Testing hypotheses and answering research questions

4.1 Introduction

In Chapter 2 we showed how it is possible to use pre-existing large corpora to support the qualitative stylistic analysis of text fragments. When such an endeavour is carried out in pursuit of insights into the significance of particular stylistic choices in the text, this practice constitutes corpus-informed stylistics. In this chapter we broaden the parameters of corpus stylistics by exploring the possibilities of subjecting the target text itself (i.e. the text we are particularly interested in analysing) to corpus analysis. That is, rather than comparing a small section of our target text with a pre-existing corpus (i.e. a reference corpus), we instead treat the target text (or, indeed, texts) as a corpus in its own right. This allows us to use corpus analytical techniques to discover patterns in the text(s) that would not be discernible through qualitative analysis alone. To this end, in this chapter we begin to utilise the insights from Chapter 3 concerning sampling and representativeness, as we consider how to analyse not just single extracts but also collections of texts.

We refer to this practice of treating the target text (or collections of texts) as a corpus in its own right as corpus-based stylistics. When analysing a corpus stylistically, although it is possible to approach the task with no particular agenda concerning what we want to investigate, we advocate concentrating any analysis on answering a specific research question or addressing a specific hypothesis. This chapter, then, builds on the corpus-informed approach described in Chapter 2 by discussing the philosophy behind corpus-based stylistic analysis. We explain
methods for generating hypotheses and candidate research questions before moving on to consider the practice of developing a robust and replicable analytical method to test/answer these. Here we focus particularly on methodological issues shared by both corpus linguistics and stylistics, namely objectivity, rigour, replicability and falsifiability. We discuss how stylistics has dealt with these issues and the extent to which corpus linguistic methods increase the potential for addressing these. We also make clear that these issues apply equally to qualitative analysis and that here stylistics has a contribution to make to corpus linguistics. We illustrate these points with reference to our third case study, a corpus-based analysis of Ernest Hemingway’s authorial style. In so doing, we introduce two more software packages for corpus analysis, WordSmith Tools (Scott 2016) and Wmatrix (Rayson 2009). WordSmith provides valuable descriptive statistics for uploaded corpora, while the particular value of Wmatrix is its capacity for automatically annotating texts for part-of-speech (POS) and semantic information. In this chapter we make particular use of Wmatrix’s POS tagger; and in so doing we lay the groundwork for the more complex Wmatrix analyses that follow in subsequent chapters. We begin, though, with a brief discussion of the nature of a corpus. In corpus stylistics, some licence is often taken with how a corpus is defined, and it is worth being clear on both the pros and cons of doing this. We then clarify our use of the term corpus-based through consideration of the distinction between this and corpus-driven linguistics.

4.2 Corpora in stylistics

In Chapter 1 we discussed the definition of a corpus, noting that representativeness is a key criterion. Corpus linguists tend to be primarily concerned with making generalisable claims about language use based on the analysis of a sample of language data. For generalisability to be possible, such samples have to be representative of whatever language variety is under scrutiny. It is for this reason that Sinclair (2005) claims that an individual text cannot constitute a corpus. However, in stylistics, it is not always the case that we want to make generalisable claims of the kind made in corpus linguistics. Of course, corpus stylisticians do sometimes construct corpora that are intended to be representative of a particular stylistic variety (see, for example, Green’s 2017 Corpus of the Canon of Western Literature, as well as the corpora analysed in Chapters 8 and 9). And in the related field of stylometry, representativeness is just as important a criterion as it is in corpus linguistics (see, for instance, Hoover 2017). But it is also true that what many
Stylisticians are interested in what Short (1996a) calls text style; that is, the style observable in a single specific text. This may or may not be connected to authorial style. For example, Stubbs’s (2005) corpus stylistic study of Joseph Conrad’s *Heart of Darkness* is concerned with providing textual justification for well-established interpretations of the novella, and offering new insights into the language and meaning potential of the text. In this respect, Stubbs’s focus is on explicating issues concerning this text in particular, such as how the story is told and how the characters are projected. He does not seek to make claims about Conrad’s authorial style in general terms. Because of his specific interest in *Heart of Darkness*, there is no need for Stubbs to construct a corpus that is a representative sample of Conrad’s output as a novelist.

Despite being focused on one text in particular, Stubbs’s (2005) analysis does make use of other texts in the form of a reference corpus. In this respect, Stubbs’s (2005) study is an example of what Adolphs (2006) calls inter-textual analysis. But some studies of text style do not even utilise separate reference corpora. Adolphs (2006) calls such cases intra-textual analysis, where the focus is entirely on extrapolating particular information from one text in order to generate insights relating to that text alone. An example of a corpus stylistic analysis of intra-textual variation is Culpeper’s (2009) study of keywords (as well as key parts-of-speech and key semantic domains) in Shakespeare’s *Romeo and Juliet*. We introduce the notion of keyness in full in Chapter 5, but in brief a keyword is one that is more frequent in one text than it is in another. Keyness studies are often inter-textual in nature, in that a target corpus is compared against a reference corpus. Culpeper’s (2009) study, however, compares the dialogue of individual characters against the dialogue of all the other characters in *Romeo and Juliet*. In this respect, the play serves as both the target text and the reference corpus. Among Culpeper’s (2009) findings are the fact that *if, or, would* and *yet* are all keywords for Juliet; that is, they occur more frequently in Juliet’s speech than they do in the speech of the other characters. Culpeper (2009) notes that the findings from his keyness analysis are relevant to understanding characterisation in the play. For example, Juliet’s keywords reflect her tendency as a character towards uncertainty and anxiety.

Both Culpeper’s (2009) and Stubbs’s (2005) studies illustrate the value of using corpus analytical techniques on single texts. So we might ask: what precisely is the nature of Sinclair’s (2005) objection to the use of the term *corpus* to describe single texts? This is partly that its dimensional restrictions do not allow for generalisations to be made.
about the language variety that it represents. But in addition, it is also possible to be knowledgeable about a single text in a way that it is impossible, given its size, to be knowledgeable about the contents of a corpus. This has potential analytical implications. For example, it may well be that a corpus-based stylistic analysis of a single text is also informed (consciously or unconsciously) by the researcher’s qualitative familiarity with it. To the extent that this might impact on the objectivity of the analysis, Sinclair (2005) is right to draw attention to the distinction between corpora and single texts. Nonetheless, for ease of reference we do on occasion use the term corpus when discussing single texts; but in such instances we make quite clear that our focus is indeed on a single text and not a corpus in the traditional sense. In such cases, the potential complicating factors associated with the corpus stylistic analysis of single texts should be borne in mind. In theoretical terms, we agree with Sinclair in cautioning against extending the term corpus to encompass single texts too.

4.3 The concept of corpus-based and corpus-driven linguistics

The first section of this chapter introduced the concept of corpus-based stylistics and defined it as the practice of analysing a target text using corpus methods (as opposed to using a pre-existing corpus to support or challenge a qualitative analysis of a text extract). Before we proceed any further with this concept, however, it is worth briefly discussing the origins of the phrase corpus-based, since our usage here differs somewhat from that to be found in mainstream corpus linguistics.

Tognini-Bonelli (2001) contrasts corpus-based with corpus-driven linguistics. The distinction between these two analytical procedures (as described by Tognini-Bonelli 2001; see also Francis 1993) is that the former uses corpus analysis to validate existing theories and hypotheses, while the latter uses the results of corpus analysis to formulate new, corpus-derived theories and hypotheses. For example, the 100 million word BNC contains part-of-speech information that was added to the corpus automatically using a POS tagger. But the POS information that the tagger contains is based on a non-corpus-derived grammar. The upshot of this is that what we learn about the grammar of English by studying the BNC will always be constrained by the model of grammar that the POS tagger uses. Retrieving information about prepositions, for instance, will only ever allow us to observe those lexical items that have been identified as prepositions by the tagger. We will not be able to observe any instances of potential prepositional meaning in words that have not already been tagged as having this function. This, then, is
corpus-based linguistics, that is, using non-corpus-derived theories and analytical frameworks to inform a corpus analysis.

By contrast, corpus-driven linguistics is not constrained by existing theories or analytical frameworks. Rather, the aim is to identify patterns in the raw (i.e. unannotated) corpus data that might then be used to formulate new theories of whatever aspect of language we are studying. Sinclair’s work on lexis and meaning (1991; see also 2004) is a prime example of corpus-driven linguistics, and it is worth briefly explaining this so as to be clear about the distinguishing characteristics of a corpus-driven approach to data analysis.

Sinclair (2004) explains that a particular problem in linguistics – one that arises as a result of the excessive attention paid to written (as opposed to spoken) language – is that we have a tendency to see words as discrete units. This, obviously, is because words on a page are separated by spaces. Words, of course, are formed from morphemes but whether we see words or morphemes as the basic unit of language, the important point is that, traditionally, linguists have focused on how these basic units combine, that is, their interest has been in grammar. Meaning, by contrast, was (and still is in some quarters) assumed to be a separate property of words, distinct from this focus on structure. This view, however, ignores the significant role that structure plays in generating meaning. As Sinclair points out, ‘[w]ords enter into meaningful relations with other words around them, and yet all our current descriptions marginalize this massive contribution to meaning’ (2004: 25). For example, dictionaries often treat phrasal verbs (e.g. give in) as anomalies to be dealt with separately from single unit lexical items. Accounting for the meanings of phrasal verbs is difficult using the standard hierarchical model of language which sees morphemes combine to form words, words combine to form phrases, phrases combine to form sentences and so on. This is because phrasal verbs seem to cut across these category boundaries. And while the meanings of some phrasal verbs seem to be entirely different from the meanings of their component parts (e.g. brush up on), others do not (e.g. look for). To explain such meanings, we need to understand how words relate to each other semantically. Since we are used to examining constraints on syntactic structure, there is no reason in principle why we should not extend this consideration to constraints on lexical relations (Sinclair 2004: 28). For example, strong tea is an acceptable formulation whereas *powerful tea is not.\(^1\) The reason we can have strong but not powerful before tea is not down to syntactic constraints (after all, English allows adjectives to pre-modify nouns). It must, then, be due instead to lexical restrictions.
The notion of lexical constraints forms the core of collocational studies (see, for example, the case study in Chapter 2) and is at the heart of the concept of a unit of meaning. Sinclair’s (2004) point is that it is only through corpus-driven analysis that we can identify such structures. As an example, Sinclair (2004: 30–5) shows how the meaning of the expression *naked eye* can only be explained by determining the unit of meaning in which it is typically found. Through observation of concordance lines for *naked eye*, Sinclair explains that the collocational and colligational patterns, in conjunction with patterns of semantic preference, suggest the following unit of meaning: ‘visibility + preposition + the + naked + eye’ (2004: 33). This unit of meaning generates such usages as ‘too faint to be seen with the naked eye’, ‘is not really visible to the naked eye’ and ‘could be seen by the naked eye’. In addition, Sinclair (2004: 33) notes an associated semantic prosody of difficulty. The point is that the meaning of *naked eye* arises from the sum total of the elements of the unit of meaning; it is not inherent in the phrase itself. And the constraints on lexical relations predict that a usage such as ‘delicious to the naked eye’ would be unlikely and unacceptable (except in deliberately creative formulations).

On the basis of such corpus-driven analysis, Sinclair (2004) formulates what he terms the idiom principle. According to this principle, words have meanings which derive from their collocational patterns. That is, meaning is not necessarily inherent in individual words but is dependent on the particular patterns that words form. Consequently, constructing sentences is not a matter of inserting individual words into open grammatical slots (as suggested by the slot/filler model of structuralist grammar), or of generating sentences according to an underlying set of grammatical rules (as proposed by generative grammar). According to the idiom principle, speakers do not have a free choice in which words they use (e.g. speakers are constrained to say *strong tea* as opposed to *powerful tea*). Instead, they select from a set of pre-existing patterns. In essence, we form utterances by relying primarily on the idiom principle. Only when this principle fails to offer an option do we fall back on what Sinclair (2004) calls the open-choice principle.

Sinclair’s corpus-driven work on lexis has been hugely influential, giving rise to, for instance, lexical priming theory (Hoey 2004) and pattern grammar (Hunston and Francis 1999; Hunston 2015), the grammar that underpins the Collins COBUILD series of dictionaries and grammar books (see Chapter 1). Hunston and Francis (1999: 22) further point to the role of the idiom principle in the interpretation of utterances. They explain that an example like *I must confess* can be interpreted either as a single item, under the idiom principle (in which
case it means something like, ‘I am about to tell you something embar-
rassing to me’), or as a string of individual items, according to the
open-choice principle (in which case it means ‘I am obliged to inform
you of some wrongdoing on my part’). It is down to the hearer to
decide which principle generates the most appropriate interpretation.
To take a similar example, consider the expression I have to say. Under
the open-choice principle, this would be interpreted as meaning ‘I am
required to inform you.’ But under the idiom principle, the meaning is
something like ‘I wish to emphasise the following statement.’ This can
be seen in the concordance lines from the BNC, shown in Figure 4.1.

The difference between the two is that the meaning generated by the
idiom principle interpretation belongs to the sequence as a whole, which
means that its component words cannot be altered without affecting
this meaning. For instance, changing the pronoun and tense (e.g. he had
to say) would preclude the idiom principle interpretation.

The point behind this explanation of Sinclair’s work and the work
of Hunston and Francis (1999) is that the insights into grammar that
they offer could only have been gained through corpus-driven analysis.
A corpus-based analysis of an automatically POS-tagged corpus would
very likely have failed to reveal these patterns because of the influencing
effects of the non-corpus-derived grammar used by the tagger. This,
then, is the distinction between corpus-based and corpus-driven lin-
guistics. Nonetheless, while this distinction is useful in heuristic terms,
it should be borne in mind that in practice there are often overlaps
between the two approaches. Hardie and McEnery (2010) see the
distinction as the difference between viewing corpus linguistics as a
methodology (the corpus-based position) and seeing it as a distinct
sub-discipline of linguistics (the corpus-driven position). Hardie and
McEnery (2010) refer to the first view as the methodologist position and
the second as the neo-Firthian position (based as it is on the theoretical
work of J. R. Firth; see Chapter 2). They further point out that the two
traditions have an equally long history and suggest that the distinction
between them seems likely to diminish over time as neo-Firthian dis-
coversies such as pattern grammar converge with cognitive linguistic
theories. Construction grammar, for instance, is a cognitive linguistic
theory of language structure that shares many similarities with pattern
grammar, as suggested by Hank’s definition of a construction as ‘a lin-
guistic element that cannot be broken down into smaller units without
loss of meaning’ (2008: 226; for more on construction grammar, see
The difference is that pattern grammar is based solely on observable
instances of language in use, whereas construction grammar is focused
Okay. Yes, that will. And I have to say my experience it leaves me to doubt, I know I'm a cynic sometimes place over the (unclear) (pause) erm (pause) we have immediately re-secured, city football club, because they are concerned, and have been yesterday, to which the answer, I mean I'm afraid organized is different. It, again I have to, being in this (unclear) together. Yes, Yes, see if we can erm, if that, because, because important and I think it's, there's er actually, ith the brick (unclear) and the flints. are (pause) two of the most important for people. Well it, I really enjoy that. Meetings (pause) once again the meetings (pause) please. Erm (pause) and that's about all training staff, your, your residential staff get trained. But there is an over-supply of residential accommodation, not evenly distributed, military term. To fill that gap for us immediately. And I have to say my experience it leaves me to doubt, I know I'm a cynic sometimes there's been about fifty erm (pause) tickets out to re-secure those (pause) those, relatively er in an erm unbiased as unbiased as they could be. Er for many years, so it's not unusual. (unclear) the th-- the glib answer is, why are you telling me you need a that although you know part of reading (-----)'s letter, erm, I er I detect (-----) in (-----)'s letter because I am quite sure that there's a little bit grieving in the way (pause) Oh, yeah, that as a result of last night's session I have a much healthier view (pause) the old, old (pause) chestnut (pause) please can we start at eight o Mr Chairman. (unclear) need a proposer (unclear)2. for Shropshire, it means one I think, er, per year, and that, that it varies, for example in Shrewsbury there's a very sizable to you again that the home care service is the one that is most under

Figure 4.1 Concordance lines from the BNC illustrating the idiom principle interpretation of *I have to say*
on modelling mental representations of language. Nonetheless, the clear similarities between them would suggest a converging of corpus and cognitive theories of language (for examples of the connections between pattern grammar and construction grammar, see Hennemann 2015; Hunston and Su 2017).

As Hardie and McEnery (2010) point out, then, in practice there are overlaps between corpus-based and corpus-driven approaches in corpus linguistics generally. The same is true in corpus stylistics. For this reason, it is worth clarifying our own use of the terms. We use *corpus-based stylistics* to refer to any corpus stylistic study that treats the target text as a corpus, whether annotated or not (cf. corpus-informed stylistics, which uses existing large corpora to generate insights into single short texts). We reserve the term *corpus-driven stylistics* for those studies that are exploratory in nature (e.g. those that are carried out for purposes of selecting a sample for qualitative analysis, or to generate potential research questions).

Let us now turn to another concept that has been widely discussed in both corpus linguistics and stylistics: intuition.

### 4.4 The value of intuition

The notion of hypothesis testing is a mainstay of stylistic analysis, though it is often not discussed in such explicit terms. In Leech and Short’s (1981) seminal *Style in Fiction*, for instance, there is no reference at all to the concept of a hypothesis. What there is, however, is considerable discussion of the notion of intuition, and of the value of using stylistic analysis to test the validity of our initial impressions of a text. In effect, this is what we did in Chapter 2 when we examined concordance lines of the phrase *with intent* in order to gain an insight into the potential effects of the phrase ‘strolling with intent’ in the John Le Carré passage. We began with an intuition that this was a stylistically interesting phrase and we then examined concordance lines and collocates to determine the effects likely to spring from its use. Testing our intuitions is very similar to the concept of testing a hypothesis. The difference is that hypotheses are formulated more stringently than intuitions, which tend by their nature to be more general. Here, for example, are Leech and Short’s initial impressions of a passage from Joseph Conrad’s *The Secret Sharer* (1910):

> Our first impression of this passage is of a meticulously detailed setting of the scene for the story. The description is clearly etched, so that we can reconstruct, in our mind’s eye, the whole
topography. But more than this, we have a vivid sense of the loneliness of the human observer, set apart from his surroundings, and of ‘a mind energetically stretching to subdue a dazzling experience outside the self, in a way that has innumerable counterparts elsewhere in Conrad’ [Ohmann 1966: 152]. (Leech and Short 1981: 83–4)

The impressions described in the above quotation include specific feelings that, according to Leech and Short, the text imbues in the reader. In addition, there is reference to the claims of another critic (Ohmann 1966) about Conrad’s style. Leech and Short (1981: 84–90) then go on to explain the linguistic source of these impressions, pointing out, for example, the distancing effect caused by the syntactic subordination in the passage, and how this might relate to their initial impression of loneliness. Leech and Short’s (1981) linguistic analysis is detailed and replicable. What is harder to evidence is a link between the stylistic features they observe and their effects on other readers. One reason for this is that the claim for an effect of loneliness is unfalsifiable. That is, if another reader claims not to encounter this feeling as they read, then it is impossible for us to say that they are wrong. The most that we can do in such a circumstance is to claim that stylistic effects in the text give rise to an impression of loneliness for us, while accepting that for other readers this emotional response may be different. Research in cognitive stylistics has made considerable headway in accounting for such differences of interpretation, not least through the use of schema theory to explain the source of literary impressions (see, for example, Jeffries 2001; Semino 2001; Stockwell 2002: 82–7; Lang 2009). Moreover, Short et al. (2011) have demonstrated that what are described as different interpretations would in most cases be better described as variations of the same higher-order interpretation.

The other issue with testing the kinds of intuitions observable in the Leech and Short (1981) quotation is that such impressions rarely provide much guidance on what might constitute their linguistic source. In this respect, looking for linguistic evidence for a particular impression of a text can be likened to searching for a needle in a haystack, or trying to locate a light switch in a darkened room. For these reasons, instead of using our general impressions of a text to underpin a stylistic analysis, it is often of greater value (not to mention more manageable analytically) to use them to formulate specific hypotheses which are then amenable to more reliable testing. We should not dismiss the value of intuition in stylistics; we should simply be mindful of its limitations.
The same is true of intuition in corpus linguistics. Again, intuition has value. As McEnery and Hardie note, ‘without some ability to introspect, it is doubtful whether a linguist could ever formulate a question to ask of a corpus’ (2012: 26). This is a point also made by Aston (2011: 5; see also McEnery and Wilson 2001: 7). But of course, our intuition is limited, particularly concerning certain aspects of language use. As Hunston (2002: 20–1) clearly explains, we are notoriously unreliable when it comes to making intuitive judgements about the frequency of words, phrases and other linguistic structures. We are similarly unreliable at making judgements about collocations, semantic prosody and phraseology. But testing our intuitions is one of the prime purposes of doing corpus analysis. Again, this means using our intuitions to generate hypotheses and research questions. And as Aston points out, ‘because intuition is unreliable, it is a good idea also to pose questions to which you think you already know the answers’ (2011: 6). This is a particularly pertinent point, given one of the criticisms that is often levelled at stylistics by non-linguistically inclined literary critics; namely: what is the value of doing detailed linguistic-stylistic analysis if the results tell us nothing that a sensitive literary critic could not have told us without going to the trouble of doing such work? The answer, as Aston (2011: 6) notes, is that while we might intuitively feel that we know the effects of a particular linguistic structure, without testing this empirically we can never know for certain. This attitude has long underpinned stylistics, and is perhaps summed up most famously by Spitzer, who notes that ‘to make our way to an old truth is not only to enrich our own understanding: it produces new evidence of objective value for this truth – which is thereby renewed’ (1948: 38). Introspection, then, is important to corpus stylistics. But crucially its value is as a source for generating hypotheses and research questions, not as a source of data in its own right (cf. the role of intuition in early generative syntax, such as Chomsky 1957).

4.5 Hypotheses

Hypotheses might be formulated based on our intuitions, or they might stem from the literature on whatever topic we happen to be interested in. The key point is that a hypothesis aims to explain a particular phenomenon. Moreover, a hypothesis must be testable; that is, it must be formulated in such a way as to make it possible, after appropriate analysis, to reject it as an explanation of the phenomenon under investigation. For example, we might want to claim that the reason Charles Dickens is so lauded as a novelist is because he had a larger
vocabulary than those of his contemporaries. This is an explanation for a particular phenomenon (that Dickens’s work is distinctive) and one that is also open to testing, at least in theory. To decide whether our hypothesis is correct or not, what we would need to do is compare the size of Dickens’s vocabulary with the size of the vocabularies of other novelists of the time, and see whether it is indeed the case that Dickens used a greater variety of words. A key point, though, is that we can never prove a hypothesis to be right. While this might sound counter-intuitive, the reason is that we can never know whether an additional test might prove the hypothesis to be wrong. For this reason, the scientific method focuses on proving hypotheses to be wrong, not right (this approach to hypothesis testing is a consequence of the seminal work of Popper [1959] 2002; see Chapman 2008: 28–48 for an outline of Popperian falsification in relation to linguistics, and Hughes and Sharrock 1997: 77–81 for a summary in relation to the sociology of science more generally). For example, if on comparing Dickens’s vocabulary with those of his contemporaries we find that he does indeed use a greater variety of words, we still cannot say for sure that Dickens has a larger vocabulary. This is because we cannot know for certain whether another novelist, one whose work we have not examined, uses an even wider variety of words. However, if we compare Dickens’s vocabulary with the vocabularies of his contemporaries, and we find that one of those contemporaries has a larger vocabulary than Dickens, then we have proved beyond doubt that our initial hypothesis was wrong (notwithstanding some of the problems associated with null hypothesis significance testing, which we discuss further in Chapter 5). We only need one instance of a nineteenth-century novelist having a larger vocabulary than Dickens to be sure of this. Such a result would then allow us to reject our initial hypothesis as a potential explanation for why Dickens’s work is distinctive.

Much corpus linguistic work relies on statistical hypothesis testing (i.e. using statistical tests to determine whether there is indeed a difference between the phenomena under investigation; see Chapter 5 for a fuller discussion of statistics in corpus studies). In such cases, it is appropriate to state both a null hypothesis and an alternative hypothesis. The null hypothesis (H₀) always posits that there is no difference between the phenomena under investigation. In the case of our Dickens example, the null hypothesis would be that there is no difference between the size of Dickens’s vocabulary and those of his contemporaries. The alternative hypothesis (Hₐ) is that there is a difference between the two phenomena. We could, in fact, formulate two alternative hypotheses:
H$_1$ Dickens’s vocabulary is larger than those of his contemporaries.
H$_2$ Dickens’s vocabulary is smaller than those of his contemporaries.

Statistical analysis can then be employed to determine whether it is possible to reject (i.e. prove wrong) the null hypothesis in favour of one of the alternative hypotheses. For example, if after analysis we find that one of Dickens’s contemporaries had a larger vocabulary than him, then we can reject the null hypothesis in favour of H$_2$ (until such time as other evidence becomes available). Alternatively, if we find that there is no significant difference between the vocabulary sizes of Dickens and his fellow novelists, then we have simply failed to reject H$_0$.

Whatever the outcome of hypothesis testing, it is important to bear in mind that observing a correlation between two phenomena does not mean that one phenomenon causes the other. For example, if after experiment we are able to reject the null hypothesis in favour of H$_1$ (i.e. Dickens’s vocabulary is larger than those of his contemporaries), this does not necessarily mean that vocabulary size is what causes Dickens to be so celebrated as a novelist. There may well be other variables that impact on this, which we would need to take account of.

It should be clear from the discussion above and in section 4.4 that hypothesis testing (or, in its weak form, intuition testing) is a mainstay of stylistic analysis. In the next section we explain the concept of research questions, how these should be used to direct a stylistic analysis, and the role they play in allowing us to reject hypotheses.

### 4.6 Objectives and research questions

The initial hypothesis stated in section 4.5 (strictly speaking, an alternative or experimental hypothesis) was that Dickens had a larger vocabulary than his contemporaries did, and that this is what makes his work so distinctive. Underpinning this hypothesis is a broader research question (what we will call an over-arching or global research question): what makes Dickens’s prose distinct from that of his contemporaries? This question, however, raises a number of other questions which we would first need to answer before we are in a position to (1) either reject or fail to reject our hypothesis, and (2) posit an answer to our over-arching research question. In order to determine what these sub-questions are, we can consider the objectives (O$_s$) we would need to achieve in order to answer our global research question, which might be as follows:

O$_1$ Determine what makes Dickens’s prose distinct from that of his contemporaries.
O2 Collect a sample of Dickens’s prose.
O3 Collect a sample of the prose of Dickens’s contemporaries.
O4 Count the number of words in the two samples.
O5 Determine the number of different words in the two samples.

What should be noticeable about these objectives is that they each have a different status within the project as a whole. O1 might be described as the project’s overall intellectual objective. Os 2 and 3, on the other hand, are methodological objectives, while Os 4 and 5 are analytical objectives. In effect, to achieve these objectives, we need to answer the following research questions (RQs):

RQ1 What makes Dickens’s prose distinct from that of his contemporaries? [Addresses O1]
RQ2 What counts as a representative sample of Dickens’s prose? [Addresses O2]
RQ3 What counts as a representative sample of the prose of Dickens’s contemporaries? [Addresses O3]
RQ4 How many words are in each sample? [Addresses O4]
RQ5 How many different words are in each sample? [Addresses O5]

It should also be clear that the answer to RQ1, our global research question, can only be posited after the other research questions have been answered. For this reason, we might refer to the other questions as local research questions. What should be apparent is that answering the local research questions enables us to achieve the objectives we need to meet in order to decide whether to reject or fail to reject our initial hypothesis. And whether we reject or fail to reject our hypothesis determines the answer to our global research question.

4.7 How big was Dickens’s vocabulary?

The discussion in the previous two sections, concerning the size of Dickens’s vocabulary relative to the vocabularies of his contemporaries, might not be particularly interesting in stylistic terms, but it is a relatively straightforward issue to investigate. We simply have to achieve the objectives described in section 4.6 in order to be able to answer the associated research questions. Since RQs 4 and 5 concern quantitative issues, we will posit both a null hypothesis and two alternative hypotheses. Recall that these are as follows:

\[ H_0 \text{ There is no difference in the vocabulary sizes of Dickens and his contemporaries.} \]
H₁ Dickens’s vocabulary is larger than those of his contemporaries.
H₂ Dickens’s vocabulary is smaller than those of his contemporaries.

RQ2 asks what counts as a representative sample of Dickens’s prose. We do not have access to everything Dickens wrote in his lifetime (this would include all his personal letters, notes, shopping lists, etc.) and so whatever we examine will be just a proportion of his total output. To achieve as complete a corpus as possible, however, we used the Dickens Complete Corpus (DCC; Alsuweed 2015). As the name suggests, this is a corpus containing the entirety of Dickens’s published work. Alongside Dickens’s novelistic output, it also includes personal letters, poems, plays, journalism and short stories. We used everything in the DCC except the non-prose outputs. As a reference corpus, we used the HUM (Huddersfield-Utrecht-Middelburg) Corpus which aims to be a representative sample of nineteenth-century fiction. So far, then, we have answered RQs 2 and 3 and achieved Os 2 and 3. The next stage is to count the words in the two corpora. However, simply counting the total number of words is not particularly informative, since (1) the two corpora are of different sizes so we are not comparing like with like, and (2) knowing the total number of words in each corpus is of less value than knowing the total number of different words. In effect, what we need to know about are word types rather than word tokens. A token is any instance of a word in a corpus. Hence, a one million word corpus contains one million tokens. A type, on the other hand, is a unique word. In a one million token corpus, many of those tokens will turn up more than once. There will, then, always be fewer types than tokens. To answer RQ4 (How many words are in each sample?) what we really need to know is the number of types in each corpus. The figures for the DCC and the HUM corpora are given in Table 4.1.

Of course, because the two corpora are different sizes we cannot rely on raw figures to determine which corpus has the highest number of types. What we can do instead is calculate a type-token ratio (TTR). To do this, we simply divide the number of types by the number of tokens and multiply by 100. Table 4.2 shows the results.

As can be seen in Table 4.2, the TTR is higher for the DCC corpus than it is for HUM. What this indicates is a greater degree of lexical

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCC</td>
<td>50,119</td>
<td>6,144,502</td>
</tr>
<tr>
<td>HUM</td>
<td>83,962</td>
<td>13,168,635</td>
</tr>
</tbody>
</table>
variation in DCC than in HUM. However, we have to be careful of reading too much into these results since the difference between the two TTRs is not great. This is a problem that arises when calculating TTRs for large corpora and is explained by the fact that in such corpora, grammatical words (e.g. articles, prepositions, etc.) are inevitably repeated, resulting in low TTRs. For this reason, a better measure of lexical variation is a standardised type-token ratio (STTR). This can be calculated using WordSmith Tools (Scott 2016), a concordancing package for corpus analysis (see Appendix 2). The STTR is calculated by measuring the TTR for the first 1,000 words in the corpus, then the TTR for the next 1,000 words, and so on. Finally, a mean average is calculated for the TTRs, which constitutes the standardised type-token ratio. Table 4.2 also shows the STTRs for both corpora, and what is apparent from these scores is that there is even less difference between them than there is between the TTRs.

On the basis of these results, we have failed to reject the null hypothesis. In effect, our results show that there is no difference between the vocabulary sizes of Dickens and his contemporaries. This, of course, is one measure of vocabulary size. Were we to investigate this further, we may want to look at, for example, average word length, word origins, and so on, before coming to a final conclusion about whether to reject H0. And, of course, there are myriad other factors we could consider, such as average sentence length and sentence complexity. Nonetheless, however we choose to investigate these issues, the key point is that we need to do this in an objective way.

### 4.8 Objectivity in stylistics

The investigation above into the size of Dickens’s vocabulary demonstrates a fundamental principle of stylistics. This is the concept of aiming to be objective in analysis. That is, rather than simply asserting that Dickens’s vocabulary was larger than those of his contemporaries, we tested the claim using a replicable method. In the natural sciences, the notion of trying to be objective analytically is uncontroversial. However, in some areas of the social sciences and particularly in the
humanities, the concept of objectivity has been the subject of considerable debate. Stylistics in particular has been criticised for its commitment to objectivity (see, for example, Bateson 1967, 1968; Fish 1980; Mackay 1996, 1999), and so it is worth briefly clarifying what it means to be objective.

The principle objections to the stylistic enterprise have come from literary critical quarters. In some cases (e.g. Bateson 1967, 1968; Mackay 1996, 1999) they have rested on the assumption that literature as a creative artefact is inaccessible to objective study (for an explanation of why this view is mistaken, see McIntyre and Price 2018a). In others (e.g. Fish 1980), they stem from a conviction that stylistics is not the objective discipline that its practitioners claim. However, what all of these complaints have in common is that they rest on a misunderstanding of the nature of objectivity. In essence, objectivity equates to clarity. If we want to answer a research question with any confidence, then we need to be clear about the methods we intend to use to go about answering it. We need to be open about the data we are using and we need to apply our analytical methods to the data systematically. We then need to report our findings honestly. The aim behind these processes is to produce results which are falsifiable; that is, results which are open to being proved wrong. This might sound a counter-intuitive thing to want to do, since surely we want to be right! We do, of course – but the only way to achieve that in the long term is to allow other analysts to question our results. If, after being as clear as we possibly can, other linguists fail to disprove our findings, then our results stand. If, on the other hand, other linguists do manage to falsify our results, then we are wrong – and to pretend otherwise would be delusional. The only alternative to this objective approach is to make subjective assertions and to argue through force of rhetoric. The problem with this is that subjective assertions are unfalsifiable. They are not open to being proved wrong so we can never know whether a subjective assertion is accurate or not. The alternative to the objective approach, then, is no alternative at all.

Despite the problems with making subjective assertions, these are still particularly common in non-linguistically oriented literary studies. Mackay (1996) even argues in favour of a subjective approach to stylistics, missing the fact that a subjective stylistics is, by its nature, paradoxical. His reasoning, however, is based on a misunderstanding of the nature of objectivity, which leads him to accuse stylisticians of not practising objectivity at all. For example, in response to Short’s claim that ‘[n]o analysis is objective in the sense that it is true for all time’ (1996a: 358), Mackay writes:
I read this as meaning that an analysis can be objective but not in the particular sense ‘that it is true for all time’. So in what sense are we to assume that an analysis is or can be objective? How long is its shelf-life? Five years? One year? Can such an analysis be true for a fortnight? (Mackay 1999: 61)

As Short and van Peer (1999: 272) point out, Mackay’s first sentence is accurate. It is indeed the case that no analysis is objective in the sense of ‘true for all time’. (Short and van Peer 1999 note that it is even possible to imagine circumstances in which analytically true statements – such as ‘all triangles have three sides’ – might no longer hold.) Mackay’s subsequent questions, however, are seriously misguided. The reason should be clear from our discussion of hypotheses in section 4.5. Recall that we cannot prove hypotheses right; we can only reject or fail to reject them. In the case of the analysis of Dickens’s vocabulary size, we failed to reject the null hypothesis. However, this does not mean that we proved $H_0$. It only means that it seems reasonable to accept it until such time as further evidence comes to light that disproves it (perhaps a larger, more representative reference corpus would give us different results). In this respect, Mackay’s questions in the quotation above are futile. There is no need to specify a ‘shelf-life’ for objective analyses; the findings from an analysis can be accepted until such time as new evidence comes to light that shows them to be false.

The other reason that Mackay (1996) opposes the objective approach is because stylisticians freely admit that achieving objectivity is impossible. Again, however, Mackay’s interpretation of this point reveals a faulty understanding of their position. Mackay (1996) points out what he claims to be inconsistencies in Carter’s (1982) endorsement of the objective approach. Briefly, Carter notes that we can only ever be ‘relatively objective’ and that there is ‘no such thing as an objective criticism’; nonetheless, he is clear that stylistics is ‘at least more objective’ than literary critical approaches to text analysis (1982: 6, 16, 82). Mackay takes issue with these assertions as follows:

From interpreting texts in a ‘relatively objective manner’, we move to the bold statement that there can be ‘no such thing as an objective criticism’ and then, finally, to the assurance that the ‘practical stylistic approach is at least more objective’ than its rivals. In the absence of ‘an objective criticism’, it is difficult to see how one could judge the relative objectivity of different approaches to interpreting literary texts. (Mackay 1996: 82)
What Mackay misunderstands, however, is that when Carter says there is ‘no such thing as an objective criticism’ (1982: 16), what he means is that no approach to criticism can be 100 per cent objective. There will always be variables that we cannot control for that might impact on our analysis (indeed, this is the reason for triangulation as a methodological tool; see Miller and Brewer 2003: 326; Baker and Egbert 2016). Crucially, though, this does not undermine the objective approach, despite Mackay’s implications to the contrary. This is because it is better to aim for objectivity and fail to fully achieve it than to reject it entirely in favour of subjective assertions which can never be proved wrong. The aim for objectivity, then, is an attempt to produce falsifiable claims via replicable analytical methods, in order to avoid making subjective assertions. In the final section of this chapter, we will explore how this works in practice in an analysis of elements of the authorial style of Ernest Hemingway.

4.9 Case study: Hemingway’s authorial style

Having described the basic tenets of corpus-based stylistics and the principles upon which it is founded, in this section we illustrate its practice through an analysis of a corpus of the novels of Ernest Hemingway. We begin with a brief discussion of how our hypotheses were formulated, before going on to spell out our research questions and associated analytical process. This involves a description of Wmatrix (Rayson 2009), the software we used to carry out our analysis. Finally, we discuss the value of corpus-based stylistics to the analysis of literature particularly, following Mahlberg’s (2013b: 11–12) suggestion that the corpus-based approach should be integrated into what Spitzer (1948) refers to as a ‘philological circle’ (see the discussion in Leech and Short 2007: 12).

Ernest Hemingway is widely acknowledged to be one of the foremost prose stylists of the twentieth century, having won the 1954 Nobel Prize in Literature ‘for the influence that he has exerted on contemporary style’ (Nobel Foundation 2018). Nonetheless, for all their discussion of Hemingway’s style, very few critics offer any concrete claims about the linguistic forms from which it is generated. All too often, any claims about Hemingway’s use of language descend into the vacuous, as in Cain’s assertion that: ‘[t]he structure of Hemingway’s sentences makes the reader keenly aware of the words that he has selected and, just as much or more, the countless other possibilities that he has not selected’ (2014: 80). The problem here is that this is an unfalsifiable claim. Cain offers no way of determining whether Hemingway’s sentence structure does indeed make readers ‘keenly aware’ of the constituent words of his
sentences. Nor is it clear what bearing this might have on Hemingway’s style. Cain continues in this vein when he says:

> The sentence that Hemingway writes must be this way: it is as it is. Yet the sentence need not be this way: it appears stable and sounds strong but it is fragile, precarious, and vulnerable to endless revision. Each sentence is closed yet open, perfectly complete in the midst of other possibilities. (Cain 2014: 82)

Again, this is nothing more than a series of vacuous, tautological and unfalsifiable claims. In short, these assertions are of no value to the corpus stylistician since there is nothing here that can be tested; there is no way we can attempt to prove Cain wrong on these points. And if there is no way of testing whether Cain’s assertions are wrong, there is no reason for us to accept that he is right. In essence, we cannot form any testable hypotheses from the claims that Cain makes. Unfortunately, this is all too often the case in literary criticism.

There does, however, appear to be something of a consensus among literary critics in their subjective assertions about Hemingway’s style. Often, for example, it is characterised as ‘repetitive’ (Ryan 1995: 232; Balaev 2014). Meyers, for instance, asserts that: ‘Hemingway’s style was characterised by clarity and force. He stressed the function of the individual word, wrote five simple sentences for every complex one, used very few similes, repeated words and phrases, emphasised dialogue rather than narration’ (Meyers 1985: 140). According to Meyers (1985: 74), Hemingway’s authorial style was an attempt to achieve the Imagist style associated with Ezra Pound. This, perhaps, is what Cain means when he claims that Hemingway sought ‘complexity through stringent simplification’ (2014: 83).

Similarly, Trodd writes that ‘[w]ithin his form, Hemingway embedded a further commentary upon language’s depleted capacity for expression. For example, his paratactic syntax – which juxtaposes clauses and like syntactic units without subordinating conjunctions – creates static, abrupt sentences that seem to stammer or bark’ (2007: 8). Trodd’s comments contain the same kind of tautological assertions as found in the quotations from Cain (2014), above. What does it mean, for example, to suggest that language has a ‘depleted capacity for expression’? However, Trodd does at least make reference to formal features of language, which opens up at least some of his claims to testing. If, for instance, Trodd is claiming that paratactic syntax is a distinctive feature of Hemingway’s style, and that subordinating conjunctions are noticeable by their absence, then we might hypothesise
that subordinating conjunctions are significantly under-represented in Hemingway’s writing.

While the mainstream literary establishment often seems loath to make falsifiable claims about Hemingway’s style, non-academic discussions of the author’s work often go to the opposite extreme, making falsifiable assertions but offering no evidence to support them. For instance, the Wikipedia site for Ernest Hemingway claims that Hemingway’s syntax ‘lacks subordinating conjunctions’,3 while the website Literary Devices goes even further, stating that: ‘Ernest Hemingway was famous for his short, declarative sentences. He rarely even used adjectives and almost never used adverbs.’4

What we can observe in the literary critical commentary on Hemingway is a tension between unfalsifiable subjective assertions and unevidenced objective claims. Despite its shortcomings, then, we might use such literary critical commentary as a point of departure for formulating some research questions and testable hypotheses. We can express these as follows:

Global research question

RQ1 Are literary critical claims about Hemingway’s style accurate?

*Answer is dependent on answers to local RQs.*

Local research questions

RQ2 Are Hemingway’s sentences shorter than those of other writers?

*Answer is dependent on rejecting or failing to reject the following hypotheses:*

- H₀ There is no difference in sentence length between the work of Hemingway and that of other writers.
- H₁ Hemingway uses shorter sentences than other writers.
- H₂ Hemingway uses longer sentences than other writers.

RQ3 Does Hemingway repeat words more than other writers do?

*Answer is dependent on rejecting or failing to reject the following hypotheses:*

- H₀ There is no difference in STTR between the work of Hemingway and that of other writers.
H₁ Hemingway has a lower STTR than other writers.
H₂ Hemingway has a higher STTR than other writers.

RQ4 Does Hemingway use adjectives less frequently than other writers?

*Answer is dependent on rejecting or failing to reject the following hypotheses:*

H₀ There is no difference in frequency of adjectives between the work of Hemingway and that of other writers.
H₁ Hemingway uses fewer adjectives than other writers.
H₂ Hemingway uses more adjectives than other writers.

RQ5 Does Hemingway use adverbs less frequently than other writers?

*Answer is dependent on rejecting or failing to reject the following hypotheses:*

H₀ There is no difference in frequency of adverbs between the work of Hemingway and that of other writers.
H₁ Hemingway uses fewer adverbs than other writers.
H₂ Hemingway uses more adverbs than other writers.

RQ6 Does Hemingway use fewer complex sentences than other writers?

*Answer is dependent on rejecting or failing to reject the following hypotheses:*

H₀ There is no difference in frequency of subordinating conjunctions between the work of Hemingway and that of other writers.
H₁ Hemingway uses fewer subordinating conjunctions than other writers.
H₂ Hemingway uses more subordinating conjunctions than other writers.

Having outlined our research questions, we can now consider what data we need to examine in order to answer them, and what analytical methods to use. Note that we use the term *methods* as opposed to *methodology*. In stylistics, and in other areas of the humanities and social sciences, these terms are often used interchangeably. This, however, is a mistake as there are important differences between them.
A methodology consists of the theoretical ideas, principles and values that underpin a particular set of methods. As Kaplan puts it, methodology refers to ‘the study – the description, the explanation, and the justification – of methods, and not the methods themselves’ (1964: 18; see Stuart 2017 for a full discussion of the difference between the two terms). For our purposes, corpus linguistics constitutes a methodology which underpins and informs the use of a number of analytical methods, including, among others, frequency profiling, collocation and keyness.

RQs 2 to 6 require, in Adolph’s (2006) terms, inter-textual analysis. That is, to answer them, we need to compare a sample of Hemingway’s writing with a sample of the work of other writers. The first consideration, then, is what counts as a representative sample of Hemingway’s writing, and what constitutes a suitable reference corpus for comparison.

We decided that our sample of Hemingway’s work would be made up of his novelistic output, including novels and novellas. The only one of Hemingway’s novels not included in our corpus is *The Torrents of Spring* (1926), as we were unable to obtain an electronic copy of the text. This is a pragmatic issue in the compilation of corpora that researchers often have to face, as Hunston (2002: 27) notes. The eight books that comprise our Hemingway corpus are therefore as follows:

- *The Sun Also Rises* (1926)
- *A Farewell to Arms* (1929)
- *To Have and Have Not* (1937)
- *For Whom the Bell Tolls* (1940)
- *Across the River and into the Trees* (1950)
- *The Old Man and the Sea* (1952)

In total, the Hemingway corpus consists of 783,551 words. The reference corpus we chose was formed from two existing corpora: the LOB Corpus of 1960s written British English and the BLOB Corpus of 1930s written British English (see Appendix 1 for more details of these corpora). Together, these comprise approximately two million words. We chose to combine these two corpora on the grounds that Hemingway’s novel-writing career spanned the mid-1920s to the mid-1980s. Two million words of 1930s and 1960s English therefore seems an appropriate set of data against which to compare his work. Ideally, it would have been useful to compare Hemingway’s work against a corpus of fiction, though for copyright reasons these are hard to come by. Wmatrix (the software used to answer some of our research
questions) is pre-loaded with the BNC Written Imaginative sampler of fiction, but this is only 222,541 words in size. We therefore decided against using this on the grounds that it is smaller than our Hemingway corpus, and generally speaking reference corpora ought to be larger than target corpora (Baker 2004). In order to answer our research questions, we needed to use two pieces of software – WordSmith Tools and Wmatrix (see Appendix 2) – and a number of different methods. We describe these as we discuss the results of our analysis, below.

RQ2 Are Hemingway’s sentences shorter than those of other writers? To answer this question we used WordSmith Tools (Scott 2016). WordSmith’s WordList function provides a variety of descriptive statistics about uploaded files; we used it to compare Hemingway’s sentence length against sentence length in BLOB-FLOB (see Box 4.1). The results are given in Table 4.3.

What these results indicate is that Hemingway’s average sentence length is indeed lower than that of other writers. Consequently, we can reject the null hypothesis in favour of H1. This result would appear to confirm the subjective intuitions of mainstream literary criticism concerning Hemingway’s sentence lengths.

RQ3 Does Hemingway repeat words more than other writers do? To answer this question, we again used the WordList function in WordSmith, specifically the STTR calculation it provides. The results are given in Table 4.4.

Table 4.3 Mean sentence lengths in Hemingway and BLOB-LOB

<table>
<thead>
<tr>
<th></th>
<th>Hemingway</th>
<th>BLOB-LOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean sentence length</td>
<td>10.65</td>
<td>18.81</td>
</tr>
</tbody>
</table>

Table 4.4 Standardised type-token ratios in Hemingway and BLOB-LOB

<table>
<thead>
<tr>
<th></th>
<th>Hemingway</th>
<th>BLOB-LOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>STTR</td>
<td>35.84</td>
<td>39.75</td>
</tr>
</tbody>
</table>
Box 4.1 Using the WordList function in WordSmith Tools (Version 7)

WordSmith Tools is a licence-only software package (see Appendix 2 for details of how to obtain a copy). We used it to generate statistics for the Hemingway corpus. For copyright reasons, we cannot make the Hemingway corpus publicly available. However, the steps below explain how to generate a wordlist, which is the first step in extracting descriptive statistics about a corpus.

1. Open WordSmith Tools and click on **WordList**.

2. The WordList tool will now open up. Click on **File** and **New**. You will see the following dialogue box:
3. Click on **Choose Texts Now** and you will see the following screen:

4. Use the drop-down menu in the top left of the window to locate your corpus files. Then drag and drop them from the left-hand pane to the right-hand pane and click **OK**.

5. Click **Make a wordlist** now and you will see the following screen:

6. Click on the **Statistics** tab at the bottom of the screen to view descriptive statistics for the file(s), including word count, average sentence length and (S)TTR.
RQ4 Does Hemingway use adjectives less frequently than other writers?

Since this question asks about a particular part-of-speech, it was necessary to examine POS-tagged versions of both our target and reference corpora. To do this, we used Wmatrix (Rayson 2009). Wmatrix is a web-based corpus analysis tool that includes a built-in POS tagger known as CLAWS (Constituent Likelihood Automatic Word-tagging System; see Garside 1987). Once a text or corpus has been uploaded to Wmatrix, CLAWS applies POS tags to each constituent word before a second tagger, the USAS semantic tagger (described in detail in Chapter 9), applies tags to mark what semantic category each word belongs to.

In order to answer the question of whether Hemingway uses adjectives less frequently than they are used in BLOB-FLOB, one option was to calculate a normalised frequency for this POS; that is, to determine how many adjectives there are per million words. However, a better measure is a keyness analysis. We noted in the introduction to this chapter that keywords are words that turn up in one corpus more (or, indeed, less) than they do in a reference corpus. Keyness analysis can also be applied to parts-of-speech, where it works on the same principle. Calculating keyness requires a statistical test to determine the significance of the difference between frequency in the target corpus and frequency in the reference corpus. One test that Wmatrix uses is known as log-likelihood. We discuss keyness in full in Chapter 5. For now, we can simply note that a log-likelihood score that exceeds 15.13 is significant in statistical terms (we also explain the reason for this seemingly arbitrary value in Chapter 5). What this means is that if an item has a log-likelihood score higher than this, we can be confident that it is either significantly more prevalent or significantly less prevalent in our target corpus than we would expect it to be, based on its frequency in the reference corpus. More specifically, its frequency is higher (or, in the case of negatively key items, lower) than we would expect it to be based on chance alone.

Since RQ4 asks whether Hemingway used adjectives less than other writers, we used Wmatrix to see whether adjectives are negatively key in the Hemingway corpus; that is, to see whether they turn up less frequently than we would expect them to do based on their frequency in BLOB-LOB. (See Box 4.2 for details of how to upload a text to Wmatrix and carry out a key POS analysis.) The results are shown in Table 4.5 and Column O1 shows the frequency of the item in the target corpus (i.e. the Hemingway corpus), while column O2 shows its frequency in the reference corpus (i.e. BLOB-LOB). Columns %1
and %2 shows what percentage of the target and reference corpora are made up of the item in question. The LL column gives the log-likelihood score. What we can see from the results in Table 4.5 is adjectives of all kind are under-represented in Hemingway’s work, compared with their frequency in the reference corpus. We can, as a consequence of this, reject H₀ in favour of H₁.

### Box 4.2 Uploading a text to Wmatrix and generating a key part-of-speech list

Wmatrix is accessible with a licence at [http://ucrel.lancs.ac.uk/wmatrix3.html](http://ucrel.lancs.ac.uk/wmatrix3.html). Wmatrix can only deal with single .txt files so if your corpus consists of multiple files, these will need to be merged into one. This is a simple process that can be done using the command prompt in both Windows and Mac operating systems. A relevant Google search will generate explanations of how to do this.

1. First, upload your target corpus. To do this, click **Tag Wizard**.
2. Type your chosen filename in the **Enter new folder name** box. Use the **Browse** function to locate your file.
3. Click **Upload now**. Your file will now be tagged for POS and semantic tags. Repeat stages 1 to 3 for your reference corpus.
4. Once tagged, your file is accessible via the **My folders** option. Clicking on the relevant folder will take you to Wmatrix’s analysis options. To generate a key part-of-speech list, go to **Key POS compared to** in the **Keyness analysis** column and select your reference POS compared to: in the drop-down menu. Now click **Go**.

<table>
<thead>
<tr>
<th>POS tag</th>
<th>O1</th>
<th>%1</th>
<th>O2</th>
<th>%2</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>JJ (general adjective)</td>
<td>31,464</td>
<td>4.89</td>
<td>125,377</td>
<td>6.62</td>
<td>2,426.13</td>
</tr>
<tr>
<td>JJR (general comparative adjective, e.g. older, better, stronger)</td>
<td>872</td>
<td>0.14</td>
<td>3,859</td>
<td>0.20</td>
<td>127.69</td>
</tr>
<tr>
<td>JK (catenative adjective, e.g. able in be able to, willing in be willing to)</td>
<td>98</td>
<td>0.02</td>
<td>548</td>
<td>0.03</td>
<td>39.37</td>
</tr>
<tr>
<td>JJT (general superlative adjective, e.g. oldest, best, strongest)</td>
<td>499</td>
<td>0.08</td>
<td>1,971</td>
<td>0.10</td>
<td>36.13</td>
</tr>
</tbody>
</table>
5. Your key part-of-speech list will now appear and can be filtered for either positive or negative keyness using the filter option above the list.

RQ5 Does Hemingway use adverbs less frequently than other writers? To answer this question, we used the same key part-of-speech analysis as for RQ4, this time comparing adverb usage. The results are shown in Table 4.6.

<table>
<thead>
<tr>
<th>POS tag</th>
<th>O1</th>
<th>%1</th>
<th>O2</th>
<th>%2</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGR (comparative degree adverb, e.g. more, less)</td>
<td>165</td>
<td>0.03</td>
<td>1,987</td>
<td>0.10</td>
<td>449.58</td>
</tr>
<tr>
<td>RGT (superlative degree adverb, e.g. most, least)</td>
<td>80</td>
<td>0.01</td>
<td>1,179</td>
<td>0.06</td>
<td>312.80</td>
</tr>
<tr>
<td>REX21 (two-part adverb introducing appositional constructions, e.g. for instance)</td>
<td>13</td>
<td>0.00</td>
<td>402</td>
<td>0.02</td>
<td>154.95</td>
</tr>
<tr>
<td>REX (adverb introducing appositional constructions, e.g. for instance)</td>
<td>4</td>
<td>0.00</td>
<td>212</td>
<td>0.01</td>
<td>95.02</td>
</tr>
<tr>
<td>RR31 (three-part general adverb, e.g. and so on)</td>
<td>29</td>
<td>0.00</td>
<td>373</td>
<td>0.02</td>
<td>89.23</td>
</tr>
<tr>
<td>RA (adverb, after nominal head, e.g. else, galore)</td>
<td>51</td>
<td>0.01</td>
<td>464</td>
<td>0.02</td>
<td>78.53</td>
</tr>
<tr>
<td>RGQV (wh-verb general adverb, e.g. wherever, whenever)</td>
<td>2</td>
<td>0.00</td>
<td>67</td>
<td>0.00</td>
<td>26.54</td>
</tr>
<tr>
<td>RG21 (two-part general adverb, e.g. up to, ever so)</td>
<td>8</td>
<td>0.00</td>
<td>107</td>
<td>0.01</td>
<td>26.41</td>
</tr>
<tr>
<td>RR41 (four-part general adverb, e.g. all of a sudden)</td>
<td>9</td>
<td>0.00</td>
<td>99</td>
<td>0.01</td>
<td>20.60</td>
</tr>
<tr>
<td>RRT (superlative general adverb, e.g. best, longest)</td>
<td>56</td>
<td>0.01</td>
<td>293</td>
<td>0.02</td>
<td>17.56</td>
</tr>
</tbody>
</table>
What Table 4.6 shows is that a large number of adverb types are indeed under-used in Hemingway’s work. However, in this case, this does not provide us with the requisite information to reject $H_0$ in favour of $H_1$. This is because although we find many adverbs under-represented in Hemingway, we also find other types to be over-represented. These are shown in Table 4.7.

Table 4.7 demonstrates the paucity of intuition regarding the question of adverb usage in Hemingway. Concerning this question, we cannot reject any of our hypotheses. Instead, the data necessitates us taking a more nuanced look at adverb usage in the corpus.

**RQ6 Does Hemingway use fewer complex sentences than other writers?**
Since the definition of a complex sentence rests on it containing a dependent clause, to answer this question we once again used Wmatrix’s key POS function, this time to search for subordinating conjunctions in both our target and reference corpora. The results are shown in Table 4.8.

The results in Table 4.8 show that coordinating conjunctions are over-represented in the Hemingway corpus. Given $H_1$ (Hemingway uses fewer subordinating conjunctions than other writers), this result may seem understandable. If subordinating conjunctions are under-used,
then perhaps as a consequence coordinating conjunctions are being overused. However, what Table 4.8 also shows is that not only does Hemingway use subordinating conjunctions, they are over-represented in his work. We can therefore reject the null hypothesis in favour of H2.

### Table 4.8 Keyness of coordinating and subordinating conjunctions in Hemingway and BLOB-LOB

<table>
<thead>
<tr>
<th>POS tag</th>
<th>O1</th>
<th>%1</th>
<th>O2</th>
<th>%2</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC (coordinating conjunction, e.g. <em>and</em>, <em>or</em>)</td>
<td>26,069</td>
<td>4.05</td>
<td>57,584</td>
<td>3.04</td>
<td>1,435.46</td>
</tr>
<tr>
<td>CCB (adversative coordinating conjunction, e.g. <em>but</em>)</td>
<td>3,975</td>
<td>0.62</td>
<td>8,555</td>
<td>0.45</td>
<td>257.35</td>
</tr>
<tr>
<td>CS (subordinating conjunction, e.g. <em>if</em>, <em>because</em>, <em>unless</em>, <em>so</em>, <em>for</em>)</td>
<td>6,485</td>
<td>1.01</td>
<td>15,621</td>
<td>0.82</td>
<td>181.16</td>
</tr>
</tbody>
</table>

4.10 Conclusion

In this chapter we have explored particularly the practice of determining clear research questions. We have also discussed the importance of intuition in corpus stylistics and its value as part of the hypothesis formation process. In particular, we showed how hypothesis testing can work as a method for identifying the linguistic source of impressions about a text, thereby systematising the notion of an analysis being used to support or invalidate readers’ interpretations. What the analyses in the previous section demonstrate is that by taking a methodical approach to the testing of falsifiable hypotheses, we are able to test the assertions of literary critics concerning Hemingway’s authorial style. In some cases (RQs 2, 3 and 4), our evidence supports literary critical claims. In others (RQs 5 and 6), it doesn’t. Ultimately, then, our answer to RQ1 (Are literary critical claims about Hemingway’s style accurate?) must be ‘yes and no’. And here is where Spitzer’s (1948) notion of the philological circle comes to the fore. Spitzer (1948) argues that understanding literary meaning involves a constant movement between linguistic description and literary appreciation (see Leech and Short 2007: 1–4 for a summary of his position). It is not enough simply to prove literary critical positions wrong, as we have done here. We need to use such analyses as starting points for further investigation. One of our findings, for instance, was that, counter to literary critical expectations, Hemingway does use subordinating conjunctions; in fact, he overuses them compared with their frequency in our reference corpus. An interesting question at this point, then, is: why do literary critics perceive that subordination is lacking in Hemingway’s work? The potential
explanations for this can be used to generate further hypotheses to be tested. Is it the case, perhaps, that the overuse of coordinating conjunctions has the effect of backgrounding the subordinating conjunctions, causing them not to be perceived as salient, despite the fact that they are used just as much? Or could it be the case that subordinating conjunctions are only used in particular parts of Hemingway’s novels: direct speech, perhaps, or stretches of first-person narration? In many cases, the answers to our research questions send us back to the corpus and its constituent texts to investigate particular issues further.

Given the value of corpus approaches in answering such questions, Mahlberg (2013b: 12) suggests an adaptation of Spitzer’s (1948) philological circle, adding ‘corpus linguistic description’ between ‘literary appreciation’ and ‘linguistic description’. Clearly, corpus methods have much to add to the stylistic analysis of literary (and non-literary) texts, though we see corpus linguistic description as simply a sub-type of linguistic description, and for preference would not view this as an additional stage in the philological circle. Nonetheless, the point is clear: corpus-based stylistics has much to offer to both the stylistician and the literary critic. In the next chapter, we discuss a key element of corpus-based stylistics: statistics.

Notes

1. This famous example comes from Halliday (1966). It may be the case that lexical constraints vary according to dialect, of course.
2. Other alternatives are available for such a study. For example, a corpus of Dickens’s novels and reference corpora of nineteenth-century fiction are publicly accessible via the CLiC (Corpus Linguistics in Context) project website (Mahlberg et al. 2016). CLiC offers a web interface for studying a wide variety of corpora of nineteenth-century fiction, particularly that of Dickens.