on the grounds that people employ it with the objective of deceiving others. While this does not seem to be far from the truth, people generally admire others who use language masterfully. For one to reach a level of using the language with great skill, one needs to learn various ways of expressing their ideas and arguments elaboratively. Throughout human history, people have developed different linguistic structures to

convey their messages. Following each people's unique course of history, languages have evolved in a plethora of ways, each one employing different ways of expressing essentially the same idea. One such structure is *it*-clefts which will be the main reference point for the paper at hand. To be more specific, we will be comparing its use by the general population and a limited number of British politicians. The three hypotheses we will be testing is (a) that politicians do not use the cleft more than the general public; (b) that there is no difference between the use of clefts in speeches of MPs and the general public spoken discourse; and (c) that other social factors (e.g., gender and political affiliation) do not play any significant role in the use of the clefts.

1.1 Structure of an *it*-cleft

Following Hedberg (1990), the structure of an *it*-cleft can be described as follows:

Cleft pronoun + copula + clefted constituent + cleft clause

Although, according to Patten (2012), *it*-clefts 'have a non-standard structure which appears not to conform to the general rules of the language', we will try to visualise their structure. Let's consider the following example:

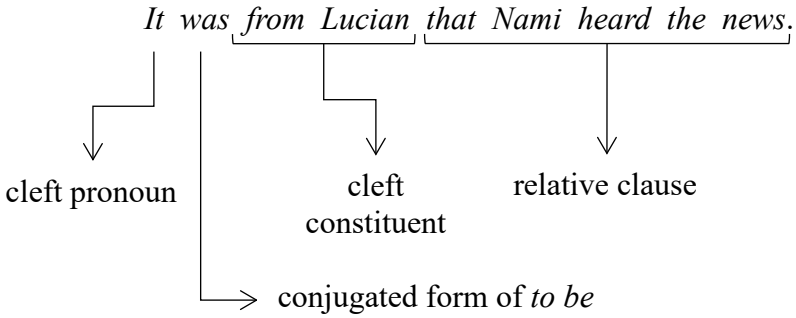

Figure 1: Breakdown of an *it*-cleft.

Generally speaking, the cleft consists of a pronoun (usually *it*, ergo the name of the particular type of clefts), followed by (a) a conjugated form of the verb *to be*; (b) the so-called cleft constituent (can be a noun, prepositional, or adverbial phrase); and (c) a relative clause (starting with a *wh*-pronoun or *that*).

1.2 Semantics of an *it*-cleft

Turning now to what meaning an *it*-cleft conveys, Patten (2012) highlights four different pragmatic properties thereof. First, as Lambrecht (2001: 489) notes, *it*-clefts are "used to prevent unintended predicate-focus construal of a proposition". Looking

back at our example, if it was not syntactically structured as a cleft, the sentence would be phrased as such:

Nami heard the news from Lucian.

This, however, allows for two different readings of the same clause, as shown below:

(a) *Nami* ***heard the news*** *from Lucian.* (b) *Nami heard the news* ***from Lucian***.

Instead, by employing a cleft clause, the focus – one property of an *it*-cleft that is elaborated upon in Section 3.1.3 – becomes unambiguous.

The second property is the presupposition. In our first example, we already know that *Nami heard the news* but not who she heard if *from*. In the case that Nami had not heard it *from Lucian*, it would mean that it was someone else who she heard the news from. In addition, the proposition that Nami heard the news is taken as a precondition to the assertion that she heard it from Lucian. The same would also apply in the case of negation.

Next, in addition to its fixed focus structure and existential presuppositions, the use of the *it*-cleft also signals exhaustiveness. In simple terms, the fact that Nami heard the news from Lucian in our example is indicative of him being the only one that would let her know about it. If the example was phrased in a non-cleft manner, then it would be indicated that maybe there was also another person that Nami could have gotten the news from. As Declerck (1988: 30) highlights, "it is clear that […] exhaustiveness is nothing else than 'exhaustive listing'".

The last property of the *it*-cleft, although more prevalent in negative clauses, is "contrast". The sense of contrast becomes greater if the set of potential values is limited and clearly defined. Our previous example would become: *It wasn't from Lucian that Nami heard the news*. This immediately implies that it was someone else who let Nami know, subsequently turning our example to: *It wasn't from Lucian that Nami heard the news – but from Shen*.

Taking all the above into consideration, we can see that there is also an interesting semantic aspect of the cleft which has to do with the factuality of the conveyed message. In other words, clefts are not necessarily consistent with the hearer's knowledge; It is the information that is presented to the hearer as factual, instead. This will be of importance to take into account when analysing the potential clefts of the corpus.

2. Corpus and data

2.1 Origin of the data

To explore the potential differences of how MPs and the public use *it*-clefts, data from two corpora were used. For the general population, the British National Corpus (BNC)[1] was used, while for the politicians, the Hansard corpus[2] was used. The former was compiled by Oxford University Press during the 1980s and early 1990s, consisting of 100 million words from a wide range of genres (e.g., spoken, fiction, magazines, newspapers, and academic). The latter, considerably larger with around 1.6 billion words in total, is a collection of the majority of the speeches delivered in the British Parliament from 1803 until 2005. However, in this study, we focused on a subset of 10 MPs, with a total number of approximately 660,000 words from almost 4,000 texts.

2.2 Cleaning and preprocessing

The two corpora were loaded into AntConc 3.5.6 (the latest – at the time of writing – 4.1.x version resulted in continuous crashes). It is important to highlight that the clefts found in the 'spoken demographic' category of the BNC corpus was selected, on the basis that (a) the MPs corpus contains also spoken data; and (b) the total count of the clefts found are relatively close, whereas in the case of 'spoken context' the available clefts would outnumber those found in the MPs corpus by 3,5 times. The concordance query used was a regular expression to match the syntax shown in Figure 2. The query returned 718 matches. After closer inspection of the potential clefts, only 96 were identified as real cases, as the majority of *it* were either referential or extraposition, which falls outside of the scope of this study, and are therefore ignored. This information can be found in the 'is cleft' column. A basic overview of the corpora can be found in Table 1.

		Clefts frequency		
	***N* words**	**Absolute**	**Relative**	**Normalised**
MPs	662,950	96	144.807	168.689
BNC	4,205,960	70	16.643	34.829

Table 1: A basic overview of the two corpora (BNC spoken demographic and the subset of the Hansard corpus containing the speeches delivered by 10 members of the British Parliament).

[1] Available at http://bnc.phon.ox.ac.uk/transcripts-html/.
[2] Available at https://www.clarin.ac.uk/hansard-corpus.

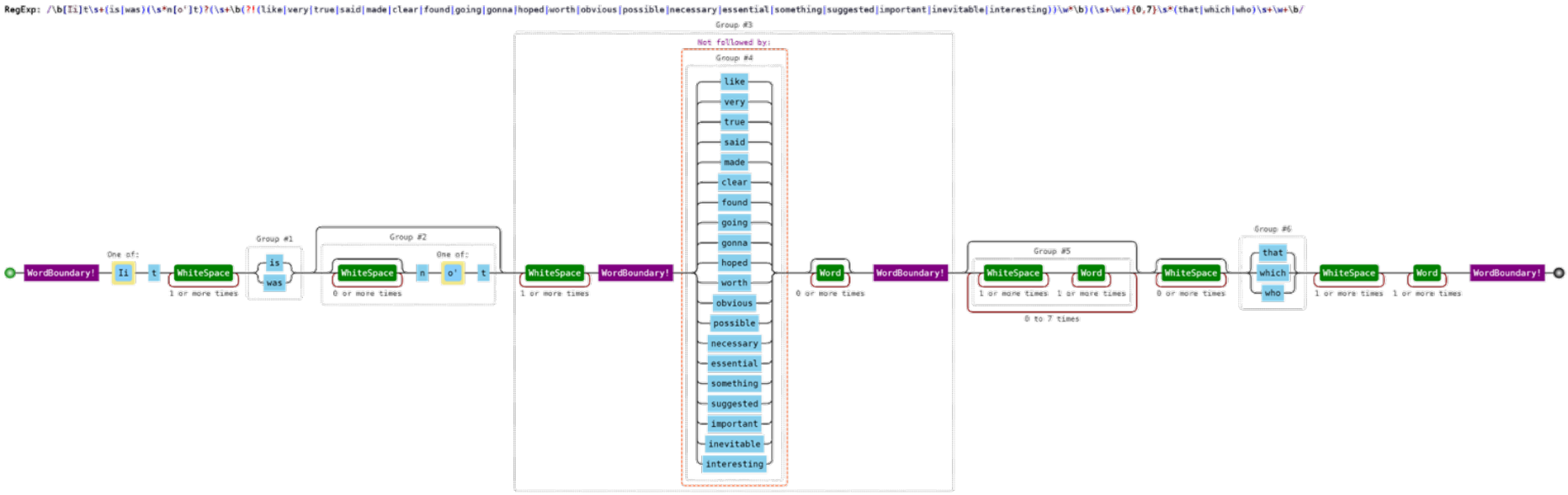

Figure 2: The regular expression used in AntConc to match the potential clefts in the Hansard corpus. Visualised using Regulex (https://jex.im/regulex/).

3. Methodology

3.1 Annotation of *it*-clefts

Following the manual annotation of the total number of matches, 96 of them turned out to be actual cleft structures. These clefts were then annotated in terms of four properties: givenness, factuality, type, and focus. We will now proceed to briefly analyse each property and define the types thereof.

3.1.1 Givenness

Givenness refers to what information is given in the relative clause compared to the already provided in the preceding spoken discourse. The types of givenness include:

- **Active**: the information present in the relative clause is also present in the immediately preceding context.
- **Inactive**: the information present in the relative clause is also present in the preceding context, but further behind.
- **Partial**: only a part of the information present in the relative clause is also present in the immediately preceding context.
- **New**: the information present in the relative clause is entirely new.

3.1.2 Factuality

As far as factuality is concerned, it refers to whether what is said in the cleft is truth or not in the ears of the speaker and hearer. Hence, the types of factuality are the following:

- **Fact**: the information provided in the cleft is considered true by both the speaker and the hearer, i.e., generally consensus on the statement.
- **Non-fact**: the information provided in the cleft is considered true by the speaker, yet the hearer has doubts as to whether it is true.
- **Unclear**: we cannot be sure about the validity of the information, or whether the speaker and the hearer consider it as true or not.

3.1.3 Focus

Focus has to do with the core of the cleft, particularly with what syntactic type the core is right before the relative clause starts. The available categories for focus are:

- **Noun phrase (NP)**: the core of the cleft is a noun or a noun phrase.
- **Adverbial**: the core of the cleft is either a prepositional phrase or an adverb.

- **Stranded presupposition**: the core of the cleft is an object with a stranded preposition in the relative clause.
- **Clause**: the core of the cleft is another subordinate clause.

3.1.4 Type

Type is the last property of the clefts that we will deal with in this annotation task.

- **Canonical**: the information within the cleft reveals new information about an aspect of the preceding discourse.
- **Informative presupposition (IP)**: the information introduced within the cleft is entirely new.
- **Performative**: the information within the cleft is a delicate issue and thus marked in a background manner to put emphasis thereon.
- **Unclear**: the information provided within the cleft is not enough for the type to fall under any of the above categories.

3.2 Frequency of *it*-clefts

Before diving into deeper analysis of the use of clefts, a frequency-based comparison between the two corpora is made. To get the numbers shown in Table 1, we used the following formulas to calculate the relative frequency of the use of clefts in each corpus. Then, we applied normalisation to account for the vastly large difference with regard to the number of words (X) present in the two corpora:

$$Relative\ freq. = \frac{number\ of\ cleft\ matches}{total\ number\ of\ words} \quad (1)$$

$$X_{normalised} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

3.3 Tools for analysis

Last but not least, a range of coding tools were used to extract plots that are later essential for the quantitative part of this analysis. Both Python and R programming languages were used, each to carry out a different part of the pipeline.

To begin with, the output from AntConc was transformed into a comma-separated values (CSV) file, using a script provided in class. This tabular transformation resulted in facilitating the Exploratory Data Analysis (EDA) on the data, to get a better understanding of how it is structured along with merging the matches with the metadata

of the speech they were found in. To this end, the DataFrame containing the exported results and the DataFrame with the word counts and the metadata on the MPs were merged using the Pandas `merge()` method in Python. Then, a Jupyter notebook with an Anaconda R virtual environment was created to further analyse the annotated clefts, as explained in the following subsections.

3.3.1 Mosaic plots

According to the R documentation[3], mosaic plots are used to visualise the standardised residuals of a loglinear model for the table by colour and outline of the mosaic's tiles. If the data passed in the function are 2-dimensional or more, they are then taken as a contingency table (Friendly, 1994), which will be helpful in the next subsection when dealing with the association statistics of each corpus with each of the clefts' properties. The same type of plots will be later used when analysing the potential impact of non-linguistic factors (see Section 4.3) on the use of clefts and the properties thereof.

3.3.2 Association plots with Pearson's residuals

Although mosaic plots are a very useful kind of visualisation, the differences displayed do not directly imply independence or not. To see whether there is a deeper correlation between what is shown in the mosaic plot, we employ the `assoc()` function, courtesy of the Visualise Categorical Data (`vcd`) library in R. This function is used to indicate deviations from a specified independence model in a contingency table.

For a contingency table, which applies in our case here as well, the signed contribution to Pearson's x^2 for cell $\{ij \ldots k\}$ is:

$$d_{ij\ldots k} = \frac{\left(f_{ij\ldots k} - e_{ij\ldots k}\right)}{\sqrt{\left(e_{ij\ldots k}\right)}} \tag{3}$$

where $f_{ij\ldots k}$ and $e_{ij\ldots k}$ are the observed and expected counts corresponding to the cell. Additionally, the residuals can be coloured depending on a specified shading scheme (Meyer et al., 2003). The `vcd` package also offers a range of residual-based shading. Some of them allow, e.g., the visualisation of test statistics, which we will be using to extract information about the significance of each of the clefts' properties and how they are used in the two corpora at hand.

[3] See more at https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/mosaicplot.

4. Results

After performing normalisation to properly rescale the frequencies of clefts in the two corpora (Table 1), we can conclude that MPs use clefts approximately 5 times as much as the general demographic (168.689 vs 34.829, respectively) in spoken discourse, hence disproving our first null hypothesis. It is also interesting to point out that even the other two major categories found in the BNC corpus (`spoken_context` and `written_to_be_spoken`) have also lower normalised frequencies (53.4693 and 45.6448, respectively). It is safe then to say that the presence of clefts in political speeches is a lot more frequent than in general spoken discourse.

4.1 Analysis on the MPs corpus

As far as the subset of the Hansard corpus we were assigned to is concerned, we are unable to run any mosaic plots showing the different uses of the cleft structures, as a mosaic plot, as highlighted earlier, requires at least 2-dimensional data. However, we can gather all occurrences of each category of cleft property we annotated and calculate the absolute and relative frequencies thereof. Both are shown in Table 2 below.

Givenness				**Factuality**		
Active	Inactive	New	Part	Fact	Non-fact	Unclear
49 (0.510)	20 (0.208)	10 (0.104)	11 (0.115)	**38 (0.396)**	27 (0.281)	30 (0.312)
Focus				**Type**		
NP	Adverbial	Clause	Stranded	Canonical	IP	Performative
57 (0.594)	30 (0.312)	8 (0.083)	1 (0.010)	**47 (0.490)**	39 (0.406)	4 (0.042)

Table 2: The absolute and relative frequencies (in parentheses) for each category of each cleft property. Note: The category with the most occurrences in the corpus is shown in bold.

What can be interpreted from the above table is that the most expected categories that are also the most common in cleft structures are present in the Hansard corpus. What remains to be seen is whether the other categories are different from the ones annotated in the BNC spoken demographic corpus and, if so, whether the differences are significant. To go about this, we loaded the second corpus in the Jupyter notebook and run a cross-corpus analysis consisting of mosaic plots (for visualisation) and association plots (for x^2 independence testing).

4.2 Cross-corpora analysis

The admittedly most interesting part of this project is the direct comparison between the two corpora: a small subset of the Hansard corpus, which contains *it*-clefts derived from the speeches of 10 MPs of the British Parliament, and the spoken demographic category of the BNC corpus. This will ultimately be the deciding factor as to whether we will reject our null hypotheses, as formulated in Section 1.

4.2.1 Cleft type

In Figure 3, we can clearly see that the canonical type is very similar between the two corpora, yet the other two categories are significant but opposite to each other; Whereas in the MPs sub-corpus, there are positive Pearson's residuals for the unclear type and negative for the informative presupposition (IP), the exact opposite applies for the BNC sub-corpus. This shows us that the political speeches examined as part of this project are mostly referring to the same thing as the previous context. The low p-value (< 0.001) confirms the significance of this difference and disproves our second H_0. If we were to attribute this to something, it would probably be due to the politicians often repeating their arguments using other utterances to give emphasis. On the other hand, a higher IP in BNC spoken data is to be expected as it could be due to information introduced in the cleft structure being entirely new – the train of thought in spoken discourse is often lost in spontaneous moments.

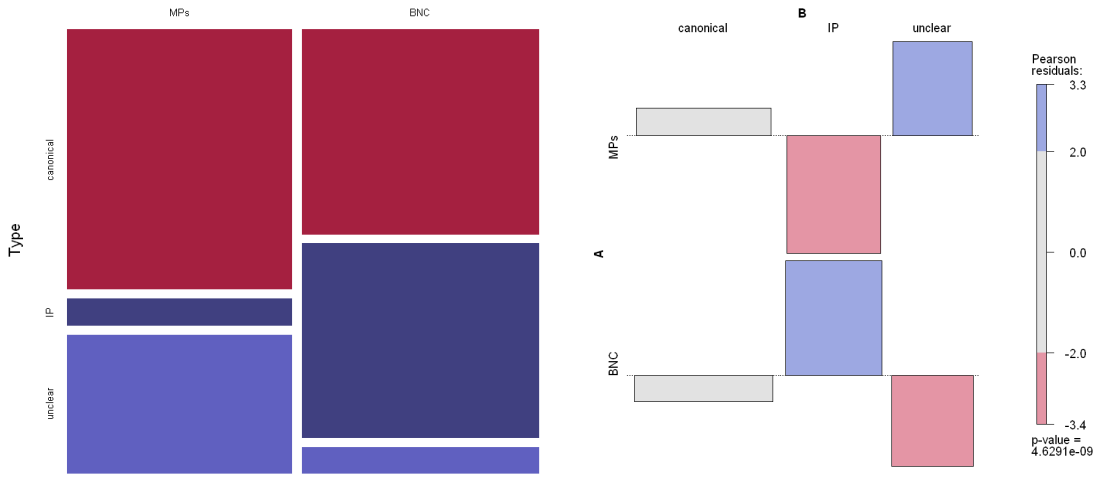

Figure 3: The mosaic plot (left) and the association plot (right) on the three main categories of the cleft type. The canonical type shows no significant difference, but that cannot be said for the IP and unclear types, where an opposite difference between the two corpora is evident.

4.2.2 Cleft givenness

For the second property of the clefts, givenness, we will follow the same path as before. By plotting the mosaic and the association plots, an interesting pattern emerges: There are positive Pearson residuals for the active givenness in the MPs data, and the new and partly givenness of the BNC data, along with negative ones for the active givenness in the BNC corpus. Again, the preparation that MPs have done prior to addressing the Parliament Bodies seems to be enough to differentiate this property in a significant way. Note from before that active givenness essentially means that what is being mentioned in the relative clause of the cleft structure is directly associated with the previous context. In the case of political speeches, that seems to almost always be the case, as they structure their arguments in such ways to feel natural and logical when uttered. Conversely, the general population talks in mostly unstructured and spontaneous ways, which is confirmed by the association graph, along with a very low *p*-value (< 0.001).

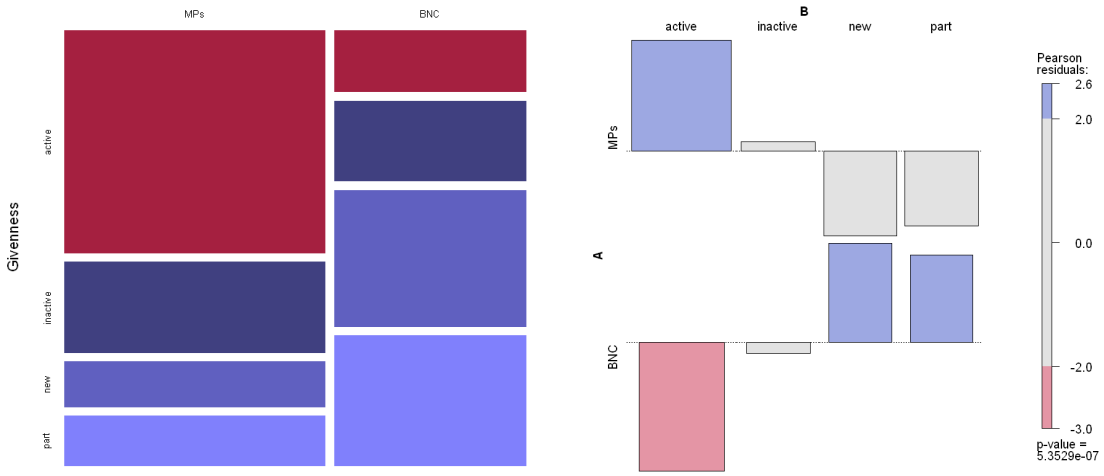

Figure 4: The mosaic plot (left) and the association plot (right) on the four main categories of the cleft givenness. The inactive category shows no significant difference, but the opposite is true for the active givenness, and the new the partly givenness for the BNC corpus.

4.2.3 Cleft factuality

Next is the factuality, which is the most challenging to annotate. What is considered as a fact is very subjective in general, especially when dealing with the topic of politics. Having taken that into account, the plots in Figure 5 suggest that, given the low *p*-value (< 0.001), there is significant difference in the unclear category. In other words, the context before and after the cleft match is not enough for us to determine whether what is being uttered is actually true or not. As pointed out earlier, fact and non-fact could be

alternatively translated into ‘consensus’ or ‘no consensus’ between the speaker and the hearer. In our case of the Hansard sub-corpus, the only examples that were classified as facts are those that contain figures, historical events, or data. Every other case should be considered as non-fact when it is obvious that there is no consensus (e.g., “It is only the Labour party that is still way back in the tired old 1960s”), or unclear when there is not enough data to support neither the presence not the absence of consensus (e.g., “Far from being a smokescreen, it is a response to the representations that we have received as a result of the introduction of national testing”). It is interesting, nonetheless, that the occurrences of unclear clefts in the case of the MPs are far less than the ones in BNC. On the contrary, there are negative residuals for the unclear clefts of the MPs and positive ones for the same category in BNC. This could be attributed to the fact that in our case there is a significantly lower number of unclear clefts regarding the MPs which is also confirmed by looking at the Table 2.

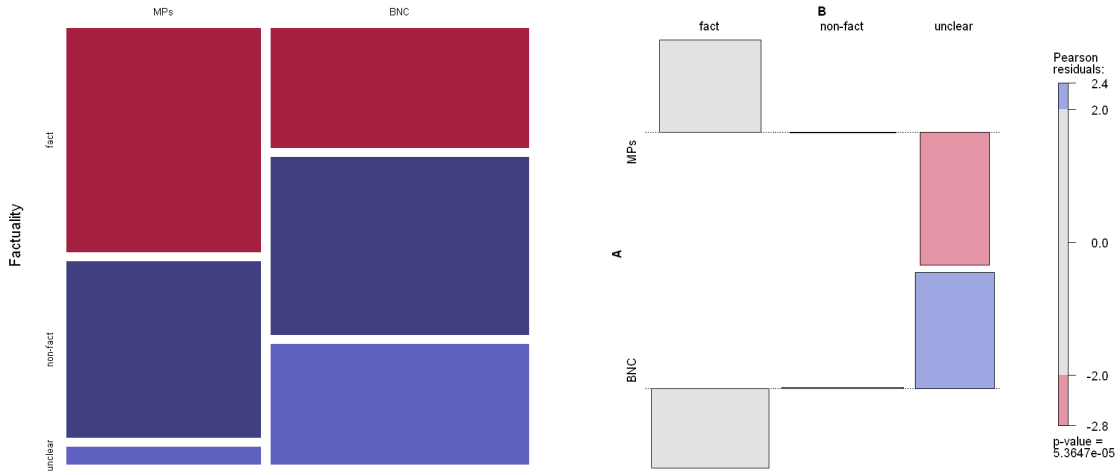

Figure 5: The mosaic plot (left) and the association plot (right) on the three main categories of the cleft factuality. The fact and non-fact categories show no significant difference, but the unclear one is significantly different for both corpora.

4.2.4 Cleft focus

Last but not least, we have the focus property, which constitutes the sole syntactic aspect of the four properties we are examining. It essentially has to do with the term that the cleft refers to (*cleft constituent*), found between the cleft pronoun and the conjugated form of *to be*, and the relative clause that follows it. this property is the easiest to spot, as it does not display any complications with the semantics of the context but, rather, with the grammar and syntax of the terms which the cleft consists of. In the case of

focus here, conversely to all the aforementioned properties, we get a relatively low *p*-value ($0.001 < 0.003 < 0.005$) which indicated that the differences are still significant, albeit no Pearson residuals are spotted in the association plot (Figure 6). However, our attention is this time drawn to the mosaic plot, as the MPs show a clear preference for clauses and adverbial phrases whereas the general population prefers using noun phrases (NPs) instead. This is expected to some extent, as, again, politicians prepare their speeches as well as are generally more proficient language speakers than the average population. Additionally, by using adverbial phrases – which, as previously mentioned, consist of adverbs, prepositional phrases, and very seldomly of adjectives – MPs are in a position to convey their message in an even more profoundly masterful way to impress and convince the audience.

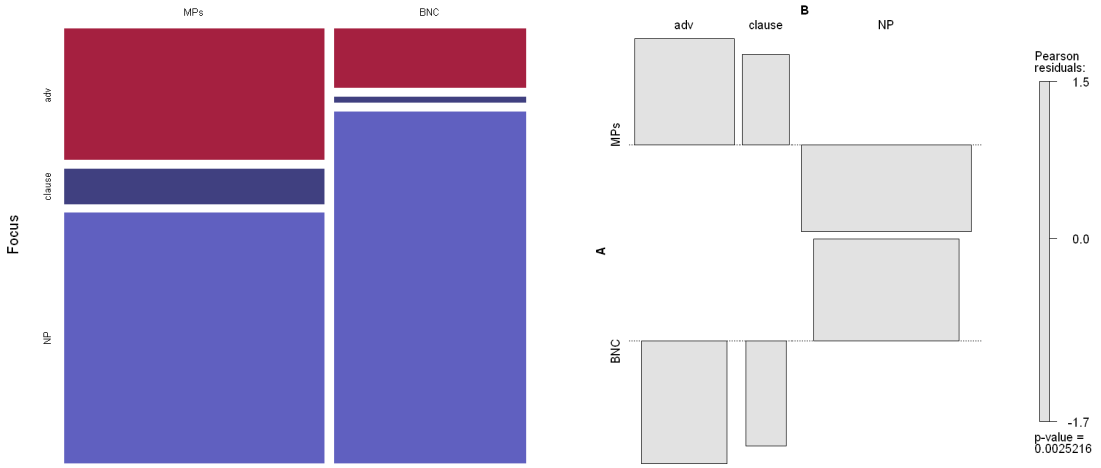

Figure 6: The mosaic plot (left) and the association plot (right) on the three main categories of the cleft focus. Although no residuals are evident, there is still something to be said about the choices of the syntactic patterns found in each of the two corpora, where MPs use more complex structures such as prepositional phrases and clauses compared to the general demographic.

4.3 Analysis of social factors

Finally, we attempt to show whether other social factors affect are correlated with the use of clefts in the Hansard sub-corpus. We will examine the gender and the political affiliation of the MPs, again using mosaic and association plots, as before.

4.3.1 Gender

The gender variable is a binary one (male or female) and is very common to examine in linguistic annotation tasks, as it is believed that it plays a role in how language is used by males and females. Unfortunately, in our case, the data are extremely imbalanced: 83 male vs 13 female speeches. By plotting both mosaic and association

plots, the only relatively low *p*-value we get back has to do with the cleft type (0.08). Although it is enough to make us reject our third and final H_0, it is crucial to point out that we need to take these results sceptically, as the imbalance in the data is massive.

4.3.2 Political affiliation

Last, we will take a look at whether the party each MP belongs to plays any significant role in how they use clefts. In our Hansard sub-corpus, there are 77 cleft occurrences derived from speeches of Conservative MPs and 19 from Labour MPs. The imbalance in this regard is evident here as well. The only *p*-value that indicates significance in said case is 0.038 regarding the factuality, yet no residuals are observed.

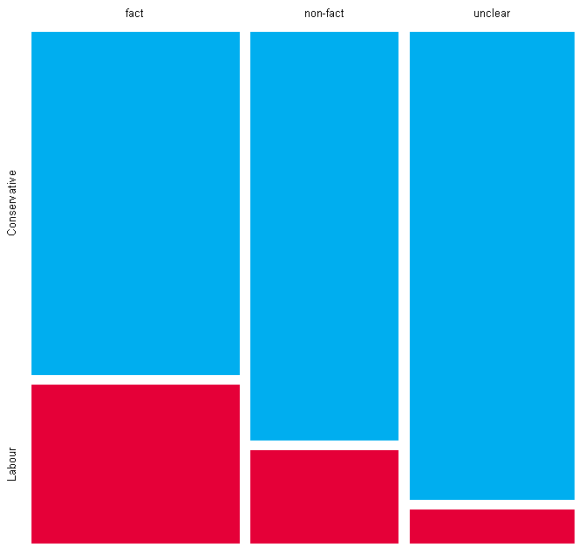

Figure 7: Mosaic plot showing the cleft factuality relative to the two parties found in the sub-corpus.

5. Discussion of issues

Most of the issues in this project have to do with the annotation process. Especially when dealing with the factuality of the cases, it was not trivial to assign the proper type of factuality, as most of the times the surrounding context was not enough for us to fully comprehend the issue being discussed. However, it raises another important question, whether the cleft is being used deliberately to deceive the hearer in this regard. In other words, politicians may employ this structure to make their point more difficult to grasp in terms of semantic meaning, as we saw that it can be used to convey slightly different tone of the utterance. Another caveat is the large number of false positives, which could be potentially solved by leveraging syntactic dependencies in combination with RegEx.

6. Conclusion

Taking all the above into account, we are able to say with certainty that the cleft is a much more complex structure that it initially seems to be. Its syntactic and semantic variations result in it requiring a lot of attention and annotation, including many false positives due to its primarily common (or simple enough) structure. We set out to give a satisfactory answer to three main questions. First, we proved, thanks to the normalisation frequencies, that politicians tend to use the *it*-cleft a lot more often than the general public. Second, we showed that there is a plethora of differences in how MPs and the general population use said structure, even significant ones along the way. The clefts focus and factuality turned out to be the most significantly different properties of this complex linguistic structure. Finally, we tried to show that there are also non-linguistic factors that affect the use of the cleft, without yielding equally convincing results but enough to reject the third null hypothesis.

This study is an interdisciplinary one, as it focuses on three pillars upon which the analysis takes place: a linguistic, a statistical, and a programming one. This goes to show that, when combining Humanities with modern tools and quantitative approaches, one can take a deeper grasp of what it is to use the language in a particular way and reveal the underlying patterns that more often than not never meet the eye. The way people use this living organism we all call ‘language’ may not be too different after all, but it portrays how people think and feel the need to belong in a larger group, nonetheless. Our sample of politicians, albeit very limited, seems to indicate that they feel the need to show higher speaking skills along with the power of their arguments by using the *it*-cleft, among other things of course. This could mainly be attributed to the objective that each group uses language. The primary aim of a political speech is to guide the public opinion, and it seems that employing the *it*-cleft structure is a means of achieving this in the eyes of the Members of Parliament and the unsuspecting hearer.

References

Friendly, M. (1994). Mosaic Displays for Multi-Way Contingency Tables. *Journal of the American Statistical Association*, 89(425), 190–200. https://doi.org/10.2307/2291215

Meyer, D., Zeileis, A., Hornik, K. (2003). Visualizing independence using extended association plots. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, K. Hornik, F. Leisch, A. Zeileis (eds.), ISSN 1609-395X. https://www.R-project.org/conferences/DSC-2003/Proceedings/

Patten, A. L. (2012). *The English it-Cleft: A Constructional Account and a Diachronic Investigation*. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110279528