Journal of
language | works

Vol. 10, No. 1, 2025

Than Meets the Eye:

A Corpus Study of English Comparison Particles

Anna-Merete Thinggaard
Aarhus University
201906951@post.au.dk

Abstract

This paper analyses data from the BNC to examine the distribution of the comparison particles than and as in their use as either prepositions or as complementisers, signalled through the case on the following pronoun. The comparison particles are used as prepositions in 85.0% of the occurrences, but considerable variation is found across the individual pronouns. Notably, the 1st person singular is barely used in the nominative case, i.e. as a complementiser, and the 3rd person singular feminine and masculine pronouns exhibit the highest rate of nominative uses. It is suggested that the limited use of 1st person singular nominative may be attributable to genre effects of the data available in the BNC whereas the preference for 3rd person singular nominative remains unclear. Additionally, the data suggests that both formality levels and genre affect the distributions of prepositional and complementiser use as more formal texts have a higher percentage of nominative pronouns and less formal text and spoken language a lower percentage.

Keywords: corpus linguistics, BNC, English comparison particles, complementisers, prepositions

Introduction

In English, when a comparison particle is followed by a pronoun, that pronoun can be either nominative or oblique. For example, if my friend finished a degree in biology last year, does she know more about moss *than I* (nom.) or *than me* (obl.)? Some people may have a strong preference for one of these pronoun cases, but both versions exist in the English language. Moreover, this variation has long been the subject of prescriptive debate (e.g. Lowth 1763, 157–59; Brittain 1788, 81–82) – a debate that continues to spark confusion about which form to use, as reflected by its continuous treatment in online fora, blogs, and popular articles (e.g. Dauenhauer 2021; Britannica 2024; Doyle 2012; Perlman 2019). In this paper, I consider the linguistic variation found in English comparison constructions and how it may be investigated with the help of corpus linguistics. Specifically, I focus my investigation on the distribution of nominative and oblique pronouns following comparisons with *than* and *as* in the British National Corpus (henceforth the BNC). The overall distributions, variation between individual pronouns, and variation across genre are all taken into consideration. In section **Fejl! Henvisningskilde ikke fundet.**, I provide the theoretical background for my paper, both as relates to corpus linguistics and to comparison particles. In section 0, I discuss the methodology, and in section 0, I present the results which I discuss in section **Fejl! Henvisningskilde ikke fundet.**. In section 0, I provide a conclusion.

Corpora, Corpus Linguistics and English Comparison Particles

What Is a Corpus?

The first question to ask when defining corpus linguistics, is what a corpus itself is. A very simple definition of a corpus is a collection of texts in written, spoken, or signed form (Jones and Waller 2015, 5; Voelkel and Kretzschmar 2021, 135). However, if this was all, corpora would not actually be that useful for researchers, and corpora to be used for linguistic research will typically be more elaborate than that. For a prototypical linguistic corpus, the texts need to be authentic, systematically collected, representative and balanced to mirror the intended language variety, register, or genre. The corpus should preferably contain various metadata about the texts, and these texts should be linguistically annotated (Voelkel and Kretzschmar 2021, 139). As for modern linguistic corpora, they are machine-readable and digitized, making it much easier for the researcher to make use of the data. There is a multitude of corpus types. For example, corpora can be comprised purely of one type of data, such as written or spoken data, or be multimodal; they can be mono-, bi- or multilingual; and they can be sample corpora, fixed in size and context, or monitor corpora, which are continuously updated and expanded. Corpora may also be either synchronic or diachronic as in the case of historical corpora. I will not go into detail with each aspect, but in section 0, I give a more detailed description of a well-known corpus, i.e. the BNC, including its design, size, and composition.

The most difficult thing to accomplish when creating a corpus is making it both representative and balanced. For a corpus to be representative means that it reflects the entire span of variation present in the language variety that the corpus is meant to be representative of. Manning and Schütze (1999, 119) define a sample as representative if what we find when looking at the sample will also hold for the general population that it is meant to represent. That means including careful considerations of the age, gender, social class etc. of the people the texts originate with but also including a wide variety of text genres and sources of the language variety in question. Mukherjee (2004, 114), Love (2020, 34), and Leech (2007, 140) all share the view that representativeness is not actually achievable, but rather an ideal to optimistically strive toward, and demographic data can help us in that pursuit. For a corpus to be balanced means that it "is stratified so that it corresponds to the proportion with which the respective strata would emerge in the language" (Voelkel and Kretzschmar 2021, 139). This means that the distribution of the texts corresponds to the natural distribution of the language variety. In practice, that is not easy to achieve. For one, it can be difficult to know exactly how language is distributed. Second, it can be difficult to obtain samples matching this exact distribution. For example, most language is spoken, but spoken data is more challenging and time consuming to obtain (Voelkel and Kretzschmar 2021, 143). Moreover, while bigger is sometimes better when it comes to representativeness since it includes more data, over- or underrepresentation will still skew the balance, which in turn could skew findings based on the corpus. Davies (2011, 68) notes that this issue is especially challenging when it comes to creating historical corpora, because the different genres do not necessarily exist in the same ratio across time. Additionally, historical corpora face the challenge of being able to properly verify the dating of texts as well as procuring texts of the less formal kind. Failing to maintain the same ratio across time will mean that we cannot properly compare the data, and the results and following analysis will be skewed, effectively making the corpus much less useful. These issues are, however, not just important in terms of diachronic corpora, but also when comparing corpora synchronically as any proper comparison needs comparable data.

What Is Corpus Linguistics?

Corpus linguistics can be defined as "(i) the study of the properties of corpora or (ii) the study of language on the basis of corpus data" (Gries 2011, 83). Here I will be focussing on the second definition, that is, corpus linguistics as a means of studying language by using corpora. There is some debate about whether corpus linguistics is more of a methodology or a distinct subbranch of linguistics on par with syntax, sociolinguistics etc. (Gilquin 2010, 5–7; Voelkel and Kretzschmar 2021, 133–34). Here, I take the approach that it is a methodology along with for example Leech (1992, 105–6), Gries (2011, 83) and Davies (2011, 65) as I, following the definition above, consider it a way of obtaining linguistic data to be analysed according to a specific theory. Furthermore, it can be combined with different theoretical approaches, and studies can be carried out with the help of corpus linguistics within various fields of linguistics such as morphosyntax (e.g. Hundt 1998; Willis 2017; Ehlers 2020) or pragmatics (e.g. Taylor 2016; Culpeper and Gillings 2018) etc.

English Comparison Constructions

In English, when the comparison particles *than* and *as* are followed by a determiner phase (DP), this DP can either have nominative or oblique case, depending on the function of the comparison particle as can be seen from examples such as 0 and 0.

(1) a. He is taller [CP than [IP I.NOM / *me.OBL (am)]]. (adapted from Jäger 2019, 5)
b. He is taller [PP than I.NOM / me.OBL].

(2) a. He is as tall [CP as [IP I.NOM / *me.OBL (am)]].
b. He is as tall [PP as I.NOM / me.OBL].

The variation between the two case forms depends on whether the comparison particle functions as a complementiser (or subordinator, subordinating conjunction) introducing an IP (a clause) with the pronoun as its subject as in 0a and 0a or as a preposition taking a DP complement as in 0b and 0b. This variation of the two different usages of the comparison particles has been present for a long time. According to Visser (1963, 1:249–50), the prepositional option may have been introduced into the English language as early as the second half of the 16th century. Furthermore, this variation has often been subject to debate over grammatical correctness, though the arguments can at times seem somewhat arbitrary. Lowth (1763, 157–60) for example only considers constructions such as 0a 0a with nominative forms correct, but makes an exception for the interrogative pronoun *who*, which he prefers in its oblique form *whom* following *than*. He gives the example repeated in 0 where the entire prepositional phrase *than whom* has been moved to the front of the relative clause.

(3) Beelzebub, [PP than whom], except Satan, none higher sat.

(1667, J. Milton, *Paradise Lost*, ii. 299 adapted from Lowth (1763, 160))

In a written survey study, Quinn (2005) sought to investigate the distribution of pronoun case forms in various constructions in New Zealand English. Her survey concerning *than*-comparisons was answered by 41 participants in 1997 (Quinn 2005, 78–79). The questions were contextualized to elicit responses that would be natural in a casual, conversational setting in an attempt at minimizing the influence of perceived grammatical norms in the answers (Quinn 2005, 94). The result from this survey is repeated in Table 1.

Table 1: Distribution of pronoun case forms following *than* by New Zealand English survey participants according to Quinn (2005)

	NOM	OBL
I/me	12.20% (5)	87.80% (36)
she/her	7.32% (3)	92.68% (38)
he/him	0.00% (0)	100.00% (41)
we/us	0.00% (0)	100.00% (41)
they/them	0.00% (0)	100.00% (41)

Table 1 shows that the New Zealand participants displayed a strong preference for oblique pronouns after *than* with only 1st person singular and 3rd person singular feminine having any nominatives at all. This is consistent with the broader findings of Quinn's study, which also examined coordinate DPs, *it*-clefts, and modified DP-subjects. In these contexts, it was found that people tended to prefer nominative for 1st person singular *I* over the rest of the nominative pronouns, but that the 3rd person singular *he* and *she* were preferred over both 1st and 3rd person plural *we* and *they* (Quinn 2005, 147). The present study investigates whether these conclusions hold for British English as well as for various genres and contexts.

The BNC

The British National Corpus (BNC; 1994) is the online version of the BNC available through English-Corpora.org, and it is the version of the corpus which will be used and discussed here. The BNC is a sample corpus consisting of 100 million words and is composed of text samples up to 45,000 words. New text is not added. As with all sample corpora, a disadvantage is that the data quickly becomes dated and there is always the risk that the language may change somewhat between the time the corpus is created and its use for later linguistic enquiries. It is a synchronic corpus, and a general corpus which is meant to represent British English as a whole, not any specific variety, register etc. For this reason, it is also a monolingual corpus, meant to only contain English produced by speakers of British English. Finally, it is a mixed corpus, as it contains both spoken and written language.

The distribution of spoken and written language is not equal, however. Out of the 100 million words, 90% is from written text, and 10% is transcriptions of spoken data, which is hardly representative of the language actually produced as most people tend to speak more words than they write. The limited amount of spoken text was largely due to constraints on time and budget since recording and transcribing speech is costly and time consuming, and it was decided that 10 million words would still suffice for viable empirical statistical data (Meyer 2023, 44; University of Oxford, n.d.). When collecting the spoken data, it was decided that it should include partly informal encounters which were socially stratified as well as more formal encounters, including broadcast interviews, lectures, meetings etc. (Aston and Burnard 1998, 31). Though great effort was made to ensure the representation of different age groups, genders, social class and region, the distribution was not balanced in proportion to the actual population; instead an equal number of speakers from each group was included (Aston and Burnard 1998, 31; Meyer 2023, 24–25; Love 2020, 36). The written data was chosen according to the three criteria of domain, time, and medium with a target percentage set for each subsection. Most of the data included were from books, but also included periodicals, manuscripts and miscellaneous sources (Aston and Burnard 1998, 29). Due to the variety included, the written data is largely as representative as the spoken data, but because of this attempt at proportionalising the included texts, the written data, which constitutes 90% of the corpus, is more successfully balanced.

METHODOLOGY

In order to consider the advantages and disadvantages of corpus linguistics, I will be examining the distribution of sentence final *as* and *than* followed either by a nominative or by an oblique personal pronoun, since this can show us whether they are used as complementisers or prepositions. This will serve as a preliminary study of the distributions of the two constructions in the BNC (1994) across pronouns and genre for the purposes already stated.

According to Jespersen and Haislund ([1949] 1974, VII:236), the oblique case is so universal that it must be considered the normal use, and in the survey of New Zealand English by Quinn (2005, 79; 143) pronouns in *than*-comparisons are almost always used in oblique. Nevertheless, a handful of speakers did choose to use the 1st person singular nominative *I* (5 of the 41 speakers) and the feminine 3rd person singular nominative *she* (3 of the 41 speakers). Quirk et al. (1985, 661) argues that prescriptivists tend to prefer the nominative in these constructions and according to Huddleston and Pullum (2002, 1113), nominative is used in formal style. Since the BNC contains texts from various domains and as a result should include examples of varying levels of formality, we should expect to find a reasonable number of occurrences with both nominatives and oblique pronouns, though I, for the reasons outlined above, expect a majority of oblique pronouns in my searches. With reference to the results of the survey study, I would also expect more nominatives for singular pronouns, especially 1st person singular, than for the other pronouns.

I searched for sentence final occurrences of three common types of comparison constructions. The first is *as/so* ADJ *as* PRO, which expresses equality between two entities. An example of this type of comparison from my data is given in 0a. The second type searched for involves an adjective in the comparative degree and *than* as the example in 0b shows. The third type consists of *more/less* ADJ *than* PRO. An example of this type is provided in 0c. Both constructions with *than* express inequality between two entities.

(4) a. You must be at least [as intelligent as me]. (all examples from the BNC)
b. Some of them were even [younger than he].
c. Can't get [more common than me]!

The reason for having the comparison particles be followed by a simple personal pronoun-DP is that the personal pronouns are the only DPs where case is visible in English. However, since the 2nd person singular and plural pronouns (*you*) as well as the neuter 3rd person singular pronouns (*it*) are ambiguous between nominative and oblique, I have left them out of my search. That means I have searched for the nominative pronouns *I, we, she, he, they* and the oblique *me, us, her, him, them*. One of the reasons for searching for sentence-final occurrences is that searching for sentence-medial occurrences will include too many non-applicable examples such as 0 in which *as* is used in another one of its functions introducing a time adverbial.

(5) [$_{IP}$ [$_{IP}$ He looked so frail] [$_{CP}$ as [$_{IP}$ I watched Gavin help him out of the car]]] (BNC)

Further, this restriction helps provide the examples where there is optionality, since the comparison particles must be complementisers if they introduce a full IP, that is, if the DP is followed by a verb such as in 0.

(6) Is it as important [$_{CP}$ as [$_{IP}$ I [$_{VP}$ suggest]]]? (BNC)

Unfortunately, it is not possible to search for something being sentence-final in general in the BNC. That means that such a search must be conducted manually, as it is still possible to search for general punctuation. For my purposes, a sentence can be defined as something “ending with a ‘.’, ‘?’, or ‘!’”

(Manning and Schütze 1999, 134). As a result, I searched for punctuation that was not a comma, which was the best option available, but it meant that 15 occurrences with colons, semicolons, or brackets as punctuation had to be removed as they were not actually sentence-final. Furthermore, one occurrence was removed because it was a quote which was used in two different articles, simply repeating the same occurrence.

After conducting the searches, I divided the results into four categories depending on their genre tags. These categories were: 1: Formal and academic writing; 2: Fictional and literary texts; 3: Media, news, and popular science; and 4: Spoken language. The only two genre-tags that were not automatically added to a category was W_misc (miscellaneous written texts) and W_religion (written texts related to religion) because they may fall under very different categories depending on their style and purpose. Two instances of W_religion and one of W_misc were added to Category 1, and 1 instance of W_religion and 8 instances of W_misc were added to Category 3. One example of W_non_ac_humanities_arts (non-academic texts relating to humanities and the arts) was added to Category 2 instead of Category 3 because it contained an example of how to use dialect in creative writing. Though the categories are not of equal size, they manage to represent the variety of genres found in my dataset and create a spectrum of formality which should help investigate whether the assumptions by Quirk et al (1985, 661) and Huddleston and Pullum (2002, 1113) hold. The expectation is that the highest number of nominatives are found within Category 1, with the lowest number to be found in Category 4.

Results of the Corpus Searches

The searches in the BNC resulted in 226 relevant occurrences after the adjustments described in section 0 above. Table 2 shows the distribution of occurrences with a nominative (*I*, *she*, *he*, *we*, *they*) and an oblique (*me*, *her*, *him*, *us*, *them*) personal pronoun following *than* and *as* in the three types of comparison constructions presented in 0 above.

Table 2: BNC sentence-final *than*/*as*+NOM/OBL

Construction type	**NOM**	**OBL**	**Total**
***as*/*so* ADJ *as* PRO**	38.5% (15)	61.5% (24)	100.0% (39)
ADJ.CMP *than*	8.4% (14)	91.6% (152)	100.0% (166)
***more*/*less* ADJ *than* PRO**	23.8% (5)	76.2% (16)	100.0% (21)

The results in Table 2 show that there is some variation between the three types of constructions I searched for, though all constructions predominantly occur with oblique pronouns. Out of the three constructions, the one with *as* shows the highest frequency of nominative with 38.5%, whereas the most common search with 166 occurrences out of 226, a comparative adjective + *than*, has the strongest preference for oblique with 91.6%. Table 3 shows the distribution of nominative and oblique following each individual personal pronoun and in total for all the searches combined.

Table 3: BNC sentence-final *than*/*as*+NOM/OBL (all searches)

	NOM	**OBL**	**Total**
I/me	1.0% (1)	99.0% (96)	100% (97)
she/her	31.7% (13)	68.3% (28)	100% (41)
he/him	26.9% (14)	73.1% (38)	100% (52)
we/us	16.7% (2)	83.3% (10)	100% (12)
they/them	16.7% (4)	83.3% (20)	100% (24)
Total	15.0% (34)	85.0% (192)	100% (226)

Out of the 226 occurrences, 85.0% of the comparison particles were followed by an oblique personal pronoun and the remaining 15.0% were followed by a nominative. This is not completely unexpected with regards to the expectations outlined in section 0 above as it shows a clear preference for oblique pronouns but still contains a fair number of nominatives. Additionally, the results in Table 3 show that there is a difference in the distribution of nominative and oblique with respect to the individual pronouns. For the 1st person singular, only 1.0% of the occurrences had the nominative form *I*. For both the 2nd and 3rd person plural pronouns, 16.7% were the nominative *we* and *they*, and 83.3% were the oblique *us* and *them*. However, these were also the two categories with the least total occurrences, 12 and 24 respectively. As it is, the difference between having 1 or 2 occurrences in nominative might not be as big as the percentages show. To be surer of the distributions of the pronouns that occur with lower frequency, further research is needed, perhaps using a second, comparable corpus of British English. Such a comparison, would, of course, also increase the overall reliability of the distributions of the rest of the pronouns. For the masculine 3rd person singular, 26.9% of the occurrences are nominative and 73.1% are oblique, and for the feminine 3rd person singular 31.7% are nominative and 68.3% are oblique. While the use of *she* to some extent was expected in line with Quinn (2005), the amount of nominatives is quite high at almost a third of the occurrences. The same applies to the masculine 3rd person singular with almost a quarter of the occurrences being nominative. There are clearly most nominatives with the 3rd person singular pronouns overall. In fact, taking them out of the equation would leave a distribution of nominative and oblique at 5.7% nominatives and 94.3% obliques. Most surprising however, the pronoun that Quinn found the highest nominative use for was 1st person singular which is the one with the least nominative uses in the BNC at only 1%.

Finally, Table 4 shows the distribution of the results across genre, divided into the four categories outlined in section 3.

Table 4: BNC sentence-final *than*/*as*+NOM/OBL across genre

Category	NOM	OBL	Total
Category 1: Formal and academic writing	21.7% (5)	78.3% (18)	100.0% (23)
Category 2: Fictional and literary texts	19.8% (24)	80.2% (97)	100.0% (121)
Category 3: Media, news, and popular science	12.2% (5)	87.8% (36)	100.0% (41)
Category 4: Spoken language	0.0% (0)	100.0% (40)	100.0% (40)

In line with expectations, the highest proportion of nominative pronouns occurs in Category 1 at 21.7% of the cases. However, the difference between Category 1 and Category 2 is relatively small, with the latter consisting of 19.8% nominative pronouns. Category 3 contains only 12.2% nominative pronouns, while Category 4 contains none. This indicates that formality and genre affect the choice of whether to use the comparison particles prepositionally or as complementisers. Furthermore, although it was expected that Category 4 would have the lowest proportion of nominative pronouns, it is notable that there are none at all. This suggests that using the comparison particles as complementisers is, at the very least, rare in spoken language.

As Expected?

The data from the BNC shows us the difference in the usage of the comparison particles *than* and *as* as prepositions and as complementisers in British English. As expected, there is a predominance of oblique personal pronouns following the comparison particles, but there are still nominatives present. This means that while there is a clear preference for using the comparison particles in the prepositional function, the complementiser function continues to be in use. At the same time, the data shows that there is a difference between *than* and *as*, as well as between the individual pronouns. This variation is interesting as one particle or one pronoun should not be better suited to a specific function than the others.

While there is a difference between the distributions of nominative and oblique between the three types of constructions searched for, the one with the highest number of results, a comparative adjective followed by *than*, is also the one with the highest number of obliques. Though the search with the second-largest number of results is the one with *as*, which also has the lowest number of obliques, it is still possible that there is a frequency effect since the two other searches both contain *than*. In any case, it seems that *than* is more likely to be used prepositionally than *as* is.

As regards the individual pronouns, the results in the current study differ from those reported by Quinn (2005) for New Zealand English. Higher nominative percentages were found for 1st and 3rd person plural, but 1st person singular was much lower than Quinn's results showed. The reason for these differences could be the fact that her survey study only included New Zealand English speakers, whose language will vary to some extent from British English, with its speakers perhaps being less inclined to use the comparison particles as complementisers. However, taking a quick look at the comparison constructions in the corpus of Global Web-based English (GloWbE, Davies 2013) the data does not suggest that to be true. At the same time, it should be noted that there are only 81.4M words of New Zealand English in GloWbE, whereas there are 387.6M words of British English, and I have not looked closely at the balance and representativeness of the data there. However, in GloWbE, web pages are used, ideally representing informal language with 60% of the data from each language being from informal blogs and 40% of them from a variety of more formal genres (Davies 2015). Of course, while both corpora are sample corpora, the data in the BNC, and that of Quinn's study, is older than that in GloWbE, which means that the balance between New Zealand English and British English may be different now than it was then. Still, the GloWbE data should be enough to give us an idea of whether the preference toward prepositional use with oblique could be stronger in New Zealand English than in British English. It would also be possible to replicate the current study in GloWbE to reveal whether there is substantial variation across a number of varieties of English, not only including British and New Zealand English, but also for example American, South African, or Indian English.

Adding together the frequencies of the three constructions outlined in section 0 above followed by a nominative personal pronoun provides a combined frequency per million words of 0.27 for British English and 0.37 for New Zealand English. Of course, there is no saying whether the construction occurs exactly the same number of times in both corpora, but at least it does not indicate that comparison constructions with nominative (that is, in the complementiser function) should be that much rarer in New Zealand English than in British English. If anything, according to this, there ought to be more.

Why there should be this difference then, between Quinn's findings and the results from the BNC, particularly as regards the 1st person singular, may instead be due to differences in data type and genre. Whereas Quinn's study is based on a written survey targeting informal language, the BNC contains many different types of language. As can be seen from the results in Table 4, genre does seem to have an effect on the choice between using the comparison particles as complementisers or prepositions, especially when it comes to spoken language, where the prepositional function is exclusively used. Furthermore, depending on how strong the effect of the prescriptivist tendencies toward nominative is, it may affect how people act and speak when asked to make a conscious decision as opposed to what people produce subconsciously. That some degree of nominative pronouns is present in Quinn's study could then have been due to the fact that the survey, though describing spoken language, was conducted in written form. The nature of a survey may heighten the attention to the language used, and prescriptive norms could have had some effect as well.

One of the benefits of corpus data is that the texts are authentic and that the language is usually produced spontaneously in natural settings (Voelkel and Kretzschmar 2021, 134). Though most corpora contain non-spontaneous language from speeches, books etc., this is still data produced with a purpose in mind that is not purely linguistic in contrast to surveys or elicitation studies. While the degree of naturalness will at the same time vary depending on how the data was collected, it still means that, in contrast to surveys the speakers are not actively thinking about their language use and making conscious decisions about it while producing it.

Particularly interesting in my findings was the limited use of the 1st person singular nominative and the relatively high use of the 3rd person singular feminine and masculine singular nominatives. This is unexpected because nominative *I* tends to be more common than the other pronouns in constructions where oblique forms are becoming more and more common such as postcopular *It is I*, as well as in cases of hypercorrection such as the well-documented and well-researched case *between you and I* where two coordinated DPs, at least one of them nominative, follow a preposition which otherwise assigns its complement oblique case (see for example Wales 1996, 105–7; Boyland 2001, 384–96). Nevertheless, *I* is barely found in my data.

A potential reason for the lack of *I* in the data could be the fact that while an advantage of corpus data is the naturalness of the language, this naturalness may also mean the accidental over- or underrepresentation of certain constructions and may for example cause accidental low numbers for certain constructions or elements. Accordingly, it is impossible to know whether something is not there because the grammar of the speakers does not allow it, or simply because they did not happen to use it in the texts included. While the numbers found in the BNC can provide a good indication of the use of the comparison particles as either prepositions or complementisers, there is no saying that the results correspond completely to the language used in real life. One way of mitigating this would be to make use of more (comparable) corpora.

The low number of nominative *I* is also likely to be affected by genre. As can be seen in Table 4, Category 1 Formal and academic writing contains the highest percentage of nominatives at 21.7% which was also expected since prescriptive norms tend to have a stronger effect on formal contexts. In formal (and particularly in academic) literature, it is often recommended to avoid using the 1st person singular pronouns (e.g. in guides like Hale and Basides 2023, 362). Research supports that this norm is mostly followed. In a corpus study of English abstracts and conclusions, Wang, Tseng, and Johanson (2021, 8) found that while the use of the 2nd person plural was relatively frequent, other pronouns, including 1st person singular were rarely used. Similar results were found in a corpus study by Kuo (1999, 124), who reports that *we, us*, and *our* are used to a much higher extent than the rest of the pronouns in academic journal articles.1st person singular pronouns are not used at all. In a study of the development of formality of academic texts between 1965 and 2015, Hyland and Jiang (2017, 45) found that academic writing is only becoming more informal by small margins and that the development happening is largely due to the rise of use of the 1st person pronouns. As a result, it is possible that newer data would include a greater proportion of 1st person singular nominatives, if the low number of 1st person singular is indeed caused by its lack of use in formal and academic texts.

By contrast, neither Wang, Tseng, and Johanson (2021, 8) nor Kuo (1999, 124) found anything to suggest that the 3rd person singular feminine and masculine pronouns were particularly common. On the contrary, in both cases the 3rd person plural was actually used more in the corpus data and in the journal articles that these studies looked at. However, in my data the highest relative amount of 3rd person singular pronouns was found within Category 2: Fictional and literary texts, where they made up 49.6% of the pronouns. By contrast, the number was 34.8% for Category 1, 29.3% for Category

3, and 30.0% for Category 4. Notably, Category 2, at 19.8% nominatives had almost as high a percentage of nominatives as Category 1. This is an interesting category, because language in fictional literature will vary according to the author, tone, style, narrator etc. As a result, it is difficult to determine whether the choice between nominative and oblique case reflects grammatical preferences, stylistic variation, or other literary choices.

Given these difficulties, a suggestion for a next step, apart from comparing with data from other corpora, would be to conduct a survey study which would provide conscious data to compare with the naturally produced data of the corpora, and it would be helpful to ask language users specifically for the reasons behind the choice of pronoun case, hopefully giving us a way to understand what motivates their grammatical decisions.

Conclusion

In this paper, I have conducted a preliminary corpus study on the comparison particles *as* and *than* in the BNC to investigate their use as either prepositions or complementisers. A clear overall preference for oblique pronouns was found, pointing to a predominating prepositional use. Interestingly, *than* seemed more likely to be used prepositionally than *as*. Furthermore, variation with respect to genre and to the individual pronouns was found. The lowest number of nominatives was found for the 1st person singular, and it was suggested that this is caused by genre differences. This suggestion is based on the observation that the most formal category, which altogether had the highest proportion of nominatives, is a category otherwise characterised by its lack of 1st person singular pronouns. Conversely, the surprisingly high proportion of nominatives for the 3rd person singular feminine and masculine pronouns is difficult to explain. Because the highest amount of these pronouns was found in the category with literary texts, it may also be affected by formality, but finding out would require an in-depth look at each individual text that each occurrence appeared in. One of the drawbacks of corpus linguistics is that corpora cannot tell us what goes on behind specific linguistic choices. It is clear from this brief look at the comparison particles *than* and *as* that there is much more to examine, such as whether conscious awareness of prescriptive norms has an effect and whether speakers are more likely to choose a nominative pronoun with *as* than with *than*. It would be a logical next step to look at the same constructions in different corpora and compare the results to be able to report on the issue with a higher level of certainty, as well as extending the research by including other methods such as an elicitation or acceptability study.

REFERENCE LIST

Aston, Guy, and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press. https://doi.org/10.1515/9780748628889.

BNC. 1994. British National Corpus. 1994. https://www.english-corpora.org/bnc/.

Boyland, Joyce Tang. 2001. 'Hypercorrect Pronoun Case in English? Cognitive Processes That Account for Pronoun Usage'. In *Frequency and the Emergence of Linguistic Structure*, edited by Joan Bybee and Paul J. Hopper, 383–404. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.45.

Britannica. 2024. 'Ask the Editor: "Better than I" or "Better than Me"?' The Britannica Dictionary. 2024. https://www.britannica.com/dictionary/eb/qa/better-than-I-or-better-than-me. Accessed May 3, 2025.

Brittain, Lewis. 1788. *Rudiments of English Grammar*. L.J. Urban.

Culpeper, Jonathan, and Mathew Gillings. 2018. 'Politeness Variation in England: A North-South Divide?' In *Corpus Approaches to Contemporary British Speech*, 33–59. Routledge. https://doi.org/10.4324/9781315268323.

Dauenhauer, Frank. 2021. 'How and When Does One Use "than I" and "than Me" in a Sentence?' Accessed May 3, 2025.

Davies, Mark. 2011. 'Synchronic and Diachronic Uses of Corpora: Interview with Mark Davies'. In *Perspectives on Corpus Linguistics*, edited by Vander Viana, Sonia Zyngier, and Geoff Barnbrook, 63–80. Studies in Corpus Linguistics 48. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.48.

———. 2013. 'GloWbE'. 2013. https://www.english-corpora.org/glowbe/.

———. 2015. 'Introducing the 1.9 Billion Word Global Web-Based English Corpus (GloWbE)'. *The 21st Century Text* (blog). 30 June 2015. https://21centurytext.wordpress.com/introducing-the-1-9-billion-word-global-web-based-english-corpus-glowbe/. Accessed March 28, 2025.

Doyle, Gabe. 2012. '"Than I" and "than Me"'. *Motivated Grammar* (blog). 13 June 2012. https://motivatedgrammar.wordpress.com/2012/06/13/than-i-and-than-me/. Accessed May 3, 2025.

Ehlers, Katrine Rosendal. 2020. '*Sin* og *sig* med Flertalsantecedent fra Runesten til LANCHART'. In *Danske studier. 2020*, 48–84. Universitets-Jubilæets danske Samfund, nr. 602. København: Syddansk Universiteitsforlag.

Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.3

Gries, Stefan Th. 2011. 'Methodological and Interdisciplinary Stance in Corpus Linguistics: Interview with Stefan Th. Gries'. In *Perspectives on Corpus Linguistics*, edited by Vander Viana, Sonia Zyngier, and Geoff Barnbrook, 81–98. Studies in Corpus Linguistics 48. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.48.

Hale, Adrian, and Helen Basides. 2023. *Keys to Academic English*. Cambridge University Press. https://doi.org/10.1017/9781009289009.

Huddleston, Rodney, and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, Marianne. 1998. *New Zealand English Grammar – Fact or Fiction?: A Corpus-Based Study in Morphosyntactic Variation*. Amsterdam: John Benjamins. https://doi.org/10.1075/veaw.g23.

Hyland, Ken, and Feng (Kevin) Jiang. 2017. 'Is Academic Writing Becoming More Informal?' *English for Specific Purposes* 45 (January):40–51. https://doi.org/10.1016/j.esp.2016.09.001.

Jäger, Agnes. 2019. 'The Syntax of Comparison Constructions in Diachronic and Dialectal Perspective'. *Glossa: A Journal of General Linguistics* 4 (1): 1–51. https://doi.org/10.5334/gjgl.651.

Jespersen, Otto, and Niels Haislund. (1949) 1974. *A Modern English Grammar on Historical Principles, Part VII: Syntax*. Repr. Vol. VII. VII vols. London: Allen & Unwin.

Jones, Christian, and Daniel Waller. 2015. *Corpus Linguistics for Grammar: A Guide for Research*. London: Routledge. https://doi.org/10.4324/9781315713779.

Kuo, Chih-Hua. 1999. 'The Use of Personal Pronouns: Role Relationships in Scientific Journal Articles'. *English for Specific Purposes* 18 (2): 121–38. https://doi.org/10.1016/S0889-4906(97)00058-6.

Leech, Geoffrey. 1992. 'Corpora and Theories of Linguistic Performance'. In *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*, edited by Jan Svartvik, 105–22. Trends in Linguistics. Studies and Monographs [TiLSM] 65. De Gruyter Mouton. https://doi.org/10.1515/9783110867275.

———. 2007. 'New Resources, or Just Better Old Ones? The Holy Grail of Representativeness'. In *Corpus Linguistics and the Web*, edited by Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, 133–49. Boston: Brill. https://doi.org/10.1163/9789401203791.

Love, Robbie. 2020. *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. New York: Routledge. https://doi.org/10.4324/9780429429811.

Lowth, Robert. 1763. *A Short Introduction to English Grammar*. 2nd ed., Corrected. London: A. Millar, and R. and J. Dodsley.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Meyer, Charles F. 2023. *English Corpus Linguistics: An Introduction*. 2nd ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781107298026.

Mukherjee, Joybrato. 2004. ‘The State of the Art in Corpus Linguistics: Three Book-Length Perspectives’. *English Language & Linguistics* 8 (1): 103–19. https://doi.org/10.1017/S1360674304001261.

Perlman, Merrill. 2019. ‘More than Me, or More than I? The Big Debate’. Columbia Journalism Review. 25 February 2019. https://www.cjr.org/language_corner/more-than-me-more-than-i.php. Accessed May 3, 2025.

Quinn, Heidi. 2005. *Distribution of Pronoun Case Forms in English*. Amsterdam: John Benjamins. https://doi.org/10.1075/la.82.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik, eds. 1985. *A Comprehensive Grammar of the English Language*. New York: Longman.

Taylor, Charlotte. 2016. *Mock Politeness in English and Italian: A Corpus-Assisted Metalanguage Analysis*. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.267.

University of Oxford. n.d. ‘BNC User Manual - Design of the Corpus’. Text. natcorp.ox.ac.uk. http://www.natcorp.ox.ac.uk/archive/worldURG/design.xml. Accessed 16 December 2024.

Visser, Frederik Theodor. 1963. *An Historical Syntax of the English Language*. Vol. 1. 3 vols. Leiden: Brill.

Voelkel, Svenja, and Franziska Kretzschmar. 2021. *Introducing Linguistic Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316884485.

Wales, Katie. 1996. *Personal Pronouns in Present-Day English*. Studies in English Language. Cambridge: Cambridge University Press.

Wang, Shih-ping, Wen-Ta Tseng, and Robert Johanson. 2021. ‘To *We* or Not to *We*: Corpus-Based Research on First-Person Pronoun Use in Abstracts and Conclusions’. *SAGE Open* 11 (2): 1-18. https://doi.org/10.1177/21582440211008893.

Willis, David. 2017. ‘Investigating Geospatial Models of the Diffusion of Morphosyntactic Innovations: The Welsh Strong Second-Person Singular Pronoun *chdi*’. *Journal of Linguistic Geography* 5 (1): 41–66. https://doi.org/10.1017/jlg.2017.1.