Imitation and Anonymity in Linguistic Identity Disguise

Fiona Jane Kelcher

Doctor of Philosophy

Aston University
August 2021

Aston University

Anonymity and Imitation in Linguistic Identity Disguise

Fiona Jane Kelcher

Doctor of Philosophy

August 2021

Thesis Summary

Authorship attribution can be highly accurate, but most techniques are based on the assumption that authors have not attempted to disguise their writing style. Research has found that when writers had deliberately altered their style, commonly used authorship analysis techniques only performed at the level of random chance. This is problematic because many forensic authorship cases investigate documents where it is believed that an author has tried to impersonate somebody else for criminal purposes, and has attempted to adapt their writing style to do so.

This study uses a corpus of scripts from the BBC drama, *The Archers,* to explore how authors write different characters' voices. Scriptwriters need to adapt their writing style to create the different characters' dialogues, and this fictional identity disguise is used as a proxy to examine authorship analysis techniques in forensic linguistics.

The thesis begins with a literature review exploring the nature of linguistic identity and literary characterisation. It considers the advantages and disadvantages of using fictional data to address forensic problems. There are three main studies: firstly, a quantitative analysis comparing inter-author consistency and variation of authorship analysis features; the second study is a qualitative, stylistic analysis of characterisation, exploring lexical choice, use of dialect, and (im)politeness strategies. The third study is a corpus analysis of the different pragmatic functions of shared lexical tokens.

The studies showed that as writers adapted their linguistic style to create different characters, results for commonly-used attribution techniques were observably affected. Some linguistic identities were more distinctive than others, and some authors were more clearly identifiable than others. At a pragmatic level, authors showed more inter-character consistency, and a reduced ability to anonymise their own linguistic traits. This reinforces the importance of investigating linguistic identity disguise at higher levels of language analysis, in addition to lower-level, structural features.

Keywords*: adversarial stylometry, authorship analysis, forensic linguistics, linguistic identity disguise, pragmatics, stylistics*

Acknowledgements

Firstly, I would like to thank *The Archers* scriptwriters who so generously allowed me to analyse their scripts through the lens of Forensic Linguistics. I would also like to thank the BBC, in particular, Kim Greengrass, for supporting my project and facilitating access to the scripts. At Aston, my thanks go to my supervisor, Professor Tim Grant, for his expertise, advice and encouragement; and to Dr Abigail Boucher, for her support throughout my PhD. My thanks also go to my examiners, Professor Dan McIntyre and Dr Krzysztof Kredens, for their advice and feedback on my research.

For the endless support of my family, I am forever grateful.

Table of Contents

List of Tables

List of Figures

Non-disclosure agreement

I have a Non-Disclosure Agreement with the British Broadcasting Corporation (BBC), allowing access and thesis publication rights, and the individual scriptwriters have also given their consent for their scripts to be used in this study. Writers have been anonymised and are referred to by number.

1. Introduction

1.1 Authorship Attribution and Linguistic Identity Disguise

Linguistic identity disguise is defined in this thesis as an author's attempt to create or adopt a linguistic persona: this includes both real-life cases of people assuming false identities, but also writers of fiction creating characters through dialogue. Authorship attribution is the task of comparing the linguistic features of a group of known texts to an anonymous text with the aim of determining the author(s) of the anonymous text. Computational linguistic authorship analysis, which generally uses large datasets for statistical authorship attribution, has achieved high levels of accuracy of up to 95% (Grieve, 2007:262), but is particularly vulnerable to authorship cases where the writer has deliberately altered their style, to the point where the results of the authorship attribution algorithms are no more accurate than random guesswork (Brennan et al., 2012:461). In contrast, close linguistic analysis, looking for consistency and variation between texts (for example, Grant, 2012), seems more robust when analysing texts written by an author who is disguising their style.

With the exception of Grant and MacLeod (2017, 2018, 2020), relatively little has been written on the nature of linguistic identity disguise, outside of the study of literary pastiche. This is surprising because forensic casework often arises precisely because the police believe that linguistic identity disguise has taken place: for example, the Amanda Birks case, described in Grant (2012) and the Jenny Nicholl case, described in Coulthard (2010), where suspected murderers have attempted to impersonate their victims. A significant issue with research into linguistic identity disguise is the lack of suitable data. In cases such as the murders of Amanda Birks and Jenny Nicholl, the number of text messages suspected to be impersonations is very limited compared to the larger amounts of data known to be written by the victim and the suspect. Further, in many cases of online identity assumption, such as adults pretending to be teenagers for the purpose of sexual grooming, or in cases of romance fraud, it is not always definitively known when deception is taking place, or who the true identities are of the people involved. Even if the nature of deceit can be verified, Grant and MacLeod (2018) argue that because there are so many ethical sensitivities regarding research and publication using data involving underage victims, there is “a need to be able to work more openly with less toxic data” (2018:60).

One way to address this issue is to use drama scripts. In drama, scriptwriters write dialogue for multiple characters. By writing only in the voices of the characters in the drama, the writers are carrying out a form of linguistic identity disguise. A further layer of complexity occurs in drama serials and soap operas: episodes are written by individual writers, but characters and storylines are shared across multiple episodes. This requires scriptwriters to synthesise these characters’ voices in order to create cohesive characters. My study uses drama scripts as a proxy to address the forensic issue of authorship analysis in cases of linguistic identity disguise, to explore how writers adapt their style when writing the voices of others.

1.2 Data

The drama scripts used are from *The Archers*, a long-running radio drama serial on BBC Radio 4, set in a farming community in the fictional village of Ambridge. *The Archers* originally began in 1951 with input from The Ministry of Agriculture, Fisheries and Food, in order to disseminate information to increase productivity of farms and smallholdings during the rationing years of post-war Britain. The programme has retained its agricultural setting, but no longer has a public information remit, and is described on its programme website as, “contemporary drama in a rural setting”. Each episode is approximately 15 minutes long, and the programme is broadcast six evenings a week, with early afternoon repeats the next day, and a weekly omnibus on Sunday mornings.

The data consist of 1440 studio scripts, written between 2010 and 2017, by six scriptwriters. Each script is usually 2600-2700 words long. Excluding stage directions, the characters’ dialogue usually comprises around 2000 words per script. My thesis analyses these six scriptwriters, who wrote regularly for *The Archers* for the whole period 2010-2017. The six writers are anonymised, and referred to as Writers 1 – 6. Each writer number remains fixed throughout my thesis. For example, Writer 1 refers to the same person throughout the whole thesis. Extracts from *The Archers* scripts are referenced by the Writer number and the year of transmission from which the quotation is taken, for example, “Writer 1, 2010”. For a brief character profile of twenty frequently-speaking characters in the drama, see Chapter 4 (Section 4.3).

This dataset has a number of advantages. Firstly, the corpus is over three-million words long, so provides an largescale opportunity to analyse cases of known identity disguise, albeit fictional rather than forensic. Secondly, the audio-only medium means that characterisation is necessarily conveyed almost entirely through language. Thirdly, all scripts are individually authored, unlike some TV series, where individual scripts can be written collaboratively by multiple authors. Access to the data came from my previous role producing and directing Radio Drama at the BBC, and any discussion of production processes in my thesis is based on my professional experience.

Whilst scripts are individually written, the ‘studio scripts’ which I received will have had some minor amendments made by members of the production team. Writers submit a week’s worth of scripts (six fifteen-minute episodes) at a time. The tight deadlines involved in the production schedule mean that standard practice is for script editors to correct minor continuity errors and make small alterations. For example, if characters discuss an event happening on the wrong day, or mistake a name place, this would be corrected by the script editors. If more than one or two lines in a scene requires a re-write, this will usually be sent back to the individual writer for them to alter, rather than the script editor changing it themselves. These second drafts are checked by the production staff, and may receive minor amendments, usually no more than a line or two at a time, before the scripts are formatted into a studio script which contains technical information for crew, and is distributed ready for recording.

Ideally, the writers’ first draft scripts would have been used for my analysis, but for various confidentiality and production reasons, this was not possible, and the studio scripts were used instead. However, since authorship attribution is a process of looking for repeated patterns of linguistic style, these patterns should remain observable, despite some interference during the script editing process. In other authorship attribution studies, a standard editing process for all selected writers was considered to be a sufficient control measure for writing samples that had been edited, as discussed here:

> Note that since all texts come from the same newspaper, they are expected to have been edited according to the same rules, so any significant difference among the texts is not likely to be attributed to the editing process. (Stamatatos, 2012:430)

Whilst it is not ideal that the scripts have had some interference, this was the necessary compromise to gain access to such a large dataset. The dataset, and the preparation and compilation of sub-corpora for each study is discussed in detail in Chapter 4.

1.3 Research Questions

To contribute to the research on authorship attribution in cases of linguistic identity disguise, my thesis has a superordinate research question: to what extent are dramatists able to create linguistically distinctive characters, and maintain the consistency of those characters' styles, whilst simultaneously suppressing their own authorial style? This overarching research question is explored through a number of sub-questions. These are:

1) To what extent do quantitative, structural-level analyses identify character style rather than authorial style?

2) Are writers able to identify consistent intra-character features?

3) Can higher-level pragmatic features provide a base for authorship analysis in cases of linguistic identity disguise?

Each of these sub-questions is the focus of each main study (Chapters 5-7) in turn.

Regarding 'structural level' and 'higher level' domains of language, I am following Grant and MacLeod's four domains of language (2020), where the lower-level structural domain is concerned with features such as typography, orthography, morphology, and syntax, in contrast to 'higher-level' domains of language concerned with meaning, interaction, and social behaviour. This is discussed in more detail in Chapter 3.

1.4 Outline of Thesis

This present introductory chapter is followed by a Literature Review (Chapter 2), which discusses the relevant literature on authorship attribution methods, sociolinguistic identify performance, and style. It considers advantages and disadvantages of using fiction as data for exploring forensic issues. Chapters 3 and 4 are both concerned with Methodology: Chapter 3 is a literature review of the specific analytical methods used in Chapters 5-7, for example, the Glaswegian dialect analysed in Chapter 6. Chapter 4 is a Methodology chapter, which describes my dataset in more detail, and explains how the individual sub-corpora for each study were selected and prepared. The chapter also explains the Methodology used in each study in the three main analytical chapters, and discusses the rationale behind my choices.

Chapters 5-7 comprise three separate studies, each adopting a different approach to exploring linguistic identity disguise, and each addressing one of the thesis' sub-questions, as set out in Section 1.3. The first of these, Chapter 5, is an exploration of quantitative authorship attribution, investigating the extent to which quantitative, structural analyses are able to identify character style, and the extent to which they are able to identify authorial style. The second study, Chapter 6, is more qualitative. This chapter addresses the second sub-question, and explores how consistently and closely individual writers are able to imitate characters' established ways of speaking, through three separate analyses of three of the more distinctive characters from the data. The chapter analyses one pertinent linguistic feature per character: these are lexical choice, presentation of dialect, and (im)politeness strategies. The study explores how characters are linguistically realised, and evaluates the extent to which writers are able to create and maintain character style of these shared characters. Chapter 7 is concerned with pragmatics, addressing the third sub-question, by analysing three sets of duologues between couples to explore the difference in pragmatic function between the way the writers use the word *oh*. The purpose of this is to explore whether higher-level linguistic analysis can discriminate between more pairs of authors than a structural analysis. This is then linked back to the overarching research question, to investigate whether the scriptwriters are able to suppress their own authorial style whilst creating linguistically distinctive characters. The final chapter, Chapter 8, draws together the findings of the studies into a cohesive discussion, which finds that as writers use different linguistic techniques to create multiple fictional voices, the results of commonly-used measurements of style were observably affected. Some linguistic identities were more distinctive than others, and some authors were more clearly identifiable than others. At a pragmatic level, I found that authors showed a reduced ability to anonymise their own linguistic traits. This reinforces the importance of investigating linguistic identity disguise at higher levels of language analysis, in addition to lower-level, structural features.

1.5 Ethical Considerations

I take a neutral view of 'style' and believe it is neither inherently positive or negative for writers if their authorial style is identifiable. Somers and Tweedie (2003) point out that a quantitative correlation of linguistic features can suggest a skilful linguistic imitation; it does not necessarily correlate with a pastiche that is well-received by critics or loved by readers. Therefore, the results of my analysis are not in themselves an indicator of skilful dramatic writing, so the results should not have a damaging effect on any of the participants.

On her retirement, after more than 40 years as an *Archers* scriptwriter, Mary Cutler reflected on the relationship between authorial style, and the style of the programme:

> I have had the opportunity to work collaboratively on the storylines with so many immensely creative people, including the other writers and *The Archers* production team. But after that I've had the freedom to dramatise those stories, every word of it mine, so that people could – and did – say "Oh that sounds like one of Mary's", the miracle of the show being that it still sounded like *The Archers*. (BBC Radio 4 Blog, 2019)

This fascinating relationship between the authorial voice and the voices of multi-authored characters, is at the centre of my thesis.

2. Literature Review

2.1 Introduction

The purpose of this chapter is to review the academic literature on both quantitative and qualitative methods of authorship attribution, and to discuss different forms of linguistic identity performance and linguistic identity disguise. Reviewing seminal works on style, such as Crystal and Davy (1969), Leech and Short (1981), and key publications on the nature of characters in drama, such as Short (1989), I set out my own position on the relationship between authorial style and character style and relate it to forensic research on authorship attribution and authorship synthesis. Advantages and disadvantages of using fictional data to explore forensic questions are also discussed. These topics relate back to my overarching research question of exploring linguistic identity imitation and anonymity.

2.2 Types of Authorship Analysis

Authorship analysis is described by computational scientist Juola as "one of the oldest and one of the newest problems in information retrieval" with questions about linguistic identity reaching back to the time of the Old Testament (2008:5). It is a field with a number of different purposes. One purpose is *authorship profiling*, where the analyst attempts to describe the sort of person who produced a given text. A second purpose is *comparative authorship analysis*. McMenamin describes the process:

> Cases of questioned authorship typically present the linguist with a questioned writing to be first contrasted (for possible exclusion of the author) then compared (for possible identification of the author) to a set of exemplar writings known to have been written by a writer suspected of authoring the questioned material. (2020:545)

This type of authorship analysis is often called *authorship attribution*. A third purpose is *authorship verification*, described by Koppel et al. as a process where the analyst aims to verify whether an unknown text was written by a particular author (2012:321). More recently, Grant and MacLeod introduced the new task of *authorship synthesis* (2018), which MacLeod describes as "the taking-over of an individual's online identity for the purposes of intelligence gathering and/or securing an arrest"

(2020:159). Grant and MacLeod distinguish between cases where an undercover officer is assuming the identity of a particular person, for example a victim of online abuse, and cases where the undercover officer is "legend-building", i.e. inventing a linguistic persona.

Just as authorship analysis has a multitude of purposes, there are also multiple approaches to authorship analysis from many different disciplines. Juola writes:

> Papers on authorship attribution routinely appear at conference ranging from linguistics and literature through machine learning and computation, to law and forensics. Despite – or perhaps because of – this interest, the field itself is somewhat in disarray with little overall sense of best practices and techniques. (2008:2)

In the fifteen years since Juola's statement, authorship analysis papers have continued to be published from the same wide range of disciplines, and the ever-increasing use of online communication has led to studies on new types of language use: for example, analyses of Tweets by MacLeod and Grant (2012), and Clarke and Grieve (2017, 2019) among others, an increased interest in disinformation and 'fake news' on social media, and an interest in adversarial stylometry and authorship anonymisation and imitation. As might be expected from the number of disciplines carrying out authorship analysis, there are many different approaches, methods and terminologies.

This study explores authorship synthesis and authorship attribution. It is a study of the ways that people adopt other linguistic personae, but also draws on many methodologies and questions used in authorship attribution.

2.3 Authorship Attribution

Wright (2017) summarises authorship attribution as, "the process in which linguists set out to identify the author(s) of disputed texts using identifiable features of linguistic style, ranging from word frequencies to preferred syntactic structures" (2017:213). Quantitative methods, as described by Grieve, have been central to many authorship studies:

> In quantitative authorship attribution, the values of textual measurements in the anonymous text are compared to their corresponding values in a series of possible author writing samples, in order to determine which possible author writing sample is the best match. (2007:251)

Quantitative techniques range from fairly simple, descriptive measurements of linguistic features, to sophisticated algorithms produced by computational linguists. The features analysed in these techniques are almost always at the structural level of language, such as grapheme distribution or word length. In the following sections, various models of authorship attribution are discussed.

2.3.1 Types of Authorship Attribution Problem

Juola (2008), and Koppel et al. (2012) outline a number of problems for which quantitative approaches have been used. There are different types of enquiries, based on the number candidate authors and amounts of anonymous and comparison data. Koppel et al., in their discussion of the different sorts of authorship problems, describe the ‘closed-class’ problem:

> The simplest kind of authorship attribution problem—and the one that has received the most attention—is the one in which we are given a small, closed set of candidate authors and are asked to attribute an anonymous text to one of them. (2012:317)

Often there are only two candidate authors to choose from. Grant observes that, “In forensic casework, this is perhaps the most common type of problem, at least when the linguist is commissioned by the police” (Grant, 2020:564).

The “closed-class” problem is closely linked to the “open-class” problem: “*Given a particular sample of text believed to be by one of a set of authors, determine which one, if any*”. This can even be as broad a question as, “*here is a document, tell me who wrote it”* (2008:6). By its nature, the “open-class” problem is harder to solve. Koppel et al. observe that the suspected number of candidate authors can be very large, even running into thousands of potential authors, and that the candidate authors may not even include the author of the questioned document (2012:317).

2.3.2 Idiolect

From whichever discipline authorship analysis is being explored, at the centre of the enquiry is the theory that that all speakers and writers use language in a way which differentiates themselves from others speakers of the same language. Coulthard, in his influential paper on the individuality of language use, wrote:

> Whereas in principle any speaker/writer can use any word at any time, speakers in fact tend to make typical and individuating co-selections of preferred words. This implies that it should be possible to devise a method of *linguistic fingerprinting* – in other words that the linguistic 'impressions' created by a given speaker/writer should be usable, just like a signature, to identify them. (Coulthard 2004:432)

Coulthard argues that we all have our own version of the language we speak, and that it is not just the use of individual words which make authors distinctive, but also the sequences or groupings of words they use. He contends that these patterns of co-selections make every writer or speaker linguistically unique:

> The linguist approaches the problem of questioned authorship from the theoretical position that every native speaker has their own distinct and individual version of the language they speak and write, their own *idiolect*, and … this idiolect will manifest itself through distinctive and idiosyncratic choices in texts. (2004: 432)

Coulthard's claim that, in theory, every speaker has their own *idiolect,* (a term coined by Bloch (1948:7), and contrasted to sociolect (the linguistic patterns of a group of speakers)), has been highly influential. Yet, whilst 'fingerprinting' is a useful analogy for the trail of language we leave behind us, it would be overstating the case to suggest that language can be used to identify an individual with the same levels of confidence of fingerprinting or DNA: as Grant points out, language is naturally variable (2020:570), and unlike physical features, all linguistic features are in theory available to any speaker of that language. Juola states that there are strong theoretical reasons underpinning the concept of an idiolect. He writes:

> Since every person has to learn "language" by themselves, and their experiences as language learners differ, so will the language" they learn differ in micro-aspects. On the other hand, there are also good practical reasons to believe that such fingerprints may be very complex, certainly more complex than simple univariate statistics such as average word length or vocabulary size. (2008:7)

Whilst the principle of an idiolect has been important in identifying the author of an anonymous text, others have argued that it is not necessary to prove that all speakers and writers use language in entirely different ways. Grant writes:

> Even if the first claim here, that every speaker has their own idiolect, can be sustained, there is no necessary implication from it that an individual's idiolect will be measurable in every text produced by that person, whatever its length. It would be perfectly rational to hold

> Coulthard's view and to also hold that a substantial and varied body of text would be required before manifest idiolectal features became noticeable or measurable. (2020:559)

Grant's position is that even if speakers of a language have individuating aspects to their language use, these will not necessarily feature in every text they write, or at least not frequently enough to be used to differentiate them. On this basis, he advocates an approach of pairwise comparison:

> The issue of linguistic distinctiveness between individuals has two levels, which may be independent. If it can be demonstrated that the suspect exhibits a consistent style in text messaging and also that the victim has a consistent but different style then the first level of distinctiveness will have been proved. I shall refer to this as pair-wise distinctiveness and I will argue that answering this question does not depend upon a strong theory of idiolect, but only upon the degree of consistency of style within each author and the difference which is demonstrable between them. (2020:565)

In this sense, Grant argues that the pairwise approach does not rely on the existence, or theoretical existence, of an idiolect, but instead, on observable differences between the known writing of two sets of authors (which can be repeated multiple times to compare more than two authors). This principle relies on the author remaining consistent in their style, and, of course, distinct from the paired author. With or without adhering to a theory of idiolect, it is of course possible that some pairs of writers will display more differences between them than other writers. As Grant argues, the authorship analysis may depend on the concept of an idiolect, but "practically and methodologically authorship analysis depends on the facility to detect consistent patterns of language use" (2020:559).

This brief discussion of the underpinning ideas surrounding the idiolect, consistency, and distinctiveness now leads into a discussion about specific features of linguistic style which can be used to discriminate between authors.

2.3.3 Models of Attribution

From a stylistic perspective, McMenamin breaks down authorship attribution approaches into three models: resemblance, consistency and population. McMenamin explains that the resemblance model is used when external factors limit the number of possible authors, citing the example of a parent arguing for custody of children and referring to specific details that only a handful of people would know (2020:542).

Consistency is the model used to compare texts by known and unknown authors to assess which known author's writing is most consistent with the anonymous text. McMenamin explains that, "establishing the internal consistency of a group of writings is frequently the first step in a resemblance case when external circumstances do not demonstrate common authorship of a body of questioned writings" (2020:542-3). The final model McMenamin discusses is the population model, citing the example of threatening letters to the head of a large company, where there is a high number of candidate authors. With the population model, McMenamin's method is to work through a large pool of candidate authors to exclude all but the real author. In this study, consistency is the most appropriate model, because it is a comparison of different, possible authors. In many cases of authorship synthesis, the resemblance model becomes important, because a linguistic impersonator would be expected to know details about their target's life, as part of the process of assuming their identity.

2.3.4 Quantitative Methods

Although modern computing has changed the nature and scope of quantitative authorship analysis, it is an area of research that dates back to the 19th century. Its premise is to measure textual features, and use these measurements to distinguish between writers. Discussing quantitative, computational techniques, Juola explains that:

> The arrival of modern statistics made it possible to investigate questions of authorship in a more sophisticated fashion, and the development of modern computers and large corpora have made it practical to investigate these questions algorithmically via information retrieval techniques. (2008:6)

Evaluative surveys of quantitative attribution methods can be found in Juola (2008), Koppel et al. (2009), Stamatatos (2009), and a quantitative comparison in Grieve (2007). In 1964, Mosteller and Wallace famously used quantitative methods to identify the author of the disputed Federalist Papers. Stamatatos, in his survey of modern authorship methods, refers to the seminal study, observing that:

> Since then and until the late 1990s, research in authorship attribution was dominated by attempts to define features for quantifying writing style, a line of research known as "stylometry" (Holmes, 1994, 1998). Hence, a great variety of measures, including sentence

> length, word length, word frequencies, character frequencies, and vocabulary richness functions, had been proposed. (2008:538)

Common features analysed are function words, word length, sentence length, and type-token ratio, (which measures vocabulary richness by dividing the number of *different* words in a text (types) by the *total* number of words (tokens). These features are discussed in Grieve (2007), Wright (2017), Argamon (2009, 2012), and Rudman (2016)) among others. Stamatatos (2009) outlines some of the predominant variables used by computer scientists and computational linguists in quantitative authorship attribution, including average word/ sentence length, vocabulary richness measures, function word frequencies, and word, character and parts-of-speech.

Grieve (2007) compares the efficacy of these various quantitative methods on a single dataset. Grieve selects 39 textual measurements from the many identified, and groups them into categories, and compares their relative efficacy. Categories of style-marker selected include word length, sentence length, vocabulary richness, grapheme frequency, word frequency, punctuation mark frequency, collocation and character level n-gram frequency. From these results, Grieve sets out a recommended process for computational authorship attribution. He states that the investigator should identify the possible authors by analysing external evidence of the anonymous text, then compile a corpus of possible authors. Thirdly the investigator should "test a wide range of attribution algorithms on the corpus of possible authors so as to establish which algorithms can best distinguish between that particular set of possible authors" (p.267), before testing "various weighted combinations of the best algorithms on the same corpus of possible authors" (p.267). Once this has been done, the investigator uses the algorithm to compare the anonymous text to each corpus, to try to identify the closest match. One interesting point from this set of recommendations is the need to test, for each separate case, which algorithms can best distinguish between a set of authors. In other words, Grieve's results do not suggest that textual features can be ranked in a fixed order of usefulness to the authorship analyst, but instead that the features – or combination of features – which are most useful, will vary between cases.

2.3.5 Explicability of Quantitative Methods

Discussing computational approaches, Wright observes that while quantitative methods have achieved high accuracy in attribution tasks, there is a lack of linguistic theory to underpin the results:

> This research is unquestionably valuable; there is now little doubt that by using a combination of linguistic features and a sophisticated machine learning technique or algorithm we are able to successfully identify the most likely author of a text. What we cannot do with the same confidence, however, is explain why these methods work. (2017:214)

Koppel et al. (2009) discuss the same issue, and its impact on the use of forensic linguistics in the legal process:

> The accuracy of current authorship attribution technology depends mainly on the number of candidate authors, the size of texts, and the amount of training texts. However, this technology is not yet reliable enough to meet the court standards in forensic cases. An important obstacle is that *it is not yet possible to explain the differences* between the authors' style. (2009:554, emphasis added)

Kredens and Pezik (2019) refer to the lack of explanation for why certain features perform better than others using the metaphor of the black box (a device where the inputs and outputs are known, but the inner processes within the box are a mystery). In forensic linguistic terms, this suggests a process where the textual features are known, and the results of the attribution process are also known, but there is a lack of knowledge about the process, and about why certain features perform the way they do in authorship tests. Secondly, even if the linguist understands the process, it may be problematic in a courtroom, if the lay jury does not. Kredens and Pezik write that, "The forensic linguist needs to be both certain of the validity of his/her findings and able to explain them to lay triers of fact; s/he needs to know what actually happens inside the black box" (2019). This lack of theoretical linguistic explanation remains an issue in quantitative authorship attribution.

A further issue is the amount of data required to run a statistical analysis. Koppel et al. observe that in forensic casework, there are often too many candidate authors, that the candidate author may not be included in the analysis, and finally, that the anonymous text, or the sample texts are too short (2012:284). If the samples are too short, this is likely to affect the reliability of certain statistical tests. Grieve et al. (2019) note that in texts of under 500 words, a system of presence and absence is generally more effective than a statistical analysis, although other recent papers have focused on using established quantitative techniques on texts as short as individual tweets (for example, Clarke and Grieve, 2019).

One quantitative method which has been shown to be highly effective, and produces explicable results, is a word-level n-gram level analysis. Word-level n-grams have been used in various authorship attribution studies. Coulthard (2004) discusses the importance of strings of words in authorship attribution, notably described as an author's "typical and individuating co-selections of preferred words" (2004:431). Juola (2013) uses all three-word sequences in his data, and Larner (2014) explores fixed phrases in authorship attribution. Grieve et al. (2018) use a new technique they term *n-tram tracing* to attribute to Bixby letter. Grieve et al. explain that short texts, such as the Bixby letter, are often attributed by a forensic linguist selecting linguistically distinctive features, and then comparing these to the known writings of possible authors. They point out a number of issues, including potential bias in feature selection, differing amounts of author data available, a reliance on the analyst's judgement, and a lack of replicability (p.497). They explain that, "the basic idea behind n-gram tracing is to calculate the percentage of n-grams that occur in a questioned document that also occur at least once in a possible author writing sample." Once this process has been repeated for all authors, "the text is then attributed to the possible author whose writing sample contains the highest percentage of the n-grams 250 from the questioned document" (p.499). In this methodology, n-grams can be either character-level, or word-level.

Grieve et al. found word-level n-grams were good at attributing the writings of Lincoln and Hay, but were not as accurate as character-level n-grams. They found that "the analysis of 4- to 12-character n-grams and 1- to 3-word n-grams was especially useful for distinguishing between Lincoln and Hay" (p.506). Despite their lower discriminatory power, Grieve et al. discuss some benefits of analysing word-level n-grams, noting that, "Although their discriminatory value was found to be weaker, it is more instructive to consider unique word-level n-grams rather than unique character-level n-grams, because word-level n-grams are less common, more distinctive, and more interpretable." (p.508).

Wright (2017) also investigates the effectiveness of word n-grams in authorship attribution, arguing that "they offer an objective way of capturing linguistic output of individuals and measuring similarity between texts" (p.220). Wright investigates whether word n-grams can be used to identify the author of a disputed text, and secondly, focuses on one author, to explore which n-grams were useful in identifying him as author of a text. Wright's study aims to "make a case for word n-grams as theoretically-motivated features for authorship analysis that can be used to attribute texts to their correct authors, and for which differences between authors can be explained" (p.237). A notable, and relevant, finding in this paper is that the word-gram method captured some authors more accurately than others, and that the most effective length of n-gram to attribute authorship varied depending on the author. It is likely that, as Wright (p.238) suggests, some writers may be more attributable than others using this method; in my study, it was also shown that some characters are more linguistically distinctive than others.

2.3.6 Qualitative Methods

The approach taken in my thesis is predominantly stylistically focused, with one quantitative study. McMenamin evaluates the relative benefits of quantitative versus qualitative approaches:

> Qualitative inquiry is rigorous if conditioned by careful framing of research questions, systematic observation, data that are the direct outcome of observation, reliable methods of description and analysis, valid interpretation of results, and a statement for the basis of every conclusion (Johnstone 2000). Qualitative evidence is also generally more 'demonstrative' than qualitative results, meaning the use at trial of documents, charts and diagrams to illustrate testimony of the expert witness, which is presented to prove or disproved allegations of authorship. (2020:543)

McMenamin also discusses the problems of qualitative approaches, and lists four issues. Firstly, "the selection of variables used for comparison and contrast of styles is arbitrary and subjective: the criteria for selection of style markers do not appear to be specified or justified" (2020:552). McMenamin defines *style markers* as "style characteristics" such as frequently used vocabulary, sentence structure, or spelling mistakes (1993:120), which together make up a writer's "composite" style. A second objection stated is that the frequency of occurrence of the variables is not defined, which can result in a lack of statistical rigour. Thirdly, the data being analysed may not have an available reference corpus to provide a linguistic norm. McMenamin's final objection is that the relative significance of each variable is not measurable, because of the lack of evidence to determine the levels of conscious control a writer has over their linguistic choices, as discussed above. McMenamin proposes the assumption that, "the most tell-tale markers are those least consciously used" (2020:252), but there is a lack of research into this topic, as discussed in 2.3.7.

Having outlined some methods and issues of quantitative and qualitative attribution analysis, what follows is a discussion of style markers which are used in authorship analysis. Some of the style markers (also known as linguistic variables) discussed here are frequently used in quantitative analyses; others lend themselves to a more stylistic analysis. It would be a mistake to overstate a divide between qualitative and quantitative approaches, when in practice both can be used in combination.

2.3.7 Style Markers

It is important to review some of the desirable features of style markers. McMenamin divides stylistic choices into two types: variation and deviation. In McMenamin's terms, variation is the choice between two 'correct' possibilities, for example, a writer who chooses "twenty-six" rather than "26" or "can not" compared to "cannot". Deviation from a norm includes spelling mistakes or non-standard grammatical forms. He discusses the important features for a variable in some detail:

> The most important step for systematic observation in both the description and subsequent measurement of linguistic variation is the identification of the *linguistic variable*, that is, the isolation of structural or functional linguistic units that carry significance with respect to group or individual writing style (Labov 1966a). Preferred variables as articulated by Labov (1966b: 6) are those that are high in frequency (i.e., meaningfully quantifiable), immune from conscious suppression, codable, and widely distributed throughout a particular population. The variable is a class of variants ordered along a continuous dimension as determined by extralinguistic variables, such as particular individual authors, and is referred to as a style marker in authorship studies. (2020:543)

McMenamin's preferred quality of being "immune from conscious suppression" is interesting, because as McMenamin himself notes, "it is not yet possible to determine levels of conscious intervention as stylistic choices are made in the writing process" (2020:252). There seems to be very little research which investigates the extent to which writers are conscious of specific features of their language, and therefore, conscious of suppressing or adapting them to perform linguistic identity disguise, compared to how writers or speakers adapt their language for different situations. It is well-established that writers will adapt their style depending on register (Biber, 1988), depending on context, and as Bell (1984) argued, depending on their audience. The changes in linguistic features are well documented, but there is little research about how consciously each linguistic element is adapted. Function words are often referred to by forensic linguists as unconsciously chosen, but there is little research which investigates the extent to which this is the case, or to identify other features which might be more consciously chosen by a writer or speaker. It is also interesting to compare the relative levels of consciousness for open choices (for example, "big" versus "large"), and restricted features, such as function words. It may be that investigating identity disguise is a way of exploring how consciously writers use certain features, if it is assumed that the linguistic features which authors fail to suppress when writing different characters are those which they are least conscious of using.

Another desirable quality for a style marker is that of being 'unmarked'. Larner (2014:2) observes that many linguistic features analysed in forensic cases are the result of writers doing something unusual, or marked, in their language, whereas authors using standard grammar and spelling may produce a text without any such features. Larner emphasises the importance of finding style markers which occur simply through the process of textual production, rather than relying on the writer doing something marked, such as the deviations that McMenamin discussed.

Grant and Baker (2001) summarise the many textual features which have been suggested as style markers, for example, function words (as used by Mosteller and Wallace), vocabulary richness measures, word-type frequency distributions and content analysis. Like Grieve (2007), they caution against generalising that because *X* and *Y* features discriminate between one pair of authors, this will be true for more sets of authors, making *X* and *Y* "good" features of authorship analysis. Even between the same pairs of authors, different linguistic contexts may affect which textual features a writer uses, and subsequently, which style markers become more or less useful at discriminating between them. Grant refers to this as "understanding language variation stylistically, as the interaction between habit and context" (2020:562). Many papers use a "basket of features", and McMenamin

(2002) along with many other forensic linguists argues for the benefit of using a combination of features: "Linguistic individuation is virtually always established as a combination of traits rather than a single trait, because singular unique language forms are rare" (McMenamin, 2002:51).

McMenamin compares linguists who argue for a top-down approach, such as Sinclair and Coulthard (1975) and Edmonson (1981) to those who adopt a "bottom-up" approach, such as Labov (1972) and Schiffrin (1994). A top-down approach pre-selects a feature list in advance, then identifies occurrences in the text, whereas a bottom-up approach begins with the textual analysis and uses this process to select features. McMenamin argues that there are benefits in using the data to drive decisions about what to include:

> The identification of a set of style-markers that would discriminate all writers in a given speech community would of necessity be a top-down undertaking, i.e., style-markers would be predetermined a priori in other than an empirical way. Such an approach does not allow the data of each language sample to drive its analysis. (2002:63)

The top-down versus bottom-up approach is linked to the courts' required standards for expert witnesses, and a concern that there is a necessary subjectivity when an analyst selects the features for inclusion, no matter how skilled or accurate the analyst may be. Referring to the standards required by the courts for expert witnesses, McMenamin states that compromise is crucial:

> It will be necessary to find a middle ground between those who pre-select style markers for analysis, based on criteria established without reference to the instant writings, and those who hold that the style markers used for analysis of a particular set of writings must be first observed as possible linguistic variables in those very writings. (2020:554)

This issue becomes more complicated when considering different sub-disciplines of linguistics, such as Pragmatics, Conversation Analysis, or Discourse Analysis, where defining *a priori* features is likely to be a complex process in some of these higher-level language analyses.

This literature review has so far focused on structural elements of written language, the visible features on the page, rather than the function of language. McMenamin observes:

> Furthermore, language can be studied on the two complementary and inseparable planes of form and function. Form corresponds to the structure of language, defined as a linguistic system. Function relates to a focus on language use, defined as an integral part of human social interaction. (2002:2)

More recently, Grant and MacLeod (2018) have discussed using all levels of language analysis, from structural, up to pragmatics, semantics and interactional. Progressing from identifying and measuring words that are visible on the page to the more abstract concepts of language use presents its own challenges to the attributionist, such as reliability and replicability of coding. This idea is explored in the pragmatics analysis in Chapter 7.

2.4 Style

2.4.1 Introduction

Having discussed style markers, I now relate this to the underlying concept of *style,* in order to set out how I define *authorial style* and *character style* in my thesis, and discuss this in relation to genre. McMenamin (1993) discusses the idea of style as a composite of a writer's style markers, citing Enkvist, who proposes that a writer's style can be seen as "the aggregate of contextual probabilities of its linguistic items" (1964). If *style markers* are, in McMenamin's terms, the individual elements which combine to make a composite concept, *style*, it is important to consider the concept of style itself.

2.4.2 Defining Style

The term style is widely used, but has lacked a single, agreed definition in academic research (Crystal and Davy, 1969:11; Ohmann, 1964:423; Enkvist 1973:11; Leech and Short, 1981:11; McMenamin, 1993:145), a situation that was addressed by a number of key studies investigating style in English in the 1960s onwards, for example, Ohmann (1964); Crystal and Davy (1969), and Ullmann (1973). Ohmann criticised the study of style for being "remarkably unencumbered by theoretical insights" (1964:423), and Enkvist wrote that, "Style is a concept as common as it is elusive: most of us speak about it, even lovingly, though few of us are willing to say precisely what it means" (1973:11). Nearly thirty years later, Verdonk echoed this position, observing that "the term 'style' occurs so naturally and frequently that we are inclined to take it for granted without enquiring just what we might mean by it" (2002:3). Style is a broad concept, and can be used to refer to linguistic phenomena, but also to non-linguistic phenomena, such as interior design or 'style of management' (Verdonk, 2002:3), or a range of artistic qualities, such as music, arts and literature, or lifestyle choices such as cars and food (McMenamin, 1993:141).

Focusing on linguistic style, Crystal and Davy (1969:9) observe that whilst style is a familiar word, it has a multiplicity of definitions, and argue that it is necessary to specify what is meant by style in order to carry out stylistic analysis. They outline four commonly occurring uses of the term. Firstly, they describe style as a person's "language habits", for example, Shakespeare's style, stating that style, "refers in this way to a selection of language habits, the occasional linguistic idiosyncrasies which characterise an individual's uniqueness" (1969:9). Crystal and Davy's second use of style describes a class or group, such as the language of Augustan poets, which share common stylistic features. Thirdly, style is used in an evaluative sense, for example, having a "refined" style. Finally, the term style can be used to describe literary language, as a characteristic of effective or beautiful language. My thesis is concerned with the first and second of Crystal and Davy's definitions, which relate to authorial style and genre, and in addition I discuss the concept of *character style*. The third and fourth of Crystal and Davy's definitions of *style* in an aesthetically evaluative sense are less relevant here, because I am not analysing the literary merits of my data.

2.4.3 Authorial Style

Crystal and Davy's first definition of style as language habits relates to authorial style, suggesting that a writer's style is a composite of all their linguistic habits, similar to the "aggregate" of features discussed by McMenamin and by Enkvist. However, Crystal and Davy note that it is usually impossible to analyse all of a writer's linguistic habits, and is instead more practical to concentrate on "those features in a person's expression which are particularly unusual or original" (1969:10). This is echoed in Verdonk's theories on foregrounding, discussed below. Crystal and Davy give examples of linguistic features which are indicative of authorial style. The examples they give are at the structural level of language, such as "'pet' words or phrases" (p.66) which an author might use with high frequency. They propose a method of identifying style which is context-bound. They write, "It is by no means extravagant to conclude that an aspect or aspects of the context exercises some kind of conditioning influence on the feature in question, and the notion of *situation* has been set up to describe the kinds of conditioning influence" (1969:64). They outline a process for stylistic analysis, where the stylistician looks for reasons behind the use of a particular feature:

> The linguist, having intuitively noted a particular feature as being significant in some way, attempts to rationalise the basis of his intuitive response by examining the extra-linguistic context in order to establish any situational factors which might account for restrictions on its use. (1969:4)

Crystal and Davy break down the notion of 'situation' into 'dimensions of situational constraint', including categories such as 'dialect' and 'time'. They propose that the role every feature plays is described using one or more of the dimensions they list; for example, a particular feature could be seen as a result of a particular social relationship, and therefore could be referred to as a feature of 'status'. One dimension of style that Crystal and Davy list is "Individuality". They explain that, "In unselfconscious utterance, certain features occur – relatively permanent features of the speech or writing habits – which identify someone as a specific person, distinguish him from other users of the same language, or the same variety of the language" (1969:66). They use Individuality in a wide sense and distinguish it from "Singularity", which, instead of being a relatively permanent feature, only applies to "occasional idiosyncratic linguistic features" (1969:76). These ideas of authorial distinctiveness and the consistency of authorial style have been debated in stylistics. For example, Ullmann deems the popular image of a linguistic fingerprint as misleading, commenting that, "one's fingerprints do not change whereas one's style may do so; moreover, one cannot alter one's fingerprints but one can adjust one's style to suit the circumstances; one can even modify it for the purposes of pastiche, parody, or the need to portray a character through his or her speech" (1973:64). It is notable that Ullman describes this change as 'modifying' one's style, which suggests that while a writer may adapt some elements of their style to portray a character through dialogue, they will not alter it entirely.

To analyse authorial style, Crystal and Davy provide a list of sub-questions about a text: firstly starting with Individuality, “Does it tell us which specific person used it?”, before analysing other features such as Regional Dialect and Class Dialect: “does it tell us where in the country he is from? Does it tell us which social class he belongs to?” (pp.81-82). This is useful for the principle of distinguishing authorial individuality from genre features, or from features that might relate to an author’s sociolinguistic background (see 2.6.1 for a discussion on this topic). Applying this to authorship analysis in cases of identity disguise, it is possible that what might seem to be individuating for an author, is actually a stylistic feature of the genre rather than the individual.

2.4.4 Stylistic Choice

An important element of authorial style is the idea of stylistic choice. Leech and Short (1981) caution against “overdefinition” of style noting the many “unsuccessful attempts to attach a precise meaning” to the term (p.38). Instead, they list a number of principles which inform their understanding of style, including the idea that style is the way in which language is used, and that “style consists in choices made from the repertoire of language” (p.38). They state that, “Stylistic choice is limited to those aspects of linguistic choice which concern alternative ways of rendering the same subject matter” which they define as Style$_2$, distinct from their definition of Style$_1$, which is a more general notion of style as linguistic choice.

Continuing their examination of style as “alternative ways of rendering the same subject matter”, Leech and Short point out a shortcoming of Halliday’s approach, where “even choices which are clearly dictated by subject matter are part of style” (1981:34). They argue that this position does not work when applied to non-fictional language, citing the hypothetical example of a medical textbook replacing ‘clavicle’ with ‘collar-bone’, which could be considered a stylistic choice, whereas replacing it with ‘thigh-bone’, could not, because it has a different real-world referent. Leech and Short write that this approach can also be applied to fictional works. They state that, “the referential, truth-functional nature of language is not in abeyance in fiction: rather it is exploited in referring to, and thereby creating, a fictional universe, a mock-reality" (p.35). This is particularly relevant to the data in my thesis, because there are many elements of the fictional universe which have already been established in previous episodes, or have been dictated to the scriptwriter by the production team, and therefore cannot be viewed as elements of authorial choice, although how each writer chooses to dramatise a storyline does allow for freedom of choice.

The concept of style as alternative ways phrasing the same content was explored by Ohmann (1964), who provides an early discussion on style, arguing that readers have an instinctive understanding of authorial style, which he describes "a rather loosely structured, but often reliable, feeling for the quiddity of a writer's linguistic method, a sense of differences between stretches of literary discourse which are not differences in content" (1964:423). Ohmann conceives style as "a way of doing it" (p.426), and approaches the concept using a dualist approach, where the form (the 'way') and content ('it') are separate. He uses the analogies of a pianist and a tennis player to illustrate the form-content dichotomy: in each scenario, the player must follow a certain number of rules, whether those are notes and tempo, or hitting a ball over the net. Within the set parameters, each player has a significant amount of freedom to choose *how* they will execute the various notes and shots, analogous to their authorial style. However, as Ohmann argues, "the relevant division between fixed and variable components in literature is by no means so obvious. What is content, and what is form, or style?" (1964:427). Without the equivalent of a piece of sheet music to guide a conversation, there is no precise way of defining the *it* that is being discussed, in order to compare the *how* of stylistic choice.

In a forensic context, the idea of style as choice is important too. McMenamin writes:

> As a feature of written language, style is defined in at least five different but overlapping ways in current research: the writer's *choice* of optional forms, *deviation* from a norm, *idiosyncratic* distinguishing features, recurrent *habits*, and aggregate set of total possible contextual patterns. These five conceptual approaches to the definition of style are extremely important to the analyst of style because, individually and combined, they form the theoretical paradigm within which the linguist applies the practice of descriptive and quantitative stylistic analysis (1993:147).

These features echo theoretical foundations from non-forensic works on style, such as Ohmann (1964), and Crystal and Davy's concept of "Individuality". As McMenamin points out, the categories are overlapping, asking, "What is the difference between style-as-choice and style-as-deviation? Put another way, when does variation within a norm become deviation from that norm? (1993:151). Arguably, though, there is no need to classify a linguistic feature as a deviation or a variation, in order to measure it.

In terms of linguistic identity disguise, Ohmann's approach opens up analytical possibilities for studying multi-authored corpora. Adopting the position that style is the *how* things are done with words allows a comparison of how each writer carries out the storylines (the *it*) they have been issued. Ohmann demonstrates this process using transformative grammar, showing how sentences can be re-ordered or paraphrased, exploring the question of which variations are a stylistic variation, and which versions have changed the content. Whilst acknowledging the difficulties of extricating form and content, Ohmann argues that without treating the two as separable, one reaches the monist position, with its "altogether counterintuitive conclusion that there can be no such thing as style, or that style is simply a part of content" (p.427).

Leech and Short propose an alternative to the polarity of monism and dualism, in the form of "stylistic pluralism" (1981:30), which acknowledges that, "language performs a number of different functions, and any piece of language is likely to be the result of choices made on different functional levels" (p.30). Leech and Short describe the benefits of the pluralist position, writing that the pluralist "can show how choices of language are interrelated to one another within a network of functional choices" (pp.33-34). This was also a position adopted by Enkvist, who argued the study of style should not focus on just one linguistic item, but should extend to all levels (1964). This multilevel analysis of style is applied in Chapter 7, and its methodology is discussed in Chapter 4.

2.4.5 Authorial Style and Foregrounding

Echoing Crystal and Davy (1969) Verdonk explores the idea that some features of writing are more stylistically significant than others. He describes *style* as "distinctive linguistic expression" (2002:3), and therefore stylistics as, "the analysis of distinctive expression in language and the description of its purpose and effect." Verdonk states:

> In making a stylistic analysis we are not so much focused on every form and structure in a text, as on those which stand out in it. Such conspicuous elements hold a promise of stylistic relevance and thereby rouse the reader's interest or emotions. In stylistics this psychological effect is called foregrounding. (2002:6)

Verdonk discusses possible foregrounded elements, such as word choices, grammars or sentence structure, but, echoing McMenamin (1993), also notes that 'style markers' may be "deviations from the rules of language in general or from the style you expect in a particular text type or context" (p.6). Verdonk's suggestion that a stylistic analysis focuses on "conspicuous elements" which arouse interest is somewhat at odds with approaches from forensic linguistics, where the stylistic individuation is often found in those elements of language which are least consciously used, such as function words (McMenamin (2020), Stamatatos (2009)). This disparity is not surprising, because an authorship analyst is studying the text for an entirely different purpose. However, it does suggest that those foregrounded features which "rouse the reader's interest or emotions", as Verdonk describes, are not necessarily those which are most individuating. I explore this question further in Chapter 6, where I compare the foregrounded elements of three characters, to analyse whether they can be used to discriminate between the way the different writers create stylistically distinctive characters.

2.4.6 Character Style

In my thesis, I refer to the linguistic traits of individual characters as *character style*, and discuss identifying features of 'authorial style' or features of 'character style'. However, this is a problematic distinction, because in drama the characters' dialogue is not separate from the authorial voice, but is part of it. For example, although Beatrice and Hero in *Much Ado About Nothing* have very different personality traits, they are arguably both recognisable as Shakespearean characters. Likewise, Burton

(1980) observes that people might listen to a discussion which flouts conversational norms, "and remark that it is 'Pinteresque'" (1980:14). Clearly this is not to suggest that Pinter himself would speak this way, but instead suggests that authorial style in dramas can be inferred *through* character dialogue.

A useful way to view this is provided by Short (1989:149), who describes the message of Addresser 1 (playwright) to Addressee 1 (audience / reader) as being conveyed through the message between Addresser 2 (Character A) and Addressee 2 (Character B) (Figure 1). The relationship between authorial style and character style is not mutually exclusive one, where a line of dialogue could be either in the style of the author *or* in the style of a character: one is produced through the other. Short writes:

> The important thing to notice is the general *embedded* nature of drama, because features which, for example, mark social relations between two people at the character level become messages *about* the characters at the level of discourse which pertain between author and reader/audience. (1989:149)

Character style is a problematic term because the individual characters can be analysed at the level of the talk they produce between themselves, but at the same time, as Short argues, they are also conveying a message from the writer to the audience about the drama itself. Short's model is reproduced below (Figure 1).

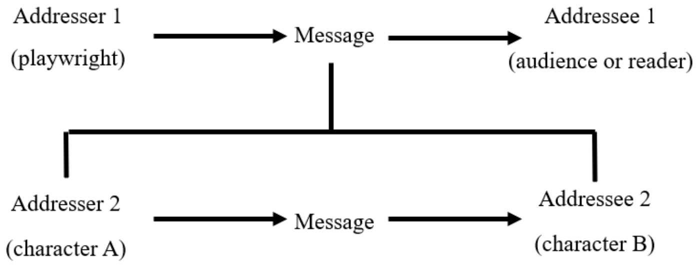


Figure 1: Short's Discourse Structure of Drama (1989:149 Fig. 8.2)

In my thesis, my focus is on the way the scriptwriters create characters' voices, rather than a literary appraisal of the scripts as a work of fiction, so in Short's terms, I am focusing on the level of the message between Character A and Character B, not the message from Playwright to Audience.

Whilst acknowledging the complicated relationship between authorial style and character style, I use *authorial style* to mean those linguistic features which are used consistently by an author regardless of which character they are writing, and by *character style*, I am referring to linguistic features which can be associated with particular characters, in their character-to-character communications in the fictional world. The term *character style* is something of a shorthand, because the character's style is, necessarily, part of the author's style, but when I use the term *character style*, I am referring to those linguistic traits which are particularly suggestive of an individual character.

2.5 Linguistic Identity Disguise

The impact of linguistic identity disguise on attribution results has been discussed. What follows is a review of literature on the different types of linguistic identity disguise, and the various approaches to authorship attribution in case of linguistic disguise. The first section considers the linguistic levels at which identity disguise can take place.

2.5.1 Levels of Linguistic Analysis

This next section begins with a brief discussion of scholarship concerned with the different domains of language at which identity disguise can operate. Grant and MacLeod (2020) state their understanding of linguistic identity as applying to four domains of language. They present these domains in a table (2020:38), which is an adaptation of the levels of linguistic analysis set out in Herring (2004:18). An abridged version of this table is presented below (Table 1).

Table 1: Levels of Language Analysis

Domains	**Methods**
Structure	Structural / Descriptive Linguistics, Text Analysis
Meaning	Semantics, Pragmatics
Interaction	Conversation Analysis, Ethnomethodology
Social Behaviour	Interactional Sociolinguistics, Critical Discourse Analysis

(abridged from Grant and MacLeod 2020:38)

Herring had included an extra possible layer, multimodality, which, due to the nature of these data, is not considered in this analysis. Grant and MacLeod advocate using all four domains of language to explore the creation of linguistic identities:

> We see identity as a phenomenon best classified at the level of social behaviour, but it must be kept in mind that the identities projected by individuals are produced with the resources available to them at all three of the other linguistic levels. (2020:39)

There is very little pragmatics research into CMD, and even less on pragmatic authorship analyses of CMD texts. Grant and MacLeod (2016) observe that in the field of identity construction, "the exploitation of pragmatic and interactional resources is comparatively under researched". Discussing the linguistic analysis of online communications, Grant and MacLeod reference Barron (2013), who "laments the scarcity of research viewing IM through a pragmatic lens, given the obvious benefits of a pragmatic approach for conceptualising and understanding IM" (2020:40). Grant and MacLeod (2016) created experimental data, asking students and undercover officers to participate in a chatroom conversation and attempt to assume the identity of another participant, with varying levels of preparation. Meanwhile another participant, in a different location, acted as judge and was tasked with identifying when the switch from interlocutor to impersonator occurred. The participants' language was coded at a number of linguistic levels, following Herring (2004), including structural features, such as lexical choice, up to the levels of pragmatics, semantics and social interactions. Grant and MacLeod found markedly different patterns at the pragmatic and interactional levels of language used by participants. They found there was:

> Variation at the pragmatic and interactional levels, as well as at the structural level. This can be observed in differences in choices and placement of speech acts, for example, and in choices not only of topic but manner of topic introduction – for example choice of discourse marker – and topic decline, for example indirectness or avoidance. (2018:64-65)

They found that untrained participants who were attempting identity disguise tended to focus "rather simplistically on the structural level, mimicking spelling, capitalisation, abbreviations and punctuation patterns", but that judges showed awareness of changes in language use at the higher domains of pragmatics and social interaction between the first interactant and the imposter. Judges were in some cases unable to fully articulate the pragmatic differences, but were instinctively aware of changes at the pragmatics level of language analysis. Grant and MacLeod also observed that linguistic leakage by the impersonator occurred at the pragmatic level, but once participants received training, they were able to improve their identity disguise in this domain (2020:105-7).

Scriptwriters do not receive linguistic training, but feedback notes from script editors often address identity at the levels of pragmatics and social interaction, for example, "why is Alice being so rude?" (*constructed*), so it is possible that the feedback scriptwriters receive has formed an informal training in honing characterisation at the levels of pragmatics and social interaction which is realised through linguistic choice, even though linguistics, as such, is not a focus.

Following on from the structural level analysis in Chapter 5, there are three remaining tiers from Herring's table, including the Meaning level, which encompasses both Pragmatics and Semantics, and is explored in Chapter 7 of my thesis. Culpeper and Haugh, introducing the field of Pragmatics, write that pragmatic meaning is "what the speaker means by an utterance and what the hearer understands by it (which could, of course, be two different things), and how these emerge and are shaped during interaction" (2014:5). The aim of my pragmatic analysis is to explore the differences between scriptwriters' contextual meanings of the same words, so that the discrimination between writers is found at the pragmatic level, rather than a structural level analysis of token frequency.

As Archer et al. (2012) observe, there is an expected overlap between semantics and pragmatics:

> Semantics in the narrow sense is concerned with the kind of meaning which belongs in truth-conditional semantics, and pragmatics with other types of meaning (e.g. the type of inferences we can make from what is said … However, the boundary between semantics and pragmatics may well be fuzzy since lexical elements are constantly drawn into the pragmatic sphere by means of changes associated with grammaticalization and particularly 'pragmaticalization'. (2012:4)

The remaining two tiers of Herring's table, Interaction and Social Behaviour, seemed less promising routes for exploration of these data for a number of reasons. At the Interaction level, analyses such as Turn Construction Units, speaker selection and Repair are problematic areas to analyse in a fictional drama because the drama is scripted, so the characters are not spontaneously negotiating their own part in the conversation, and one writer controls all speakers. The ultimate aim of this research is to explore linguistic features which could later be applied to forensic data, so researching how one author manages the topic control of multiple characters is stylistically interesting, but is less applicable to forensic settings, such as identity disguise in online fora. Ethnomethodology and Social Behaviour are likewise less obvious approaches because of the fictional nature of the data.

An alternative perspective to Herring's hierarchical structure is outlined by Schneider and Barron, (2010), who caution against viewing pragmatics as a level to be added to phonology, morphology, syntax and semantics. Instead they adopt a 'complementary view', "based on Leech's conceptualisation of 'grammar' (i.e. the language system) and 'pragmatics' (i.e. language use) as two complementary and interacting domains" (2010:240). Whichever way the interrelation between pragmatics and structure or grammar is viewed, both positions suggest that adding a pragmatic approach to a quantitative analysis of 'structure' can broaden the tools available for authorship analysis.

2.5.2 Authorship Synthesis

The first type of identity disguise considered here is that of *authorship synthesis*. MacLeod describes a scenario where this task becomes necessary:

> A caregiver discovers a child has been taking part in sexualized Instant Messaging (IM) conversations with an adult online. The police are alerted, and the victim is removed to a place of safety. An undercover officer (UCO) takes her place, engaging the adult in IM conversation in an attempt to set up a meeting to secure an arrest on suspicion of grooming under the Sexual Offences Act 2003 ... The UCO must synthesise – that is, construct from available resources – the victim's identity. (2020:159)

UCOs often have a limited amount of time to analyse and adopt the linguistic persona of their victim, and as MacLeod argues, must simultaneously suppress their own linguistic traits (2020:159). They must also absorb the factual knowledge about the target's life and opinions, described by McMenamin (2020) as a "resemblance model" (discussed in section 4.2.2). Compared to fictional characterisation, where the audience is generally willing to engage in a suspension of disbelief, online chatrooms are a low-trust environment where suspicions about interactants' identities arise quickly. One observation pertinent to scriptwriters is that identity assumption in this scenario needs to be targeted to that particular individual, and "cannot be achieved through performance of linguistic stereotypes of a particular group, or any other simplistic categorisation of a social taxonomy" (2016:60). Whereas, to an extent, figures in a drama are created by a combination of top-down schematic knowledge and bottom-up textual cues (Culpeper, 2001), authorship synthesis in an undercover operation arguably has a closer focus on textual clues than schematic knowledge.

Grant and MacLeod (2016) found that participants with lower levels of linguistic training tended to attempt authorship synthesis by focusing on the structural level, whilst ignoring higher level features such as turn-taking and topic management. This significant experiment helps explain the reason why computational methods are so vulnerable to obfuscation attacks: as discussed, quantitative approaches to authorship attribution necessarily focus on observable, measurable features, which tend to occur at the structural level of language. These are the same features that obfuscators are aware of, and adapted when carrying out linguistic identity disguise.

Computational approaches tend not to analyse language at the levels of pragmatics, stylistics and interaction, as these features are not easily quantifiable or extracted automatically. Interestingly, Grant and MacLeod found that although these features were ignored or under-utilised by the untrained impersonators, they were noticed by the judges, even those with less linguistic training, who were aware of the change in higher-level language features:

> Judges also notice failures of identity assumption which can be described through higher levels of linguistic analysis, notably at the level of interaction – 69% of the judgements in the Undergraduate group mention timing or message length as having led them to their decision of when a switch occurred. (2016:62)

This has parallels to multi-authored scriptwriting, where audiences may instinctively feel that an individual scriptwriter has not quite captured what they believe to be the voice of a character, even if it is not possible to pin down exactly why. This experimental approach strongly suggests the importance of considering different levels of language when analysing linguistic identity disguise and the construction of linguistic personae.

2.5.3 Adversarial Stylometry

Juola and Vescovi pick up on the problematic metaphor of the linguistic fingerprint when they write, “just as criminals may wear gloves to hide their fingerprints, so too may criminal authors mask their writing styles to escape detection” (2010: 115). Yet, despite this possibility of stylistic disguise, they note that “most authorship studies have focused on cooperative and/or unaware authors who do not take such precautions” (2010:115). For example, Baayen et al. comment that, “We interpret our results as supporting the hypothesis that authors have ‘textual fingerprints’, at least for texts *produced by authors who are not consciously changing their style of writing across texts*” (2002:69, emphasis added). Baayen et al.’s comments are not unusual in testing attribution methods on texts where the writer has not deliberately altered their writing style.

Researchers investigating the problem of attribution in cases of linguistic identity disguise have termed it Adversarial Stylometry. Brennan and Greenstadt (2009) divide Adversarial Stylometry into two main areas: “obfuscation, where a subject attempts to hide her identity, and imitation, where a subject attempts to frame another subject by imitating his writing style” (2009:60). Whilst imitation is fairly self-explanatory; obfuscation can involve various techniques such as running a text through multiple translator tools, then back into its original language, with the aim of reducing the number of identifiable stylometric features (Caliskan and Greenstadt, 2012).

An important question is why might authors disguise their style, and also, what effect does this have on stylometric techniques? Overdorf and Greenstadt approach this issue with the privacy of the individual in mind, rather than the investigative requirements of law enforcers. They write:

> The field [Stylometry] is inherently linked to privacy and security research as the use of it can provide or deprive users of anonymity online. The more robust stylometric methods become, the greater their threat to privacy. (2016:155)

To illustrate the issue of threats to privacy, Brennan et al. (2012) provide the example of an employee who leaks confidential information anonymously, but can be identified using a linguistic analysis which is compared to her social media account. In contrast, forensic linguists (e.g. Grant and McLeod, 2016), have written about the need for law enforcement officials to be able to negotiate deceptive or disguised writing styles, describing the "serious social problem" (2016:50) of online grooming, where false identities are used online, both by the perpetrators, and by undercover police officers attempting to gain criminal evidence. They observe that, "the issue of identity and influence within transnational online communities has become a significant social and policing concern" (2016:51).

From both perspectives – preserving privacy, and identifying and apprehending criminals – authorship analysis in cases of identity disguise is an important social and forensic area of research. The effect of identity disguise on stylometric techniques has been investigated, mainly from within the field of computational linguistics. Following on from these two perspectives, it is worth noting that there are two ways to evaluate "success" in adversarial stylometry. One is to measure the success of the method of stylometry in finding the correct author despite the disguise; the other is to measure the success of the writing in preserving its author's anonymity.

To explore the effects of adversarial stylometry on authorship attribution, Brennan et al. tested authorship attribution techniques on texts where the writers had disguised their style, and then on texts where they had not:

> We demonstrate the effectiveness of multiple methods of stylometry in nonadversarial settings and show that authors attempting to modify their writing style can reduce the accuracy of these methods from over 95% to the level of random chance. (2012:12.2-12.3)

While stylometry techniques can identify authors with high accuracy in non-adversarial scenarios, their accuracy is reduced to random guessing when faced with authors who intentionally obfuscate their writing style or attempt to imitate that of another author. One explanation might be that the lack of linguistic theory underpinning computational analyses makes the process more vulnerable to adversarial attacks than linguistically grounded, observational methods.

Clearly, linguistic disguise is a key issue for authorship analysis tasks, especially as much forensic casework is based on police suspicion that an author has attempted to impersonate or imitate another person in order to conceal a crime (such as the Amanda Birks case described in Grant, 2020). Although these computational techniques have not proved robust at circumventing adversarial attacks, it is worth noting that there is a strong track record in manual detection of deception within forensic linguistics (e.g., Grant, 2010, Coulthard, 2004, 2010), using close analysis of smaller datasets. Whilst such pairwise comparisons are not so drastically affected as largescale statistical approaches, it is still important to discover more about the nature of adapting writing style.

Brennan et al. (2012) carry out three experiments to study the effects on stylometry of obfuscation and imitation attacks, and a machine-translation obfuscation attempt. In the manual obfuscation attack, writers attempted to hide their identity when writing a short passage describing their neighbourhood. For imitation, they wrote a short article in the style of Cormac McCarthy's *The Road*. These were measured against a control sample of participants' pre-existing writing samples. Thirdly, machine translation tools were used to test the effect on stylometry. They found that imitation texts were very successful in having the author attributed to the victim of the imitation attack.

Obfuscation attempts rendered the stylometric methods used no better than random chance, and Brennan et al. noted in their conclusion that no great skill was required by the obfuscators: "the attacks were generated by participants in very short periods of time with no expert knowledge in linguistics or stylometry" (2012). While the study considered text samples from a much higher number of people than previous 'literary detection' papers have, there were some limitations on the data. Participants were asked to use formal writing, in order to avoid slang. In practice, many cases of possible deception are quite likely to include informal, conversational pieces. It would also be interesting to see these techniques applied to dialogue, since computer-mediated communication is a vast area, with clear potential for people to disguise their identities.

In a more in-depth paper on machine-translation as obfuscation, Greenstadt and Caliskan (2012) build on Rao and Rohatgi's (2000) idea of translating a text to a different language and back again, to obfuscate authorship. They tested which features were preserved through this process (which included top letter trigrams and words), and which were preserved less well (function words, letters and word length). They concluded that machine-translated tools introduce an effect on the translated text, but that the translator tool can often be identified, thus undermining the obfuscation. They also found, as would be expected, that the more times a text was translated, the less-preserved the original stylometric features become.

Brennan and Greenstadt (2009) conclude that although the attribution techniques do not perform well in cases of obfuscated or imitative writing, this does not render stylometric techniques useless; rather that techniques need to be developed or adapted to cope with cases of adversarial attacks. Meanwhile, Juola and Vescovi (2010) noticed certain features, such as function word usage, which shifted when the writing was deemed deceptive, and also found that "Character-based events (bigrams and trigrams in the analysis) appear to be more robust to obfuscative attacks than word-based events" (2010:121). However, they "were unable to find a "silver bullet" that reliably solves the hostile author problem" (2010:121).

A further element to be considered is the duration of the identity disguise or deception. Afroz et al. compare shared linguistic features between fiction and long-term deception (2012:471-2), and found differences between long-term and short-term deceptions, comparing a blog written by a 40-year-old man pretending to be a Syrian woman over a number of months, to entrants from a Cormac McCarthy imitation contest where each entry was only 500 words. Afroz et al., "found these deceptions to be more robust to our classifier but more vulnerable to traditional stylometry techniques" (2012:463). Day et al. comment that, "Recent research has shown that adversarial stylometry is not effective in concealing one's writing style over the long term" (2016:1). This is an interesting contrast to a long-running drama, where the opposite result might be expected; that a scriptwriter will improve their craft with experience, and will therefore be more successful at writing convincing character dialogue.

2.5.4 Deception Detection

Although Adversarial Stylometry is a relatively new area, there is a longer tradition of research into deceptive writing, where the content, rather than the style, is deceptive. Whilst this is a much broader area, and beyond the scope of this study, its links to adversarial stylometry are discussed by researchers exploring obfuscation. Afroz et al. explore whether it is possible to detect the presence of stylistic deception itself in documents, questioning which linguistic features indicate stylistic deception. They state that although "stylistic deception is not lying, similar linguistic features change in this form of deception" (2009:462). They ask which features are likely to change and which remain constant in adversarial attacks. They also question whether stylistic detection shares characteristics with other deceptions. Using linguistic and contextual features they were able to distinguish between stylistic deception and regular writing with 96.6% accuracy and identify with 87% accuracy whether a deception was obfuscation or imitation (2012:462). Afroz et al. concluded that some linguistic features change when people hide their writing style, and by identifying those features, it was possible to detect deceptive documents (2012:462), though not the identity of the author. They found function words to be best stylistic marker for this analysis, and showed that 'deceptive' writing could be detected with 'high accuracy' if a large feature set was used (2012:473). Writing dialogue in a drama is a kind of identity disguise, but unlike deceptive writing, is done in the understanding that the audience is aware of the fictional nature of the writing. However, the psychological process of writing outside of one's own experience may have some cognitive processes in common with the process of deceptive writing.

2.5.5 Literary Pastiche

Whilst adversarial stylometry is relatively new, identifying the author of anonymous or misattributed literary works has provided much of the basis of early authorship analysis, such as nineteenth-century speculation that Shakespeare's plays were written by Francis Bacon, and more recently, Foster's (2000) attribution of an anonymous poem to Shakespeare, although this was later discredited. Much of this research involved a literary stylistic analysis using qualitative methods, with no standardised methodology or criteria for a definition of success. Another way in which literature has formed the basis of authorship studies is through the examination of literary pastiche, where an author writes in the style of a well-known work of fiction, usually with the intention of honouring and flattering the original work. Pastiche is distinct from parody, in which aspects of the original are exaggerated for comic effect.

Somers and Tweedie (2003) examine literary pastiches to identify whether a text is an original or pastiche. They note that in distinguishing between literary pastiches and their originals, they "have not explored whether the comparative nature of these numerical results reflects the subjective ratings of critics" (2003:424). The same study analyses linguistic imitation through analysing Gilbert Adair's pastiche of *Alice's Adventures in Wonderland*. They found that standard measures of lexical richness, Yule's K and Orlov's Z, could distinguish between a Lewis Carroll novel and Gilbert Adair's pastiche. Somers and Tweedie describe Hilston and Holmes' 1993 findings that two very dissimilar pieces of work by Ian Fleming (*James Bond* novels and a children's story) showed more similarity than between Fleming's Bond novels and a Bond story written by Kingsley Amis (2003:410).

Somers and Tweedie observe that there is a "no-win" paradox for authorship attribution studies regarding the case of imitation and pastiche:

> Overall, our results send a mixed message regarding authorship attribution techniques and pastiche: if the technique succeeds in distinguishing the pastiche … can we point to this as support for the robustness of the techniques … In a similar manner, if the technique cannot distinguish the pastiche and the original … do we say that it is a measure of the pastiche writer's skill … or does the result cast doubt on the technique? (2003:423)

Literary authorship attribution has tended to focus on comparing a questioned document or pastiche to known originals. A possible response to this is to make a comparison between various pieces of imitative writing, (a multi-authored drama series), rather than comparing the attributes of one pastiche to its original. The result is not simply about whether or not the pastiche 'achieves' its aim but is a more relative comparison of which features change, and in what proportion.

There are various forms of literary imitation. One type is a continuation novel, when an author writes a new book in a literary series, usually after the original author's death. For example, new *James Bond* novels, in the style of Ian Fleming, have been written by Sebastian Faulks, Jeffery Deaver, Anthony Horowitz and Kingsley Amis (writing as Robert Markham). These were official continuations, commissioned by Ian Fleming's literary estate. Continuation stories can also be unofficial, as in the case of J.M. Barrie, who wrote subsequent *Sherlock Holmes* stories, in honour of his late friend Arthur Conan Doyle. Such stories overlap with other forms of imitation, for example "fan fiction", where the imitating authors (generally amateur fans, rather than professional writers) may change or subvert the characters and the original fictional world.

In long-running dramas, there is a comparable form of imitation, where a scriptwriter creates a new programme, and, if the programme is successful, other writers are commissioned to write subsequent episodes, often with the first scriptwriter acting as 'head writer' or maintaining some kind of creative lead in the process. Again, there is an expectation of using the original series and characters as a definitive version, from which the subsequent series are derived.

In the case of *The Archers* and other continuing dramas (e.g., soap operas or very long-running drama series, such as the BBC One hospital drama *Casualty*), the manner of imitation is somewhat different, in that characters and storylines are created more collaboratively, rather than having a definitive version of each character on which to base their imitation, although character consistency remains important.

In forensic linguistic terms, these fictional 'imitations' could be cautiously compared to identity assumption (Grant and MacLeod, 2020), and authorship synthesis (2018). MacLeod (2020:159) describes the forensic setting of undercover police officers who are tasked with linguistically impersonating an online victim in order to communicate with a suspected abuser. They have to "construct from available resources – the victim's identity" (2020:159). This comparison is tentative for a number of reasons – primarily the fictional quality of the first set of data. Secondly, in the fictional examples, the writers are attempting *character* synthesis, whereas in Grant and

MacLeod's cases, the writers are attempting *authorial* synthesis. Whilst the fictional characters are expected to have credibly consistent behavioural patterns, it is not necessary for the authorial voice to be entirely anonymised, as *Archers* scriptwriter Mary Cutler observed (Section 1.5).

For undercover officers to achieve authorship synthesis, and for scriptwriters to create new scenes for existing characters, there is a need for consistency in the way the characters think, behave and speak, so that the interactant or audience believe in the character. However, in a drama, there is also an authorial voice which operates simultaneously to these character performances. This relates to the structure of dramatic discourse discussed in Section 2.4.6. Just as Wallis and Shepherd (2002) and Short (1996) among others describe the way that plays produce a direct message from writer to audience, which is communicated through the character-to-character message, so Mary Cutler describes how her own authorial voice was recognised by listeners, existing simultaneously with the voices of the characters. Another point of comparison between fictional and forensic is in the multiple authorial process. MacLeod notes that, "it may be the case that UCOs will have to provide operational cover for one another – multiple officers may be required to operate as one specific offender or victim within an operation due to changing shift patterns, illness or leave. Similarly, a single officer may be involved in multiple concurrent operations" (2020:161). This multi-authorial approach has strong links to multi-authored drama scripts. There is also a point of comparison between the undercover officer, who has an ultimate aim of progressing the process of apprehending a suspect, and the scriptwriter who has an ultimate aim of progressing the plot: both authors have to use dialogue to achieve a goal which guides the interaction.

2.6 Identity Performance

2.6.1 Three Waves of Sociolinguistic Identity

In order to discuss linguistic identity disguise, it is important to consider the nature of identity and language, and various approaches to sociolinguistic analysis. The study of social meaning in sociolinguistic variation has been conceptualised by Eckert as three waves of analytic practice. Eckert explains that the waves are not separate eras, but rather they operate as part of a whole, and no single wave supersedes the others. Each wave "represents a way of thinking about variation and a methodological and analytic practice that grew out of the findings of the previous one" (2005). Eckert describes how the first wave, exemplified by Labov's New York study (1966), explored correlations between linguistic variables and broad demographic categories, such as ethnicity and gender. The second wave, including studies such as Milroy and Milroy (1978), and Milroy (1980), focuses more on ethnographic practices, and explores social networks and variation within local communities. Eckert (2012) sets out a theoretical Third Wave, in which linguistic variation is perceived not as a reflection of social categories, but instead is part of a social practice in which the speakers engage. Reviewing third wave studies, Eckert observes that, "work in the third wave is diverse, but it always involves a focus on style and social meaning" (2021:382).

The first wave took an essentialist approach, correlating linguistic variation with features of social class. Eckert describes how first-wave studies "established the broad patterns of social stratification of variation across the primary macrosociological categories of class, age, gender, and ethnicity" (2012:87). Key studies are Labov (1966), Wolfram (1969), and Trudgill (1974). Labov's highly-influential study of rhoticity in three Manhattan department stores, one budget store, one middle-ranking store, and one exclusive store, found that *r*-pronunciation increased as formality increased, observing that shop staff in Klein, the lower-end department store, pronounced the 'r' sound to a much lesser extent in the phrase "fou*r*th floo*r*" than their counterparts in the middle and highest-ranking department stores. Labov found that *r*-pronunciation increased on a second use, in response to an "Excuse me?", and the rhoticity was also more notable in the word 'floor' than 'fourth'. These nuances are relevant to my study: a specific linguistic feature may be more or less prominent in different context or place in a conversation, just as Labov found rhoticity increased on a second usage. It is possible that a linguistic impersonator might identify linguistic variables used by a target (for example, identifying the level of rhoticity), and then apply this over-consistently, rather than echoing the natural variations of the target's linguistic usage.

Wolfram (1969) carried out a similar study, analysing linguistic variation within ethnicity, separated by social class, investigating a number of variables, including pronunciation of final consonant clusters, such as “test”, post-vocalic ‘r’ pronunciation, multiple negation and copula absence. Wolfram found a correlation between linguistic features and social class; for example multiple negation was used by upper middle-class speakers on approximately two per cent of possible occasions, which rose to a figure of 70 per cent for lower working-class speakers. Reviewing this study, Wardhaugh and Fuller state the importance of nuance here, arguing against a mistaken assumption that some features are always used by some demographics, and never by others. They point out that, “No class uses one variant of the variable to the exclusion of the other, regardless of circumstances. Speech within any social class, therefore, is inherently variable, just as it is in society as a whole” (2014:177). This is relevant to analysing identity assumption, suggesting that it is important to assume variability among the characters, but not necessarily a binary presence or absence of linguistic features.

Trudgill’s (1974) analysis of linguistic variation in Norwich is another key first-wave study. Trudgill analysed sixteen phonological variables including the “ng”, “t” and “h” sounds. Trudgill, like Labov, recorded different results depending on where in the word a particular sound was placed, but also based on the level of formality in which it was used; for example when reading a word list, compared to casual conversation. Further, Trudgill’s results indicate that female speakers are more likely to adhere to the ‘standard’ variant. Again, these nuances are relevant for linguistic impersonation because variation may be affected by a number of factors, such as by social class and then by gender.

Eckert observes that the second wave, marked by studies such as Milroy (1980), Rickford (1986) and Eckert (1989), was ethnographically focused, exploring social networks and local categorisation: “The second wave began with the attribution of social agency to the use of vernacular as well as standard features and a focus on the vernacular as an expression of local or class identity” (2012:91). Milroy and Milroy (1978) compared language variation in three communities in Belfast with differing levels of employment, and showed a correlation between closeness of community and the use of “vernacular norms”. The results supported Milroy’s hypothesis that close-knit social networks reinforce norms, including linguistic norms. Milroy also observed that, “a closeknit network structure appears to be very common . . . in low status communities" (1980:43), and related this to participants valuing the closeness of the social network, rather than an aspiration to the 'standard', more prestigious varieties of the language. According to Milroy, the vernacular norms are "perceived as symbolizing values of solidarity and reciprocity rather than status, and are not publicly codified or recognized" (1980, 35–6).

Another highly influential second wave study is Eckert's work (1989) on Burnouts and Jocks in the pseudonymous "Belton High School" in Detroit: the 'Jocks' are broadly middle-class students, and 'Burnouts' are urban students of lower socio-economic status. Eckert observed that the differences between the two groups were "deeply ideological" and both groups used a variety of linguistic and non-linguistic strategies, including clothes, hair, make-up, but also vowel sounds, to mark their differences. Eckert, (2016), reflects on the key findings from this study:

> Most important in this study was the fact that the phonological variables correlated with social category affiliation and not with parents' socioeconomic class. Thus one cannot say that variation reflects passive social address or a system acquired at home during childhood. Rather, participation in these sound changes emerged as part of the speakers' participation in the peer-based social order as they constructed an adolescent identity (2016:10).

These observations identify key second wave concerns for the importance of social networks and communities of practice which account for social meaning beyond essentialist categories such as race, age and gender. In drama scripts, this demarcation of social difference becomes significant because there is a need for characters to be distinct from each other, both for dramatic interest, but also (especially in an audio-only format) for clarity. However, based on these studies, audiences might also expect vernacular norms to occur between characters in common social networks, and at a whole-text level, we might also expect to see vernacular norms as part of a script's style. Eckert herself (2021) emphasises the importance of style in the second-wave studies, observing that, "At this local level, style took on a new significance, as it became apparent that variables combined into styles that articulate social differences among and within categories" (p.383). Interestingly, Eckert herself identifies "Belton High School" as an ethnographic study which is not a key second wave study, but instead, "represents the transition from the Second to the Third Wave" (2016:4).

Introducing the idea of the third wave, Eckert writes: "The Third Wave is based on the understanding that language is not just structure but practice. Change is basic to human life, and as part of social practice, language must be dynamic at its core. Language does not just happen to change – language is change" (2016:11). This constitutive view of language emphasises the stylistic agency of individual speakers and groups to perform identity, and also emphasises the fluidity of speaker performance over time. Comparing the three analytical waves, Eckert writes: "The First Wave viewed variables as indexing the speaker's membership in macrosocial categories, a view that the Second Wave challenged implicitly and that the Third Wave challenges explicitly" (2016:3). She argues that "Indexical activity, on the other hand, is local and specific, and it is at the local level that we produce and recognize the social" (p.3). This becomes interesting in a drama script, where a writer needs to convey information about a character (age, socioeconomic class and so on), and needs to create characters who are credibly from the background set out in the fictional universe, but for major characters, is also likely to wish to create individuality and move beyond simply portraying stereotypical features of macrosocial categories. Hall-Lew et al. (2021:5) state that "indexicality is central to third-wave research". They describe the process of indexicality:

> While all linguistic forms have the potential to signify social meaning, a form only does so when our system of ideas and beliefs creates a link between the form and a type of social meaning (such as stance, persona or social type … At its core, indexicality is a process of association, where a linguistic form points to some dimension of its conventional context of use (2021:5).

Johnstone examined this concept through her research into Pittsburghese (2013), investigating the phenomenon of Pittsburghese, though dialect items such as "yinz". Johnstone traces the links between linguistic choice and social meaning, exploring diachronic change in the Pittsburgh dialect, but also the reasons why Pittsburghese remains important to identity. Johnstone et al. (2006) describe how a set of linguistic features came to be associated with Pittsburgh, and gradually became "enregistered" (Agha, 2003) as the Pittsburghese dialect. Drawing on Silverstein's (1976/1995, 2003) "order of indexicality", they describe how:

> "First-order" correlations between demographic identities and linguistic usages … came to be available for "second-order" sociolinguistic "marking" … of class and place, and then how certain of these indexical relations between linguistic forms and social meanings became

> resources for the “third-order” indexical use of sociolinguistic “stereotypes” … in more reflexive identity work.” (p.78)

Johnstone’s position that sociolinguistic identity is a reflexive activity could be applied to the writing process for fictional writers: they are performing an identity, both at the character level, and at a genre level. When writers (re-)create established characters, arguably they are not necessarily inhabiting every aspect of a character’s dialogue, but could instead, be said to index certain aspects, in particular, any more clearly marked aspects of a character’s sociolinguistic practice, such as a particular regional dialect, or notably high, or low, socioeconomic class. Halle-Lew et al. note that “in this model of indexicality, the social meanings identified are almost exclusively related to the main correlations measured in first- and second-wave research (e.g., persona types, such as “Pittsburghers;, or social types, such as ‘working class’.” (p.6). The idea that fictional characters, as well as writers, could be performing certain personae, and need a degree of flexibility, is an interesting one, and relevant to my study.

The ideas summarised here emphasise the importance of flexibility and indexicality to sociolinguistic identity. If characters in a long-running serial are to develop beyond stereotypes demarcated by their macrosocial categories, and continue to interest their audience, we might expect – perhaps on an intuitive level – that their sociolinguistic behaviours will reflect the flexibility and indexicality found in their real-life counterparts. However, as discussed in 8.8, there is also a burden of credibility for writers of fiction: sociolinguistic performativity and indexicality might make a fictional character more complex and interesting, but a character whose identity performance is too varied might be viewed as lacking in cohesion and credibility.

2.6.2 Identity Construction

As Joseph (2004) explains, “It has been argued that language is the most flexible and pervasive resource available for identity production, and that language and identity are ‘ultimately inseparable’ (Joseph 2004:13, in Grant & MacLeod, 2016:52). This is particularly true in online chatrooms and text messages, and indeed, in audio-only formats such as radio and podcasts when there are no visual clues to identity. Grant and MacLeod acknowledge Herring's (2004) position that "owing to the lack of a physical context, language is at its most performative in online contexts". The same can be said of radio drama: as a non-visual medium, characterisation is created through linguistic choice and vocal performance. The way that identity is perceived has important impacts for the ways authorial identity is analysed.

Much research into authorship profiling and authorship analysis been guided by principles of sociolect (linguistic varieties associated with particular groups, such as age), and idiolect (one's individual use of a language). Inherent in the concept of authorship profiling is the idea that predictions can be made about the sort of person who produced a text, such as their age, gender and ethnic identity. Sociolinguistic research has traditionally viewed language as the product of speakers' experience; for example, Lakoff, Holmes, and Coates' extensive work on gender, where language is frequently presented as a result of belonging to various social groups.

Many researchers have pointed out that people, of course, do not simply fit one category. For example, Fawcett and Hearn describe an individual who has a disability, and the effect this may have on language production: "People with disabilities are not only that; they are black, middle class, Jewish, and so on" (2004:202). Within this multifaceted persona, individuals can be selective about which parts of their sociolinguistic background they choose to emphasise: Lawler observes that individuals can prioritise the more interesting or glamorous parts of their inheritance to form a prominent part of their own identity (2000:59), or identify more strongly with one aspect of their identity than others, depending on the situation.

Bucholtz and Hall acknowledge the benefits of essentialism in highlighting the language of previously overlooked linguistic groups, but argue that essentialism fails to account for intra-group variation, and inter-group similarities (2004:374). This is a particularly important point for the analysis of fiction, where one would often expect variety and complexity between characters of similar social backgrounds, as well as inter-group variety. An absence of this would suggest writers who are reliant on stock characters, each one representative of their sociolinguistic profile. An essentialist approach to a character-driven drama does not account for the differences in language between, for example, in this dataset, Clarrie Grundy and Susan Carter, two middle-aged white, British women from the same geographical area, with similar educational backgrounds. They belong to the same sociolinguistic grouping, but in a drama, one would expect them to be linguistically distinct. This is especially important in a radio drama because of the lack of anything visual to create character; dialogue is the sole way that characters are created, so the audience needs to be able to distinguish between different characters, even when they are from similar sociolinguistic backgrounds. This can be achieved through the actors' vocal qualities, but would also be expected to occur through their characterisation in the written scripts. This could also be in the characters' psychological characterisation as well as their linguistic characterisation, although the two are, of course, not inseparable. As will be discussed more fully in Chapter 6, there is an interplay between archetypal schematic knowledge, and fleshed-out, complex characters. Despite the benefits of essentialist approaches, linguists have cautioned against a wholly essentialist approach to sociolinguistic profiling. Discussing identity, Johnstone argues that: "No matter how refined our models of the various social facts that correlate with patterns of language use – social class, gender, age, ethnic identity, social network, urban versus rural background – we cannot predict what a given person will say in a given situation, or how it will be said" (1996:8).

In their influential paper, Bucholtz and Hall (2005) describe identity in deliberately broad terms as, "the social positioning of self and other" (p.586). They argue that "identity does not emerge at a single analytic level – whether vowel quality, turn shape, code choice, or ideological structure – but operates at multiple levels simultaneously" (2005:586). In their 2004 paper, they assert that, "one of the greatest weaknesses of previous research on identity, in fact, is the assumption that identities are attributes of individuals or groups rather than of situations" (2004:376). This concept of essentialism, where language is viewed as a result of sociolinguistic experience rather than situated in individual interactions, is described as: "a theoretical position that maintains that those who occupy an identity category (such as women, Asians, the working class) are both fundamentally similar to one another and fundamentally different from members of other groups" (2004:374), which leads to the unpredictability described by Johnstone (above), and illustrated in the characters, discussed above, of Clarrie and Susan.

Bucholtz and Hall argue that better theoretical frameworks to study identity are necessary, although they simultaneously caution against fully abandoning essentialism, stating that, "a non-essentialist approach to identity within linguistic anthropology cannot dispense with the ideology of essentialism as long as it has salience in the lives of the speakers we study" (2004:375-376). This salience has parallels in schematic knowledge, discussed by stylisticians, including Culpeper (2001), and discussed in this study (2.7.2). Dividing their study into four key areas of Practice, Indexicality, Ideology and Performance, Bucholtz and Hall emphasise the importance of practice, and the continuing influences on an individual's language. The second concept they discuss – indexicality – is important in identity assumption; because it outlines the flexibility within identity; that one can emphasise certain aspects in certain situations. As discussed in Eckert's analysis of 'jocks' and 'burnouts' (1989), this indexicality is often value-laden, a point echoed by Lawler (2000, above).

The final category of performance is, unsurprisingly, relevant to the data. Bucholtz and Hall write:

> Whereas practice is habitual and oftentimes less than fully intentional, *performance* is highly deliberate and self-aware social display. In everyday speech, as in much linguistic anthropology, the type of display that *performance* refers to involves an aesthetic component that is available for evaluation by an audience. (Bauman 1977) (2004:380).

Grant and MacLeod observe that Bucholtz and Hall's views are a departure from the position held by many computational forensic linguistics such as Juola (2008) and Argamon (2007), who follow the traditional sociolinguistic position that identity is determined by belonging to certain groups. They argue that viewing language as a product of sociolinguistic experience is at odds with current scholarship in other, related disciplines, for example Discourse Analysis, where language and identity are viewed as being created through social interaction, and they set out their own, contrasting position:

> Since our approach is informed more by the linguistic ethnographic concern with how language users orient themselves to these categories (Rampton, 2010), and the processes by which they perform their membership, it is evident that there is a gap to be bridged between

> current theoretical understandings on the one hand, and practical, operational conceptualisations on the other. (2016:54-55)

Grant and MacLeod themselves adopt a position of “a ‘situated identity’ that arises out of interaction,” acknowledging that identity can be fluid, and co-constructed. In a paper analysing the moves and strategies used by predators in online grooming, Chiang and Grant set out this position in more detail:

> We investigate identity here from the constructionist perspective commonly held in contemporary identity research in the social sciences, seeing identity not as a fixed, internal ‘core self’, but as fluid and constructed, or performed, through various modes of expression, the most flexible being linguistic expression. (2019:677)

They analyse the offender’s adopted multiple identities in his attempts to groom children online. They describe the “macro-level” sociolinguistic features, such as age, sex and gender, but point out that speakers and writers are also operating at the ‘micro-linguistic’ interactional level, including such temporary roles and orientations as “evaluator, joke teller, or engaged listener (2019:695).

This last point is applicable to drama scripts: characters in an audio drama exist by doing something, mainly through dialogue, such as gossiping, arguing, telling a joke and so on. This differs from prose fiction or poetry, where characters can be described at length by the author, and their conversations can be summarised, rather than being heard directly. Furthermore, they are, on one level, engaged in an interaction between each other, but on another level are also fulfilling dramatic positions in the play, such as hero, villain or stooge. Viewing identity as constructed provides a framework for separating the identity and linguistic traits of the scriptwriter from the fictional identities of the characters they create. The writers are not bound by their own sociolinguistic background, but are constructing characters instead (discussed in more detail in Section 2.6.3).

2.6.3 Identity Construction: Resources and Constraints

This next section summarises Grant and MacLeod’s theories of resources and constraints in identity production, then applies them to drama, and then specifically to *The Archers.* The idea of language as something produced, rather than something inherited, is developed by Grant and MacLeod. Drawing on contemporary social interactionist terms they view language, “as a *resource* that is drawn on to index or perform particular identities, as opposed to a mere *product* of those identities" (2018:81). They set out a model of linguistic identity (2018, 2020), in which a speaker or writer can draw on resources and constraints, as discussed by Johnstone (1996), in order to negotiate and perform their identity.

One example they give is an earlier author profile report on a text containing phrases that are heavily associated with Jamaican English, such as "bad-minded language". They explain that "this evidenced the individual's language contact with Jamaican English, but that this indexed familiarity with a community of practice not the ethnicity of the author" (2018:91). The distinction allows for the author to be non-Jamaican, but with a high level of familiarity with that particular variety of English. This highlights the ways in which language can be viewed as a resource, rather than a deterministic factor in linguistic identity. In cases where the writer has deliberately attempted to disguise their identity, this distinction becomes very significant, and the sense of identity as something constructed becomes stronger.

Grant and MacLeod identify four categories of resources which writers or speakers can draw on:

> 1. Sociolinguistic history
> 2. Physical self, primarily their cognitions as supported by the physicality of their brain
> 3. Resources of a given interaction.
> 4. Resources provided by specific individuals and audiences and community of practice. (2018:87)

They argue that every writer has constraints as well as resources, which can affect their ability to create a particular identity. In their discussion of constraints, they outline two main types of constraint: firstly, the non-availability of specific resources, for example not being able to speak a certain language; and secondly, that selecting one type of linguistic resource can simultaneously preclude the use of another resource. They state:

> This resource model creates a powerful explanatory framework and understanding of how individuals can actively 'do' identity in different aspects of their lives. It also begins to articulate what we might understand as a unified identity and as a basis for some consistency amongst a wealth of very different identity performances. (2018:88)

The resources and constraints model can be applied to fictional data too, albeit with some notable differences. Scriptwriters have to assume the identities of many characters who are not from their own sociolinguistic background, so they must draw on resources, using imagination, research and the characters' established linguistic identities to enact that character. It would be expected that professional writers would possess good resources of research and imagination to convey a full cast of characters, and would be able to draw on their own skills of observation and writing practice, as well as the resource of the characters' own linguistic histories, to write their dialogue. Further resources are available, such as the agricultural research notes produced by farming experts so that the writers can reflect the characters' working lives.

This resources model can – to an extent – be applied to the characters themselves: to use Grant and MacLeod's example, there could be a white character who indexes a Jamaican variety of English. However, to include a character who spoke in a way that was consistent with Jamaican English, but was, for example, a white British man who had spent his childhood in Jamaica, would require a distracting level of effort from the listener to continually separate the character's perceived sociolinguistic background from the resources they drew on, and, in this particular example, would raise significant issues of representation. There is a delicate balance between characters sounding credible to the audience, based on their sociolinguistic background, and characters in a drama being rounded and able to surprise, amuse and entertain their audience. This is often exploited in comedy sketches, for example, the "intellectual scaffolders" in the TV sketch show, *Harry Enfield and Chums* (BBC, 1990-1998), will privately have erudite conversations in heightened RP accents about the arts and politics, but then to anyone else – for example if a woman walks past – will shout out sexually aggressive comments in a strong South-Eastern accent, conforming to old, well-worn stereotypes of construction workers. Comedy arises from the mismatch between schematic expectations and the scaffolders' private conversations, as well as from the rapid codeswitching between public and private.

This incongruity between expectations of identity, and the identity performance that actually takes place is further exploited in *The Armstrong and Miller Show* (BBC, 2007-2010), which has comedy sketches featuring World War Two RAF pilots. The characterisation of the pilots draws on schematic knowledge about stiff-upper lipped British officers fighting for King and Country. Whilst the characters do speak in clipped upper-class English accents, their vocabulary and attitudes represent an urban, contemporary variety of English, for example, "I bought some really nice trousers in Camden? They is well hardcore with all pockets and shit."[1] Much of the comedy in these sketches arises from hearing exaggerated contemporary, urban slang and modern sensibilities spoken in accents that carry a contrasting set of expectations about the characters' identities (as discussed in Bousfield and McIntyre, 2017). The idea of resources and constraints for fictional characters is more complicated than in naturally-occurring data partly because of credibility of characters, and a wish to surprise audiences with interesting characterisations and to exploit expectations for comic or dramatic effect. For example, in detective dramas, it is desirable for an unlikely candidate to be revealed as the murderer. The resources system becomes very complicated because there are also requirements of plot and audience expectation, and a balance to be found between credibility and intrigue.

In a drama, there are also constraints on language, for both characters and writers. For the writer, there are practical constraints: each episode must be a certain length to fit within the agreed time-slot in the network schedule. A scriptwriter does not have the same creative freedom as a novelist for their characters to converse with an unlimited number of people, since the number of speaking parts mandates the cast size, which affects the programme budget. Also, the form of a drama, in particular a soap opera, creates the strong expectation of certain dramatic structures: scenes that combine to form a number of continuing storylines; a dramatic 'hook' at the end of each episode, and a bigger 'hook' at the end of the week. Such dramatic and formal constraints affect the creation of characters and their linguistic identities. There are parallels to non-dramatic contexts: people have to fit into certain roles (for example, as an employee, as a parent, as a customer and so on) and these 'real-life' roles constrain our behaviour.

There are also expectations about clarity and coherence: *The Archers* uses clearly audible, scripted conversations, as opposed to, for instance, heavily improvised or naturalistic verbatim drama.

[1] (The Armstrong and Miller Show | Best Of The RAF Pilots - YouTube (2017).

There is also an audience-related constraint. The programme has characters ranging from newborns to nonagenarians, but the station has an average listener age of 55, and this is reflected in the topic and tone of the dialogue, which is broadly pitched at middle-aged to older adults. Radio 4's own commissioning guidelines for 2017, the final year of these data, describe the station's audience:

> The station continues to have a balanced audience in terms of gender (49% male / 51% female). The average age of the Radio 4 listener is 56 years old and skews towards an older audience. Our target audience of 35-54 ABC1 (commonly termed 'replenishers') makes up 24% of the audience. The station also continues to have an upmarket bias – 75% of those tuning in fall into the ABC1 demographic.
>
> (https://downloads.bbc.co.uk/radio/commissioning/R4_44_Minute_Drama_Audience_Pack.pdf)

This too is relevant to the language of the scripts: for example, when there are scenes with two teenagers speaking to each other, the language is generally expected to be easily understood by the radio station's average listener, so is unlikely to contain high use of colloquialisms or detailed discussions on age-specific topics, for example, music, which are presumed not to interest the station's average listener. For example, strong language is used sparingly and heavy reliance on 'in-group' terminology is avoided. Even teenagers talking to each other would always use language which is broadly acceptable in public and readily understood by all ages. There is a secondary constraint created by the programme's scheduled transmission slots, which are in the daytime and early evening. This means the dialogue has to be suitable for a broad audience, so strong language is used very sparingly.

These necessary constraints can affect the linguistic features of the characters. For example, sometimes teenage characters can seem rather older than they are, as in this example of a conversation between 18-year-old twins:

> "But for heaven's sake how can you become a full time DJ? That's what every other kid our age imagines doing."
>
> (Writer 2, 2017)

The constraints in a drama script can be external to the writer, such as the broadcaster's guidelines about appropriate language. They can also be internal, in this case, having to write teenagers' dialogue, without necessarily having direct access to the current linguistic patterns of that group. The vocabulary used which sounds too old for teenagers is an example of 'linguistic leakage', which Grant and MacLeod define as a phenomenon which, "occurs when particular aspects of the target identity are picked up and successfully emulated, but a residue of the impersonator's 'home' identity remains" (2020:78). This is the idea that when a writer attempts identity assumption, something of their own linguistic self may unwittingly remain. However, it could also be argued that the writer constrained themselves, to prioritise language which would be readily understood by the programme's core audience, rather than delve too fully into 'teen speak', which might not be understood or appreciated by the programme's broad audience, so it is debatable whether lexical choices such as "for heaven's sake" are examples of linguistic leakage, or simply tailoring a script to the perceived core audience.

Other examples of linguistic leakage in forensic cases might be spelling or punctuation but could also occur in such areas as pragmatics and discourse analysis. Discussing cases where an undercover officer (UCO) has impersonated a victim of grooming online, Grant and MacLeod state:

> Where identity assumption is unsuccessful we will expect to find hybrid identities which draw on both the home resource set of the UCO and also on those of the target identity. (2018:92)

This is comparable to computational linguistics work on Adversarial Stylometry, where the success of the writer is measured in terms of their ability to suppress their own linguistic selves and also assume the identity(s) of another. The average age of the scriptwriters is certainly much higher than teenagers, so it is perhaps unsurprising that some of the vocabulary used by the teenagers and young adults is more closely associated with people decades older, and the idea of linguistic leakage is one which will be explored in more detail in the main analyses.

There is an additional constraint caused by the audio-only format. Without any visual content, any long pauses and silences are problematic too, so writers have quite an unusual constraint, that they are continually *having* to use language to express identity. Constraints occur at a genre level too. Grant and MacLeod describe the example of a victim impact statement, where an individual is constrained by a certain context: they need to perform the identity of a victim, rather than a strong survivor. This links to the dramatic function of a character, such as the character Helen Archer being a victim in a long-running storyline on coercive control. The character is less tightly constrained than in Grant and MacLeod's example, yet it forms an important aspect of fictional identity: if a character needs to fulfil a particular dramatic role, this will affect their identity for the duration of a particular storyline.

Although Grant and MacLeod argue that identity is fluid, they propose that some aspects of identity performance remain stable, whilst other elements will change as part of the identity performance, but conclude that writers have access to both stable resources and dynamic resources. The stable resources could be age, race, or gender, whilst dynamic identity resources influence particular interactional moments. They state:

> Persistence of identity therefore does not require a static and unchanging identity. It does, however, require more understanding about which aspects of identity performance remain stable while the resources we draw on are changing in each specific interactional moment. (2018:86)

In a soap opera, identity persistence is interesting: characters evolve over a number of years. The Ship of Theseus paradox asks whether a ship which has had all of its parts replaced over time is still the same ship. In the same way, can a character be deemed a persistent identity if the scriptwriters and production crew have all changed over the years, and if the actor (as happens) has been recast with a new actor? This could be seen as an extreme example of identity being situated literally in the realm of performance, rather than the (multitude of) writers' sociolinguistic histories.

A further layer of complexity is added if one considers that the identity of the original author behind the mask is itself another constructed identity, rather than a single, fixed entity: in Chiang and Grant's data, the identity of sexual predator masquerading as pre-teen in order to groom young adolescents is not the only aspect of this person's linguistic identity. The same person may well have other identities and language patterns in their role as, for instance, husband, or neighbour. Regarding *The Archers,* if a constructivist approach is applied, each writer has arguably constructed an identity for themselves, not just as a scriptwriter, but more specifically in their capacity as a scriptwriter for an early evening Radio 4 drama. The same writer is likely to have an entirely different style in other contexts. For example, playwright Gurpreet Kaur Bhatti has written for *The Archers,* and also wrote *Behzti* (2004), a play depicting rape and murder in a Sikh temple, which led to protests and riots from members of the Sikh community at Birmingham Rep, death threats against Bhatti herself, and the abrupt termination of the production's run. This opens the question of whether Bhatti's identity as an *Archers* scriptwriter would be recognisable from analysing the text of her controversial stage play. The same question could be asked about *Archers* scriptwriters who have also written for *EastEnders*: also a BBC soap opera, but with a very different style and tone.

2.6.4 Character Consistency

Persistence of identity is an important aspect of Grant and MacLeod's model of identity construction. It can be applied to *The Archers*, where writers are required to create multiple identities: each separate character needs some consistency in order to be recognisable as a coherent fictional character. In their handbook on writing radio drama, Claire Grove and Stephen Wyatt's opinion is that characterisation is the primary interest for *The Archers'* audience:

> As with all soaps, the hook for the audience is the characters. They behave in mostly predictable ways, but they must go in different directions every now and then or the audience will get bored. (2013:19)

The requirement for character consistency links to a further point: credibility. Sacks and Heritage, among others, have argued that linguists should work with naturally occurring data, rather than fictional works. Sacks (1984) argues that, "however rich our imaginations are, if we use hypothetical, or hypothetical-typical versions of the world we are constrained by reference to what an audience, an audience of professionals, can accept as reasonable" (p.25). Further, he argues that unusual and unexpected things frequently occur in real-life conversations, which one might not think to create in a hypothetical piece. Even if these were carefully fabricated, critics could argue that "people don't speak like that." For a drama script, where speech is fabricated, the corollary is that unexpected or extraordinary speeches might well be an accurate representation of naturally occurring speech, but that if they fail to match audience expectations of 'how people talk', then they will be deemed implausible, because there is a burden of credibility on drama scripts which does not apply to naturally-occurring conversation. Additionally, while we expect characters' identities to be consciously consistent to ensure credibility, it is well known that the storylines in a soap opera lead to an implausibly higher number of dramatic events (murder, theft, rape, car accidents, affairs, and so on) than would ever be likely in one small community.

2.6.5 Identity Performance, in Performance

A further point about the performative nature of identity is made by Bucholtz and Hall, who observe that language has a social meaning as well as a referential meaning. Every linguistic choice, for example "whassup?" carries a referential meaning, that is different from, for instance, "How are you?". Bucholtz and Hall argue that "it is precisely this duality of language – its ability to convey meaning at two levels, one semantic or referential and one pragmatic or contextual – that makes it such a rich resource for semiotic production within human societies" (2004:377). In a drama script there is an added layer of complexity to the performative nature of identity: characters are interacting with each other, and may be indexing certain aspects of their identity with each other, but at the same time, there is also an identity performance between the character and the overseeing, or overhearing, audience, rather than the benefit of the characters (discussed in more depth in Section 2.4.6). For example, two characters could be having a conversation, but the audience knows that one of those characters is lying, and this additional knowledge, not available to the other fictional interactant on stage, has implications for how the identity of the first character is written, and perceived.

This occurs at a genre level too. Each character can be practising a certain kind of identity for the benefit of the other characters in the scene, but can also be fulfilling a dramatic function for the overseeing, or overhearing, audience, for example being the "villain" of the drama. In this sense, characters have a dual identity, such as Rob Titchner, the overbearing and controlling husband, who fulfils the dramatic purpose of being the villain of a long-running storyline exploring coercive relationships.

2.7 Fiction as Linguistic Data

The next three sections explore some of the features arising from the use of fictional drama scripts as the data. The first section uses Harold Love's study on authorship to discuss the process of writing fiction; the second discusses scholarship on the ontology of fictional characters, the third reviews literature on the similarities and differences between scripted drama and 'naturally-occurring' conversation, and the final section discusses ways in which analytical techniques traditionally associated with naturally-occurring conversation and non-fiction have been applied to the study of scripted drama.

2.7.1 The Process of Authorship

Having discussed the nature of identity production, it is now important to discuss the *process* of identity production. For these data, this means the process of producing scripts for a long-running radio drama. Love (2002) argues that the prevailing view of an author who sits alone writing their masterpiece, is an oversimplification. He observes that:

> Much consideration of authorial work still takes as its model the single author creating a text in solitariness – Proust's cork-lined room, Dickens' prefabricated Swiss chalet, Mary Ward's elegant study at 'Stocks'. In doing so it restricts itself not just to a particular kind of authorship but to a particular phase of that kind of authorship'. (2002:33)

Love suggests two ways in which the production of a text is more complicated than this. Firstly, the author may actually be more than one person collaborating on a text. Secondly, concentrating only on a 'particular phase' ignores the precursory work, research, sources and other influences; as well as later the revisions and suggestions, which may also influence the final text. He describes these stages of authorship in more detail. These are summarised first, and then applied to the writing process of *The Archers*.

Firstly, Love describes Precursory Authorship. This involves an earlier work, or works, which the given text draws upon. It could be a source, for example, the way that Shakespeare adapted plots from Holinshed and Plutarch, or that *My Fair Lady* is based on *Pygmalion*. Alternatively, Love argues there could be a collaborative element, especially if 'text' is extended to include genre. He gives the example of Clint Eastwood being a pre-cursory author for certain types of Westerns, which draw heavily on his style (2002:41). Herman (1995) argues that the history of the genre will shape what is possible, arguing, “Drama has its own history – other performances, other texts, other contexts of performance, other theatrical conventions – and its own contemporary constraints for aesthetic, experiment of social purposes” (Herman, 1995:10). Also, Love points out, attitudes fluctuate regarding what is acceptable use of an earlier text: what constituted respectful re-use of elegant words in the Middle Ages, might be considered plagiarism after 1700, when concepts of copyright and literary property had emerged.

The second stage of authorship outlined by Love is the Executive Author. This is the person, or people, who actually compose the text. He then discusses the Declarative Author, the (usually famous) person who puts their name to a text which they did not actually write. Love cites the example of Bill Clinton being named as author of a book his staff wrote. A common reason for having a Declarative Author is to obtain a publishing deal or greater publicity by having a well-known person “own” the words publicly, even if the writing was carried out by the Executive Author(s).

The fourth and final type of authorship listed by Love is Revisionary Authorship. This can include joint editing or workshopping of a text; for example, Ezra Pound famously edited *The Waste Land*, and Charles Dickens edited Wilke Collins’ works. Most modern novels include acknowledgements, which almost always include an editor, and often friends, family, agents and other authors, who have read and given feedback on the manuscript.

Love’s categories can be applied to the production process of *The Archers*. The programme’s pattern of scriptwriting is now outlined, with reference to Love’s types of authorship. As with any soap opera, it is written by more than one author. In *The Archers*, this is broken into weekly blocks, with one author writing a week’s worth (six episodes) of scripts at a time. Soap operas are a collaborative process, and necessarily, scripts are not written in isolation. Every script heard on air is part of a much longer process of planning, storylining, synopsis-editing and scriptwriting. Although the details of the production pattern can change from time to time, especially with the arrival of a new

Editor (the creative head of the programme), and more recently, because of the Covid-19 pandemic, what follows is a discussion of the process for many of the years from which these data are taken.

The Archers holds two ‘longterm’ script meetings per year, where the writers and production team discuss the major storylines, such as Jack Woolley’s long struggle with Alzheimer’s. Discussions sometimes reach even further ahead; for instance, looking at the long-term implications of a new baby in a particular family. Once the broad brushstrokes of these storylines have been agreed, they are written up as a ‘long-term’ storyline document: each major storyline (e.g., Jack Woolley’s Alzheimer’s) is written as a narrative, including major plot points and a rough timeline. An aim of this process is to ensure that the different strands, both serious and light-hearted, are woven together to maximum effect. Using Love’s framework, this ‘longterm’ document is a piece of precursory authorship. However, before the meeting, any of the scriptwriters and production team can submit ideas for longterm storylines, so even the precursory document has earlier texts which influence its creation.

Between the ‘longterms’ meetings, there are monthly script meetings to discuss a four-week block of scripts, attended by all the scriptwriters, regardless of whether they are commissioned to write in the next block of scripts or not. Before this meeting, the production team creates a script pack, which includes notes from the ‘longterm storylines’ document for that month, as well as a compilation of ideas suggested by writers, and research notes. The research notes could be agricultural, for example, information on lambing; they could be related to a current storyline, for example, information about Alzheimer’s; or they could be about public events which may be mentioned by characters, for instance, Brexit or a major sporting event. The notes could also be an archive note, such as a reminder that two characters have previously argued over a certain topic, or that a significant birthday or anniversary will occur in time period which the new scripts will cover. At the script meeting, the Editor leads the discussion between writers and production team to discuss the details of the storylines and sub-plots for the four-week period. Notes are taken by a producer incorporating the contributions of the attendees, and after this meeting, the production team writes up a storyline document, which draws heavily on notes taken during the meeting. The storyline document outlines the main narrative points of the month's different stories, usually broken down into character groups, e.g., "Jolene and Kenton at The Bull". The document will state which parts of the plot occur in which week, so that each writer knows when to start and end their section of the story.

Here is an example from a 2013 storyline document, describing a scene where Lilian is being driven along by her secret lover, Paul, while discussing her long-term partner, Matt:

> MATT, LILIAN AND PAUL Lilian is worried by his fury and his erratic driving. She persuades him to stop the car in a lay-by well away from Ambridge, and she listens to his ranting. It's just not normal. In the end she asks Paul whether he still loves his wife. He freaks: of course not. It's Lilian he needs and wants. If she would just commit to him.
>
> (http://www.bbc.co.uk/blogs/writersroom/entries/989184e1-10a1-3c2d-916ecfbf67c7a334. *All quotations from this section are taken from this same blog post*, accessed November 2017).

This storyline document forms part of the precursory authorship, and is the result of collaborative input, and is itself drawn from a number of earlier sources, as discussed. The writer working on the week's scripts uses this storyline document to write a full synopsis document for the week. Keri Davies, who was the writer for this particular week, explains:

> I have to decide which parts of which story I will tell on each day, which characters we need to hear, and which bits we can hear "by report". Where possible, we try to find ways to link stories, so the listener isn't simply jumping back and forth between isolated strands.

The synopsis fleshes out the storyline document, to produce a much more detailed outline, which includes locations, timings, speaking characters, and so forth.

> 4. INT. PAUL'S CAR. 1030 HRS LILIAN, PAUL Furious Paul continues to drive erratically. Worried Lilian persuades him to stop the car in a lay-by. At least this is well away from Ambridge. She listens to his ranting. Celia has turned the whole family against him, and this Frank is no way the right man for her. Lilian asks Paul whether he still loves Celia. He freaks. Of course not! It's Lilian he needs and wants.

As is standard practice, much of the text has been lifted directly from the storyline document to create the synopsis document, and lexical choices have been repeated, such as "furious", "erratic" and "ranting". Whole phrases, such as "of course not" have also been repeated. In the above synopsis document, more detail has been added – in this case about Paul's ex-wife Celia – and gives a clearer indication of what the dialogue will be. This is not always the case. In a later section, the synopsis simply states, "It escalates until Lilian can't deal any more with Paul's unreasonableness." As the writer explains, “You can see that at this point I didn’t know exactly what they were going to say as the argument escalates.”

All four writers for the block submit their synopsis simultaneously. This allows the production team, usually the Editor and producers, to read all four weeks and ensure that the plot is coherent. There are also practical considerations, such as actor availability, which forms part of the editing process. After the writer has received feedback on their synopsis, they have approximately 11 days to write their six episodes. The following extract shows the script version of the synopsis quote:

> LILIAN Okay. Fine. So let her be unhappy, and then you can say “I told you so”. I just can’t see why it’s that important to you who she marries. Unless... (BEAT)
>
> PAUL What?
>
> LILIAN Unless you still love her? Is that it?
>
> PAUL (IN HIS CONTROLLING WAY HE DOES, ALTHOUGH HE’S IN COMPLETE DENIAL ABOUT IT. SO HE FREAKS) No! Of course not! How can you even say that?
>
> LILIAN You’re getting so het up about it. You won’t let it rest.
>
> PAUL I love you! How many times do I have to say it? I love you and I need you. And you love me.

Whilst the script contains substantial amounts of new dialogue, it can be seen that certain words have remained from the storylines document written by the production team. The phrase “he freaks” appears in all three iterations of the story, albeit as a silent stage direction in the script. The phrase “Of course not!” in response to Lilian asking Paul if he still loves his ex-wife has remained. Also, there is an echo of the storyline document (“It’s Lilian he needs and wants”) in the line, “I love you and I need you.”. In another example, the phrase, “turning up at my home” can be traced from its first appearance in the production team's storyline document, when Lilian tells Paul, "she certainly won't be forced into making that decision by him turning up at her home." In the writer's synopsis, this appears as, "Lilian tells him she isn't in a position to do so, and she certainly won't be forced into making that decision by him turning up at her home." The same phrase recurs in the writer's script, along with "forced", which has also remained present in all three versions of the story: "And if you think you're going to force my hand by turning up at the house, then you're out of your mind". Although the exact formation of the words has changed, the influence of the preceding versions is clear, and whilst the story has changed forms, it is interesting to see the influence of each text on its successor.

This linear description of the production process mirrors neatly Harold Love's ideas about precursory, executive, and revisionary authorship. There are some additional aspects, which stand outside of this chronological process. Declaratory Authorship is not fully applicable here in the sense of the Bill Clinton example: the scriptwriters have largely written the work for which they are named on the script as the author, and given an "on-air" credit at the end of the week, even though – as discussed – they are not the sole creator of each episode. In this sense, each writer is the Declaratory Author of that episode, despite it being a collaborative process. Simultaneously it could be argued that the BBC, as copyright holder, is another Declaratory Author, and in the event of a particularly controversial storylines, it is usually the broadcaster who is criticised, not the individual writer.

In addition to all the storyline documents which form the Precursory Authorship, there is a much broader type of Precursory Authorship which occurs, using Love's example of Clint Eastwood being a type of precursory author, by defining the features and constraints of a genre. Soap opera has its own style and characteristics as a genre, and *The Archers* has its own style within that genre. Geraghty outlines key characteristics of a soap, and refers to the distinctive styles of different soaps:

> The audience is presented with a rich pattern of incident and characterisation – the dramatic is mixed in with the everyday, the tragic with the comic, the romantic with the mundane. The proportions will vary from serial to serial. *The Archers* sometimes seems to consist of nothing but the humdrum, while Crossroads frequently veers towards melodrama. (1981:12)

This – perhaps somewhat unkind – view of *The Archers* alludes to some of the differences in soap styles, and links back to the nature of identity, (Section 2.6.3) that a scriptwriter's identity as an *Archers* writer could be different from their identity as an *EastEnders* writer. Audience research has shown that the demographics of *Archers* listeners are very different from BBC One's *EastEnders*, so it is unsurprising that the soaps' styles reflect this. When a new Editor introduced more melodramatic storylines into *The Archers*, he was accused of turning *The Archers* into "*EastEnders* in a field" (Greenhill, 2015). This criticism highlights the strong sense of identity that a soap has, which acts as a form of precursory authorship. Using Grant and MacLeod's theories of identity, the programme's long-established style also acts as both resource and constraint for the writer.

Arguably in drama, there is an added element of Revisionary Authorship, which is realised in the production process. Once a novel has been written, revised, edited and published, there is a definite 'final form', but a script never quite has this complete status: instead it is a handbook for the cast and crew, who then convey the script to its intended audience. The performance could be included in Revisionary Authorship, but since it is beyond the point of the text being revised, it seems to be worth considering performance as a later, and separate element. In broadcast media, there is an even further element of revision: reviewers and audiences respond to the programme online. In a long-running drama, the views expressed along with traditional critics' opinions, could feed back into the production cycle. For example, an unpopular character might be written out, or a badly received storyline could be toned down in response to critical reception or audience's reactions. In this way, audience reception could be seen as a revisionary text for a transmitted episode, whilst simultaneously being a precursory text by influencing decisions for future episodes.

The importance of this process to authorship analysis is the understanding that individual writers are not creating scripts from a blank page, but are moulding their dialogue from a pre-existing text. In Grant and MacLeod's model, these precursory texts could be viewed as resources for the writers to draw on. The extent to which different writers incorporate vocabulary from these precursory documents could be individuating in itself, but has a wider importance in terms of understanding the context in which individual scripts are produced, and the ways in which some parts of the text may be more the result of textual influence rather than individuality.

2.7.2 The Ontology of Fictional Characters

Following the review of the way identity is conceived, it is worth considering how fictional characters have been viewed; both in their creation and reception. In his influential study, *Language and Characterisation* (2014), Culpeper observes that "much literary critical energy, such as it is, has been spent debating the ontological status of character" (2014:6). Culpeper draws on research in linguistics, cognitive psychology, social psychology and stylistics to examine the process of characterisation in plays. He sets out two main approaches to characterisation: humanising and dehumanising. The humanising approach, popular in the early twentieth century, argued that characters in plays were "imitations or representations of real people, or – the more extreme view – that they are actually real people." This was most famously adopted by A.C. Bradley in his 1905 work, *Shakespearean Tragedy*. In contrast, the 'de-humanising' approach denies that fictional characters have any human qualities, and argues that "characters are products of the plot or simply a textual phenomenon" (2001:6). The humanising approach to characterisation was condemned, notably in L.C. Knights' parodic essay, "How Many Children Had Lady MacBeth?" (1933), which criticised scholars such as Bradley for creating worlds and events for fictional characters beyond those which occurred in the original text. Although the humanising approach has become far less common in modern literary criticism, this is the approach which is arguably most pertinent to playwrights and their audiences. Culpeper, who advocates a mixed approach, observes that humanising characters is an important part of an audience's appreciation of drama:

> It is difficult to deny that what we all do when we watch a play or a film is to attempt to *interpret* characters with the structures and processes which we use to interpret our real-life experiences of people. We also frequently talk about characters in terms applicable to real people. Even writers who express some doubt about the humanising camp admit that you cannot entirely get away from this idea. (2014:10)

McIntyre, building on Eder et al. (2010) sets out the different views of fictional characters:

> 1. Semiotic theories consider characters to be signs or structures of fictional texts.
>
> 2. Cognitive approaches assume that characters are representations of imaginary beings in the minds of the audience.
>
> 3. Some philosophers believe that characters are abstract objects beyond material reality.
>
> 4. Other philosophers contend that characters do not exist at all.
>
> (Eder et al. (2010:8) in McIntyre (2015a:150)

McIntyre observes that the second of these positions is currently dominant in stylistics (followed by Culpeper, 2001; Culpeper and McIntyre, 2010, among others). McIntyre writes:

> In stylistics, characterisation commonly refers to the cognitive process by which readers comprehend fictional characters. In effect, characterisation is the process of forming an impression of a character in your head as you read. This includes determining the personal qualities of the character in question as well as other aspects such as their social and physical characteristics. (2015a:159)

Whilst there is an emphasis here on readers' responses, the same can be applied to the writing process; in particular in continuing dramas, where the characterisation and plot are created from the producers' and writers' comprehension(s) of the fictional characters in previous episodes.

As an aside, it is worth noting the multifunctional quality of the word "character". Culpeper uses the term 'character' to refer to the people that inhabit the fictional worlds, and 'characteristics' to describe the qualities that form a personality (2014:2), a convention which is followed here. In contrast, Pfister prefers the term 'figure' because it alludes to the functionality of characters, and because it emphasises "the ontological difference between fictional figures and real characters" (1991:160). Using the analogy of chess pieces, Pfister argues that the form of a 'figure' is inseparable from its function:

> Dramatic figures cannot be separated from their environment because they only exist in relationship to their environment and are only constituted in the sum of their relations to that environment. Social conditions can influence or determine the life of a real person, but, in drama, the fictional context serves the function of actually defining the fictional figure." (1991:161).

In a continuing drama, the relationship between figure and function becomes more complicated because the characters are pre-established before a given episode is written, so in this specific genre, characters do exist separately from plot.

Our understanding of both character, and characteristics, is of course something which varies enormously between historical periods, geographically, and in the varying styles of theatre and film. Such a discussion is outside the scope of this study, and ‘characters’ in this study refer to a broadly contemporary conception of character, in contrast to, for example the chorus of Greek tragedies or symbolic ‘characters’ in Medieval Mystery plays. Pfister describes characterisation in naturalistic and realist drama as distinct from the more emblematic figures of early drama:

> The figures are actually conceived as multidimensional individuals and not as idealised representatives of mankind. For this reason the figures’ respective levels of awareness are restricted and relativised by the emphasis on the irrational qualities of their emotions and moods, on the unconscious influences exerted by milieu and atmosphere, and on the subconscious influence of collective drives and traumatic experiences. (1991:183)

Pfister’s conception of ‘figures’ suggests that our perception of character is shaped by wider influences: the genre of the play, and the era in which it is being written and received will all affect our perceptions of the characters. The genre of the data is a soap-opera, which mandates a high frequency of melodrama, so it is not entirely accurate to describe the data as naturalistic or realistic. One of the ironies of a soap opera is that “many of the characters are recognisable, “ordinary” folk, speaking down to earth language, and dealing with the stuff of everyday life” (Thompson, 2011:3), yet the events that happen to them are extraordinary, particularly if the accumulated number of extraordinary events happening to one single person is considered. For example, the character Helen Archer has: had her older brother die in a tractor accident, had her partner kill himself, suffered from anorexia as a result; accidentally run over a pedestrian and let her younger brother take the blame; been cheated on by a later boyfriend, who had an affair with her adult stepdaughter, decided to have a baby by donor sperm, almost died during labour, become a victim of coercive control, been raped, stabbed her husband, had a second baby in prison, and been acquitted at trial. Yet, despite these extraordinary life events, Helen is portrayed as a very “ordinary” person, who lives with her two sons and works in a shop selling cheese.

This disconnect between character and events is common in soap operas, and adds to the long-running discussion about the symbiotic relationship between character and plot. Pfister describes the two as inseparable: “in drama the presentation of a figure without even the most rudimentary plot and the presentation of a plot that does not contain even the most drastically reduced form of figure is inconceivable" (1991:161). In *Poetics*, Aristotle stated that, "in drama action comes first, and that characters are foremost 'agents' of the action" (cited in Culpeper, 2014:8). Pfister argued that this undervalued the importance of characters' dialogue:

> The characters are allowed to present themselves directly in their role as speakers. It is therefore the figures' speech, and above all, their dialogical speech, which constitutes the predominant verbal matrix used in dramatic texts – something that was scarcely even acknowledged in Aristotle's essentially plot-oriented poetics, and that was all but ignored until the dramatic theories of A.W. Schlegel and Hegel gave it the recognition it deserved. (1991:7)

The interplay between characterisation and plot remains at the heart of drama. When appraising a completed drama, play or novel, it is impossible to know to which came first – character or plot. In some cases, it might be possible to hazard a guess, but it is impossible to know the extent to which the playwright created the personal attributes of the characters to facilitate events in the plot; or the extent to which the direction of the plot was constrained or enabled by the writers' sense of the "personalities" of the characters. The relative influence of characterisation to plot can vary, not just between writers, but also between plays by the same writer; as well as between characters in the same play, and also between different drafts of a play. Once a playscript is completed, it is not possible to know the extent to which plot influenced characterisation or vice versa. In a soap opera, such as *The Archers,* the situation is different. This interrelation between character and plot in the writing process is interesting because the text is – by its own nature – unfinished and continually being created. The new episodes being written are informed by the established history of the drama and its characters which has already been broadcast. New plot events, large and small, take account of the characters' established histories and personalities. The pre-existence of plot events and characterisation fits with Grant and MacLeod's resources and constraints model: for example, the character Jack Woolley' was diagnosed with Alzheimer's. This storyline provides a rich resource to the writers of possible situations and emotional trajectories to develop from this given situation; it also provides a constraint, because Jack cannot suddenly return to his previous cognitive abilities for the sake of a later storyline.

The following selected examples demonstrate how the writers and production team overwhelmingly adopt a 'humanising' approach to characterisation. Firstly, scriptwriter Keri Davies describes how every episode of *The Archers* has a programme synopsis, which is kept in the programme's archive for future reference. Character notes are also kept, including this example, which is about Ruairi, a young boy who is going to live with his father (Brian) because his mother (Siobhan) is dying:

> Ruairi likes his books "Three little pigs" and "Where the Wild Things are" and when he's watching "The Jungle Book" he needs a cuddle when Baloo dies. In summer he likes to eat outdoors and his favourite food is hot dogs made with bratwurst. Mousey is his comforter. Siobhan bought a duplicate in case the original got lost, but Brian would need to dirty it up a bit. (Davies, 2013)

The Mousey comforter, which was noted, was then referred to in an episode a number of years later when Ruairi went to boarding school.

Characters' preferences and tastes are also recorded it the programme archives, in order to maintain character consistency, as Davies recounts:

> I recently wanted a reason for Brian to have left Lilian and Matt's dinner table. Perhaps he didn't care for the dessert? Camilla [a former programme archivist] confirmed that if Jennifer said Brian doesn't really like meringue it didn't conflict with anything in the archive. *Archers* listeners care about these things. And quite right, too. (Davies, 2013)

These examples illustrate how the writers and production team (and listeners) are firmly in the 'humanising' camp of characterisation, treating characters' histories as 'fact' within the world of the drama.

This next example highlights the evolving relationship between characterisation and plot which occurs in long-running dramas. Scriptwriter Tim Stimpson, discussing the high-profile storyline about Rob's coercive control of Helen Archer, explains:

> I introduced Rob as a sort of domineering, chauvinistic type of character. Some people would find his traditional sensibilities very charming, but it would wind others up - particularly Helen's mum and dad. It was only when Sean O'Connor became Editor that he suggested taking the story further, culminating in the stabbing. (In *Woman and Home*, 2016)

This comment highlights the process of characterisation in a soap opera: characters' personalities are established in the show, and in this instance are being discussed as if they are real-people, including the effect they have on other fictional characters. From this base, characters are given storylines which can culminate in extreme actions – in this case a stabbing – which in turn reveal or explore new layers to that character's personality. Whilst the extreme storylines of a soap opera put these ordinary characters in extraordinary situations with an implausible regularity, these examples demonstrate how the characters are still thought of as real-people by the creators and audience, in terms of their reactions to these extreme situations.

The third example is a recent piece of feedback on a draft script from the programme's current Editor, Jeremy Howe, which reads, "Jakob needs to be more Jakob, Kate more Kate" (unpublished note, 2021). Again, this feedback note illustrates how characters are conceived as having a recognisable personality, which is produced at a linguistic level, through their dialogue. From this script-editing note, it can be inferred that sometimes writers are able to create more closely the shared conception of a character's personality, and that sometimes the dialogue needs a further draft to create the linguistic persona in the way that it is conceived by the producers.

Culpeper's model (1994, 2001) which combines top-down and bottom-up processing for characterisation is a useful way to think about how writers create characters, as well as the cognitive stylistic focus on readers' comprehension. Discussing this process, McIntyre writes that, "Schemas can be formed directly (i.e. as a result of personal experiences) or indirectly (as a result of reading or watching plays, films, etc.) (2015a:152). McIntyre further explains that Schneider (2000, 2001) suggests a similar approach, "in which characterisation occurs when readers combine knowledge stored in their long-term memory (i.e. prior world knowledge) with textual knowledge accumulated in their working memory" (2015:152).

In forensic contexts, schema theory could also be applied to how people impersonate others online, for example the Undercover Officer scenario, when Grant and MacLeod (2016) found that the officer impersonating a teenager used lots of initialisms and 'lol' which the teenager had not. One explanation might be that the officer had a preconceived idea of teenage text-speak, which influenced the linguistic features which they used to impersonate a teenager. There is a difference in characterisation for forensic scenarios and fictional settings – the targeted audience needs only to convince the one correspondent (in this case, the suspected paedophile), compared to needing to convince a broader audience. Arguably for the mass media scenario, there is more engagement with the 'top-down' aspect because the writer needs to convince a broad audience who may or may not have direct knowledge of a character and not realise that they are breaking the stereotype. In a sense, one has to draw on enough elements of the schemata to convince the audience that this character is credible, whereas the undercover officer only has to convince one, or perhaps a small number, or interactants, who have specific knowledge of the original interactant. With a mass audience (whether from broadcasting, publishing or theatre) there is an underlying suspension of disbelief to overcome. E.M. Forster (1927) famously observed that rounded characters should be able to surprise an audience. Yet, whilst thwarting those expectations is a key part of skilful writing, there is still the issue of credibility and needing to draw on enough of the audience's 'top-down' pre-conceived knowledge to convince them the characters are credible, but also give them the satisfaction of being surprised by the characters' actions.

These 'top-down' expectations and 'bottom-up' textual implications also unfold diachronically. Referring back to scriptwriter Tim Stimpson's description of Rob as a "as a sort of domineering, chauvinistic type of character", it is clear that he as the writer is drawing on a recognisable "type" of character. Then as the character develops over time, more layers, complexities and idiosyncrasies are planted in the textual clues to allow the character to grow and develop beyond those initial expectations and archetypes.

2.7.3 Fictional Dialogue compared to Naturally-occurring Conversation

Naturally-occurring speech has generally been seen as the preferred data source for linguistic analysis. Grant and MacLeod, in their paper arguing for the merits of experimental data, observe that:

> Since at least Labov (1966) sociolinguists and discourse analysts concerned with the description of variation in language across different contexts have had a strong focus on naturally occurring language data. Many researchers express a preference for such data (see, for example, Eysenck, 2014; Hepburn and Wiggins, 2007; Johnstone, 2000; Potter, 1997). (2016:50)

This position partly stems from a concern about using constructed examples to illustrate linguistic arguments. In a chapter advocating the use of fictional data to study pragmatics, McIntyre and Bousfield recount Stubbs' (1993) criticism of Halliday's tendency to rely on "invented sentences about aunts, dukes and teapots, or about Christopher Wren and a gazebo (Stubbs 1993:9, cited in McIntyre and Bousfield, 2017:759), and describe the enduring concern that invented examples cannot sufficiently encapsulate the complexity of naturally occurring language. They observe that consequently, the linguistic study of fictional texts has remained almost entirely within the field of Stylistics, where linguists are specifically concerned with the nature and function of literary language in itself, rather than as a way of exploring language more broadly.

Despite these long-standing preferences for naturally-occurring speech, Grant and MacLeod, and McIntyre and Bousfield, have argued for the merits of expanding the types of data used to investigate applied linguistic questions, and also for the similarities between "natural" and "constructed" data. Just as Grant and MacLeod argue there are merits in using experimentally created data, McIntyre and Bousfield also argue that "advances in corpus analytical techniques have begun to show that some fictional data is perhaps not as different from naturally-occurring language as we might first have assumed" (2017:759), a position echoed by Braber (2018). Bousfield and McIntyre cite Quaglio's (2009) analysis of the TV sitcom *Friends*, which uses Biber's (1988) multidimensional analysis method to compare co-occurrences of linguistic items against a reference corpus of the Longman Grammar Corpus conversation section. Quaglio found many features in common that reflected the shared context, such as first and second-hand pronouns, ellipses, substitute pro-forms (e.g. one / ones, do it/ that) and deictic expressions (e.g. this, that). (2009) Several differences were noted, such as the lack of interruptions and overlaps, and repair, which are generally avoided for the sakes of pace and clarity:

> Because conversation is interactive, speakers are often eager to participate in the communicative event. This cooperation often results in overlaps, interruptions, and incomplete utterances without interfering much with the flow of the exchanges. The virtual absence of these features in television dialogue is probably one of the most salient differences between the two registers. (2009:3)

There are differences between naturally-occurring conversation and scripted dialogue, but nonetheless, despite these differences, Quaglio concludes:

> Once the differences are acknowledged, the numerous similarities can be explored for different purposes. For example, the use of television dialogue as a surrogate for natural conversation for the analysis of certain linguistic features seems perfectly appropriate. As Rey (2001) has noted, the language of television dialogue is a reflection of the perception that scriptwriters (and actors) have of actual conversation. As such, the analysis of certain features – especially those that are less likely to be captured by a corpus of natural conversation – could be based on television dialogue. (2009:148-149)

Furthermore, Quaglio is comparing scripted dialogue to naturally-occurring speech, but if the point of comparison is moved to compare scripted dialogue with other forms of interactive communication, such as online chatrooms or text messaging, some of these differences may be reduced: online textually-realised communications often incorporate features of spoken language, and allow writers to edit their words at the point of writing and posting, which is similar to the process of writing drama scripts.

A final, and obvious, point regarding the pre-eminence of "naturally-occurring" spoken data, is that there is no single, homogenous type of "naturally occurring" conversation. Herman remarks that "In everyday contexts, too, variation is the norm. Dialogues in courtrooms differ from those in classrooms; social chit-chat differs from those parliamentary debates" (1995:3). There is an almost infinite variety of spoken interactions which could be used as "naturally-occurring" spoken data, so comparing scripted dialogue only to a prototypical version of conversation (such as the Longman Grammar Corpus) cannot capture the full range of comparisons available. Many speech situations (for instance teacher-pupil communication, or between shop staff and a customer) place certain restrictions or expectations on the speakers. Whilst a comparison of a prototypical drama and prototypical naturally-occurring conversation (if such a thing exists) might display certain differences, comparing the "everyday" language of soap operas to naturally-occurring conversations such as instant messaging chatlogs, might yield examples of language which are far closer. Additionally, authorship attribution techniques and methods are not completely re-invented when moving between text-types, so it stands to reason that exploring techniques which could distinguish between scriptwriters could potentially yield results that would distinguish between other forms of communication.

There are also practical benefits to using drama scripts as data. One benefit is that the data are accessible and plentiful, and whilst scripts may have copyright restrictions, they are less likely to have ethical considerations. There are plenty of digitised scripts, which cuts out the significant amount of time required to arrange, record and transcribe conversations.

Having summarised some of the ways in which scripts may be similar to naturally-occurring conversation, what follows is a brief summary and evaluation of some of those differences and issues to consider when using drama scripts as a proxy for linguistic identity disguise. One of these features is spontaneity. Kozloff argues, “in narrative films, dialogue may strive mightily to imitate natural conversation, but it is always an imitation. It has been scripted, written and rewritten, censored, polished, rehearsed, and performed” (Kozloff, 2000:18). Although this analysis considers scripts, and not the finished production, it is still the case that even the “first draft” script which has been sent to the drama producers for feedback, will have been planned, and almost certainly revised by the writer, before it is submitted to the production team. However, the same can be said of other texts which form the basis of forensic authorship attribution, such as emails and social media posts. Many texts which are the subject of authorship attribution are written texts, such as emails or social media posts, which the writer can have considered and re-drafted before posting or editing. Even once posted, many social media sites allow their users an “edit” or “delete” function.

The production schedule of *The Archers*, and indeed other soaps, and other media with a regular output, such as newspapers, mean that writers have to work rapidly. In a standard production cycle for *The Archers*, writers would have around eight days to write six episodes, which averages at approximately 1800 words per day. This is different from a novel or single play or film script, where a writer might have weeks and months to re-draft, discuss, workshop and refine their play. On a line-by-line basis, a writer could spend less time writing and editing a single line of *The Archers* than the equivalent amount of time spent writing a text, email or posting on social media. Even though the communication is less immediately synchronously bound than in online chatrooms, there are some time constraints in writing a soap which severely limit the overall time available to a scriptwriter for revising and editing their words.

Another, obvious, way in which drama scripts differ from everyday conversation is the presence of an audience. Pfister (1991), Short (1996), McIntyre (2015b), Wallis and Shepherd (2002), among others, discuss the nature of stage dialogue as being a conversation between two or more characters, but simultaneously – and primarily – a message from the playwright to the audience (see also 2.4.6 for a discussion on *character style*). Wallis and Shepherd state:

> Let us start by stating the obvious: what is said on stage is designed to be heard by the audience. When one character speaks to another, he or she sends a message – about themselves or their desires or what happened last week – to the other character, who receives this message. But this onstage communication itself, involving both sender and receiver, can be thought of as another message, one being sent from the stage to the audience – again about character, action and circumstance, but always a slightly different message from that between character and character. (1998:41)

The audience, listening in to any conversation on stage, is the primary overhearer, or intended audience of the utterances. Further, there is an authorial voice which is the primary communicator behind the speaking characters, and the two voices cannot be conflated. As Pfister argues, both the author and the speaking character are expressive subjects, and it would be naïve to confuse the two:

> The degree to which figure orientation may dominate over authorial orientation, or vice versa, varies considerably. The wit expressed in an Oscar Wilde comedy constantly draws the audience's attention to the wit of the author, whereas the plays of the naturalist school, such as those by Ibsen, are attempts to establish the absolute dominance of figure orientation and eliminate all references to the author. (1991:103)

In the case of *The Archers*, there is an established "voice" to the programme, although this has changed diachronically from an information-based programme about farming methods to a contemporary drama in a rural setting. This creates another layer of 'author', because the individual scriptwriter is partly a conduit for the overarching tone and "voice" of the programme.

Audience Design is another important area of difference between naturally-occurring conversation and scripted drama. The impact of the listening, or overhearing, audiences is discussed in Bell's influential paper (1984). He outlines the different levels of 'listener' and argues that the further away from the speaker an audience is, the less influence they will have on the way the speaker chooses their words. The concentric circles in Bell's diagram (1984:159) place the Speaker in the middle, and move outwards through the various layers of hearer (addressee, auditor, overhearer and eavesdropper). According to Bell's theory, a speaker will be influenced by their addressee, and will bear in mind any auditors in the room (for example, the presence of children in the same room, or diners on the next table in a café). An eavesdropper implies an unknown listener, so it is logically impossible to accommodate style for their presence, although a speaker might be wary of potential eavesdroppers, especially if discussing something confidential. In this framework, the audience at home is the eavesdropper, and so, according to Bell, would have the least bearing on the character's language, but this model cannot be fully applied to drama scripts, because the "eavesdropping" audience is actually the primary addressee (Short, 1989:149; Wallis and Shepherd, 1998:41).

Similar disruptions to Bell's model have been found in forensic contexts too. Haworth (2013) argues that this theory of diminishing influence is unsatisfactory when analysing police interviews. She explores the ways in which the police officer and suspect have a conversation which is discursively imbalanced because the police officer is orienting the discussion to future listeners, such as prosecutors and the court, while the suspect often orients their talk only to the police officer, and is less aware of future audiences. Haworth cites the example of a police officer (IR) saying:

> IR: okay you're shaking your head.
>
> IE: (yeah)=
>
> IR: → =for a no. okay mate. … (2013:57).

This exchange is, of course, completely unnecessary for the participants in the room at the time, and is instead making something visual explicit for the audio recording, which will be for the benefit of future listening audiences. Bell argues that this case of overhearer design, "clearly influences a speaker's style, although it is evident at macrolevels of language rather than in the quantitative shift of microvariables" (1984:177). Haworth (2013) and Stokoe and Edwards (2008) show ways in which the future listening audience do indeed influence the detail of the conversation at the level of microvariables. In these data, listeners at home are the primary audience, yet they are cast as "eavesdroppers" in Bell's model.

Similarly, in drama, Bell's model does not account for the double element of performance in drama. The entirety of a script is for the benefit of the audience, but this can be seen specifically in the "microvariables" too, as in the next example. In the scene, Helen has taken her young adult step-daughter, Annette, to terminate an unplanned pregnancy. While Annette is having the procedure, Helen steps outside to phone her own mother, Pat:

PAT   (ON PHONE) Hello. (BEAT) Is everything all right?

HELEN   (COVERING, JUST FELT COMPELLED TO RING) Yes, yes, fine! I, **er**, it was just to say thank you for last night.

PAT   That's very nice, but there's no need, you thanked me before you left.

HELEN   Yes, **but anyway**...

PAT   I thought it went well.

HELEN   Yeah, me too!

PAT   Good. (BEAT) Are you at the shop?

HELEN   **No!** I'm, **um**, going in a bit later. Kirsty's opened up.

(Writer 3, 2010)

Helen is ostensibly talking to her mother about the previous night's meal. However, the hesitations in "er" and "um", and the tailing off speech, "Yes, but anyway…" and the over emphatic "No!" are for the benefit of the audience, conveying Helen's emotions arising from the current situation, which she is attempting to conceal from Pat. Dramatic dialect is also highly likely to convey emotions in a way that naturally-occurring conversation may not. Whilst people are often adept at putting on a brave face which can convince others, this would not work here dramatically: if Helen produced a flawless performance of a happy daughter chatting on the phone, and was not signposting her inner emotions to the audience, then the scene would seem irrelevant, and would reduce any dramatic tension or jeopardy about the events occurring (out of earshot) inside the abortion clinic. For the scene to have dramatic purpose, there has to be a dual performance, one for the audience, another for the other characters in the scene.

Another example of multiple audience design if the character is speaking ostensibly to one character but for the benefit of another character. In the next example, Rob, who was eventually accused of coercively controlling his wife Helen, is talking to his step-son, Henry, but is really using the opportunity to convey a threat to Helen's brother (Tom) and mother (Pat), that he will be moving closer to Helen.

```
ROB  Well it's a senior role, supervising their
     agricultural holdings. But I'll be based here.

PAT  When – (did this happen)

ROB  So Henry... Daddy is going to be working very
     close to home. Won't that be nice?
```

(Writer 4, 2016)

Rob is apparently telling his young step-son Henry some good news, but is fully aware that Pat and Tom are listening, and will not be happy, so there is a dual audience design in his speech, to Henry, and then to Pat and Tom. There is even a third implied audience, in that the information will presumably be filtered back to Helen. Furthermore, this exchange is designed to signal information to the audience at a meta level of programme genre: signalling that the storyline of coercive control will continue. The final sentence "Won't that be nice?" is the "hook" at the end of the episode, which adds a dramatic weight for the audience, but would be outside of the characters' awareness. On this level, it operates at the character-to-character level, but is also a direct message from scriptwriter to audience. This is different from Haworth's example of the police officer and the interview tape, where the future audience is separated by time, but has echoes because the speaker (Rob) is saying something for the benefit of an overhearing audience, and also a future audience (Helen).

A further element of audience design is the intermediary audience of the production team. This works diachronically: in the production process, a scriptwriter is likely to become aware of the particular preferences of the script editor and editor, and may shape their next script accordingly, knowing that the scripts need to be “signed off” by the production team. This adds an extra layer of audience design into the writing process.

Plot is the final area of difference between naturally occurring conversation and scripted data which is discussed here. While a naturally-occurring conversation need not have any particular purpose, each scene in a script is presumed to exist for a reason. Pfister argues that:

> Since dramatic dialogue is spoken action, each individual dramatic utterance does not just consist in its propositional expressive content alone, but also in the way it is itself the execution of an act – whether in the form of a promise, a threat or an act of persuasion etc. Therefore, the performative aspect described by speech-act theory is always present in dramatic dialogue. Even at the most general level this condition of the performative aspect always applies. (1991:6)

In this way, each section of dialogue moves the plot along in some way, and must somehow change the situation, or develop a character’s perspective. TV scriptwriter Steve Wetton makes similar points from a practitioner, rather than an academic perspective, identifying the five functions of dialogue in a scene:

> 1. Push the Story Along.
> 2. Give necessary information.
> 3. Delineate character.
> 4. Have a subtext.
> 5. Set up or pay off a funny line.
>
> (cited in Thompson, 2011:98-99)

As Wetton suggests, even scenes that do not contain major plot point, will still have a purpose, for example, to show characters in a state of equilibrium, which may later be upturned.

In drama, the everyday mismatches, topic changes, repair and overlaps of actual conversation tend to be avoided in favour of cleaner dialogue that progresses the plot more efficiently (Quaglio, 2009). McIntyre observes that the non-fluency features are generally avoided in scripts (2015b:434), citing Richardson who argues that an attempt to accurately recreate naturalistic speech patterns, “would occlude the meaning of particular disfluent utterances as signs of hesitancy, embarrassment, uncertainty, disbelief, and so forth” (Richardson 2010:78, in 2015b:434). What is dismissed as disfluency in naturally-occurring conversation becomes significant in a scripted drama. This was evident in Helen's hesitations in the phone call to her mother, where the hesitations such as, "I, er, it was just to say" and "No! I'm, um, going in a bit later", cannot be dismissed as natural hesitations. Instead they are indications to the audience that – as expressed in the stage directions, she is "covering" her real feelings. McIntyre writes:

> What all of this points towards is that dramatic dialogue can never be a mirror of naturally occurring conversation, since the impetus behind dramatization requires that every characteristic of natural speech be available for use as a stylistic device, whether this be in the service of character or plot. Fictional speech can therefore never be truly authentic in the sense of reflecting naturalistic speech patterns (2015b:434)

Fictional speech may not be "truly authentic", but as McIntyre and Bousfield (2017), Quaglio (2009) and Braber (2018) argue, there are still many points of commonality. McIntyre states:

> It would seem reasonable to suppose on the basis of Quaglio's evidence that there are likely to be significant degrees of similarity between (television) dramatic dialogue and naturally occurring speech. If so, a number of consequences would follow. Such a finding would suggest that the stylistic analysis of dramatic dialogue has the potential to contribute to our understanding of the function and effects of everyday conversation. (2015b:436)

Herman argues that the important question is not so much about whether naturally-occurring dialogue is mirrored in drama scripts, but is more about the mechanics of "the exploitation by dramatists of underlying speech conventions" (1995:6). Herman describes a model of linguistic resources, which dramatists can utilise. Pre-empting the resources model of identity proposed by Grant and MacLeod (2018, 2020), she states:

> The principles, norms and conventions of use which underlie spontaneous communication in everyday life are precisely those which are exploited and manipulated by dramatists in their constructions of speech types and forms in plays. Thus 'ordinary speech' or, more accurately, the 'rules' underlying the orderly and meaningful exchange of speech in everyday contexts are the *resource* that dramatists use to construct dialogue in plays. (1995:6)

This resources model is a useful position from which to study drama scripts as a proxy for criminal identity impersonation. If a person were asked to write a scene of a play, and were also asked to impersonate someone else in a forensic context (for example, if a middle-aged person were asked to impersonate a teenager in a written interaction), it seems likely that they would draw on similar cognitive processes and resources for both tasks. Rey writes, "while the language used in television is obviously not the same as unscripted language, it does represent the language scriptwriters imagine that real women and men produce" (2001:138). As such, the way in which writers use language to impersonate other voices in fictional drama, becomes very relevant for forensic settings. Although different from everyday conversation, scripted dialogue should not be viewed as substandard data, and a poor second to a transcript of naturally occurring conversation. The aim of the analysis is to compare how writers imitate other voices. Using the data of soap opera scripts allows an examination of the same linguistic tools and resources that authors have available to create alternative identities.

2.7.4 Discourse Analysis and Works of Fiction

In addition to comparing the similarities between naturally-occurring conversation and drama scripts, it is also important to consider the ways in which discourse analytical methods, have been applied to the study of play-texts. There is a strong tradition of *discourse stylistics*, the application of discourse analytical techniques to drama scripts, as discussed by Carter and Simpson (1989). Burton (1980) discusses the similarity of "play-talk" to "real-talk" and argues that methods from sociolinguistic analysis and conversational analysis can help readers to interpret play-scripts. Discussing these discourse stylistic methods, she states:

> It is, I think, fairly obvious that if we want to consider play-talk and its degree of similarity to real-talk, then discussing sentences, phrases, alliteration, polysyllabic words and so on, is not going to tell us a great deal. The only possible linguistic level to use as a basis for such analysis is *discourse,* or, even more specifically, *conversation* – as an aspect of discourse. (1980:9)

Burton argues that using discourse analysis methods for drama scripts can provide evidence for the intuitions that audiences or readers may draw about characters. For example, Burton's analysis of Pinter's short sketch, *Last To Go*, shows how Pinter exploits the norms of conversational structure for comedic effect.

Bennison (1998), like Culpeper (2014), acknowledges that for audiences, a large part of the interest for audiences lies in the 'personalities' of the characters in the play. He quotes Downes (1988), who argues that we interpret characters in a drama in the same way that we interpret real people. Whilst noting that Downes' position was not universally accepted by critics, Bennison suggests that it can be a helpful starting point, proposing that, "it follows that the methods of analysing conversational behaviour in the real world are also readily applicable to that of the dramatic world" (1998:68).

Simpson (1989) and Bennison (1998), like Burton (1980), use analytical methods from discourse analysis and pragmatics to examine how audiences infer 'personalities' of characters from dialogue. Simpson (1989) uses Brown and Levinson's theories of politeness to track the change in power dynamics between the Professor and the Pupil in Ionesco's *The Lesson* (see 3.5.2). These two studies are able to explain and give evidence for the changes in characterisation. Analysing the conversation of Professor Anderson in Tom Stoppard's *Professional Foul*, who gets caught up with political events while at an academic conference in Prague, Bennison uses discourse analysis show how Anderson's character traits, such as his pomposity and urbanity, are inferable from his conversation. Bennison analyses features such as conversational turn-length, turn-taking and topic-shift, and then analyses Anderson's behaviour using pragmatic frameworks (Grice's Co-operative Principle, Brown and Levinson's politeness and Leech's Politeness Principle.). He demonstrates how discourse analytical approaches can use replicable methodologies to illustrate how characterisation is created through interaction.

Bennison observes that a key benefit of this method is that it "provides us with a relatively precise methodology for dealing with perceivable changes in character" (1998:81). This observation suggests a possibility and a problem for authorship analysis. In his analysis of *Professional Foul*, Bennison argues that by the end of the play, Anderson becomes more direct and less polite (1998:69). The aim of Bennison's study was to analyse Anderson's character through his conversation, so observing and explaining changes in character is a useful step. However, as a method for authorship analysis, there is no suggestion that the methodology would discriminate between changes in character development, and changes in authorship. Whilst the value of discourse stylistics has been well established, using such methods on drama scripts to explore authorship attribution is, I believe, a new approach.

2.8 Conclusion

This Literature Review has drawn together the key areas of research which inform my research and my epistemic position. I have reviewed methods of authorship attribution, both quantitative and qualitative, and discussed core concepts of style and sociolinguistic identity. I explored the difference between authorial style and character style and have defined how I use these terms in my thesis.

The next chapter also reviews relevant literature, but is focused on selected readings which have informed the analytical methods used in each separate study, rather than those topics which contribute to my thesis as a whole.

3. Analytical Methods

3.1 Introduction

These next two chapters introduce and discuss the methodology used in my thesis. Chapter 3 reviews selected academic literature specific to the analytical methods used in the three main studies in Chapters 5 – 7, and Chapter 4 describes the Methodology of each study and the corpora used.

In this chapter I summarise some of the relevant quantitative authorship attribution methods (discussed more fully in 2.3.4). This is followed by a discussion of the literature relevant to analytical methods used for the three character studies in Chapter 6. These character studies each focus on a different character and a pertinent aspect of that character. Finally, I review the literature on pragmatic noise relevant to the analytical methods used in Chapter 7.

In Chapter 4, I describe the overall dataset as well as the sub-corpora used in each study. I explain how each corpus was prepared, including the software used, and explain the reasoning behind these choices. I set out the methodology for each of the three separate analytical chapters, stating the rationale behind these choices, and conclude with a brief review of my position as a researcher, and ethical considerations of this study.

3.2 Analytical Methods for Chapter 5

The first analytical chapter addresses the first of my research sub-questions, exploring quantitative, structural-level features of language. This section discusses key literature on quantitative attribution methods which have informed my methodology.

3.2.1 Word-n-gram-based test

The first test carried out is the word-n-gram-based test. As this approach has already been discussed (2.3.5), it is only discussed briefly here. Although the discriminatory power of word-based-n-gram tests has been shown to be lower than character-based-n-gram-tests, Grieve et al. argue that, "word-level n-grams are less common, more distinctive, and more interpretable" (2019:508). Interpretability is a crucial part of being able to analyse the results to determine why some n-grams might be used more frequently than others for certain characters.

The remaining three tests, based on the categories of textual measurement in Grieve (2007) are: Average Word Length, Average Turn Length and Type-token ratio, which are discussed in this next section.

3.2.2 Average Word Length

The second variable analysed in Chapter 5 is Word Length. Grieve (2007) outlines two measurements for word length: the first is Average Word Length; the second is a Word Length Distribution study (frequency of one-letter words, two-letter words etc.). Here, the first of these measurements is used to represent the category Average Word Length.

Nini (2018), investigating authorship profiling, cites early studies, including Bernstein (1962), Kitson (1921), which found that increased average word length correlates with higher social status, a finding also supported by Bromley's (1991) analysis of descriptive essays and Berman's (2008) study on narrative speech samples (Nini, 2018:43-44). Bernstein compared working class and middle-class teenagers, and found that working class subjects used shorter word lengths. He showed that there was a difference between the two social classes, which was not accounted for by the IQ of subjects, but by social class alone. In an even earlier study, Kitson (1921) observed differences in average sentence length and average word length between magazines. Nini observes that:

> Significant differences in terms of average sentence length and average word length between magazines influenced the kind of readership and therefore the kind of social groups that would read a certain magazine. These findings suggested that there was a correlation between the average sentence length and word length encountered by a social group and its social status." In these studies, a higher average word length was also associated with increased age, and with male speakers. (2015:60)

Whilst significant changes in society mean the specific findings from these early papers may no longer apply, the studies serve to show that word length can vary between different demographic groups, even if the specifics of those groupings might change over time. As such, it is interesting to see if the scriptwriters vary the word length depending on the characters they are voicing.

3.2.3 Average Turn Length

As with Average Word Length, the Average Turn Length can be associated with higher levels of education, grammatical complexity and subordinate phrases, which are themselves features associated with higher social or educational status. Hunt (1983) examined essays and newspaper articles, analysing average sentence length, and Poole (1979) examined structured interviews, including average sentence length. Reflecting on these studies, Nini writes:

> Syntactic complexity, measured through average sentence length or number of dependent clauses per sentence, has been found to be correlated with class by Loban (1967) in both oral and written texts, Poole (1976) in life-forecast essays, Johnston (1977) in experimental elicited narratives, Poole (1979) in structured interviews, Labov and Auger (1993) in sociolinguistic interviews, and it was also found in Kemper et al.'s (1989) and Mitzner and Kemper's (2003) studies on syntax and ageing to be a good predictor of level of education. (2018:43)

There have been criticisms of sentence length as a predictor of sociolinguistic profile. Rudman (1998) observes, that "even as early as 1903, Robert Moritz pointed out major flaws in the 1888 "Sherman principle" of sentence length as an indicator of style and authorship" (1998:352). As with the literature on Average Word Length, some of these studies are not recent, so it is problematic to apply the sociolinguistic profiles to contemporary data.

Average Turn Length has been analysed on fictional texts. Analysing turn-length in Stoppard's *Professional Foul*, Bennison (1989) notes that Anderson's turn-length alters over the course of the play, as Anderson changes in response to plot developments. Bennison observes that Anderson's difference in turn length between the beginning and end of the play reveals a marked contrast between his language when weary, compared to when he is full of enthusiasm (see 2.7.4). This suggests that Average Turn Length may be a feature which varies within the corpora of individual characters, and may be affected by context, making it a less stable feature for analysing authorship.

3.2.4 Type-token Ratio

Nini includes a number of studies which point to a higher Type-Token ratio as being suggested of higher social status and higher educational levels. Bromley (1991) analysed descriptive essays' average word length; Berman (2008) studied narrative and expository speech samples and texts and found average word length in syllables; lexical density; proportion of words from Romance and Germanic origins all pointed to higher levels of literacy. According to Nini:

> The literature would have predicted that age is also positively correlated with the number of long words (or average word length) and number of rare words, that is, with variables that measure how many rare words of the English language are used in the text. Instead, the only variables that showed a positive significant increase with age were type-token ratio and Baayen's P. T. (2014:143)

A further observation was that:

> The pattern related to vocabulary size is the most consistent pattern for level of education found in the FMT corpus. Table 5-7 shows that all the variables related to average word length or to rarity and sophistication of vocabulary such as Advanced Guiraud 1000 show a significant and consistent increase with level of education. (2014:151)

Culpeper (2014) also notes the association that researchers such as Bradac have found between lower lexical variety and lower socio-economic status and communicator competence. Based on this, I would expect that those characters who are portrayed as intelligent, or of a higher social standing, would have higher levels of vocabulary richness.

3.3 Analytical Methods for Chapter 6: Three Character Studies

3.3.1 Introduction

Chapter 6 adopts a qualitative approach to authorship attribution, focusing on three fictional characters, and uses a different linguistic domain to analyse each. The three domains are lexical choice, dialect, and (im)politeness strategies; and the three characters are, respectively: retired History professor, Jim Lloyd; womanising Glaswegian agricultural worker Jack ("Jazzer") McCreary; and the village's self-appointed organiser-in-chief, Lynda Snell. This study addresses the second of the research sub-questions, asking whether writers are able to use consistent intra-character features. The first of these character studies explores the lexis used by Jim Lloyd, and 3.3. reviews selected literature about lexical choice and characterisation, which informs the analytical approach taken in my thesis. The character study on Jim Lloyd's lexis is further sub-divided into three separate tests which analyse his vocabulary. Subsequent sections review key literature informing the analytical methods used for the second and third character studies, on dialect, and (im)politeness respectively.

3.3.2 Lexis and Characterisation

The first character study investigates the role of lexis in authorship attribution. Three separate analyses are carried out within this character study. These are: lexical richness, key word analysis, and a comparison of Latinate versus Germanic words. This section reviews selected academic literature on lexis, characterisation and authorship attribution, including Culpeper (2014), which offers a comprehensive view of the role of lexis in characterisation in drama, and McIntyre's key word analysis of *Reservoir Dogs* (2010).

The importance of lexis has long been recognised in both authorship analyses, and in literary studies of characterisation. In authorship analysis research, Gibbons (2003), Winter (1996) and Coulthard (1994) all pointed to the importance of vocabulary choice. Coulthard discussed the importance of "unlikely vocabulary choices" (1994:38), although this raises methodological questions about finding a replicable threshold for "unlikely". A further way to analyse the distinctiveness of lexical choice is by studying collocations of words. Coulthard (2004:440) argues that short collocations may occur frequently enough to be useful in attribution tasks. More recent studies such as Wright (2014, 2017) and Larner (2014) have considered the importance of fixed phrases and n-grams. Wright discusses the use of content words in authorship attribution tasks:

> In stylometric approaches to authorship analysis, content words have at best been avoided and at worst received unsubstantiated criticism over the last fifty years. Within a research tradition which focuses on relative frequencies of individual words this is not surprising, as on their own all they reveal about an author is that they write more or less about a particular topic than another author. (2014: 136)

Instead, Wright argues for the possibilities of using collocations of words to explore how these content words are used. Another issue regarding the use of lexical choice in attribution tasks is that lexical choice may not be a strong marker of authorship diachronically: as Hoover argues, authors can learn new words, and also may stop using words or even forget them (2003:157). Despite these methodological issues, lexical choice remains a significant tool in authorship analysis. Larner states that:

> Lexis is generally an accepted marker of authorship in the field of forensic linguistics … whilst acknowledging that focussing on the open class set of lexical items is not necessarily the most effective marker of authorship, nor indeed the only marker of authorship. (2014:9)

Larner argues that open class lexical items are not the most effective markers of authorship in quantitative attribution studies. However, in drama, lexis is an important way in which audiences interpret characters, as Culpeper argues:

> Intuitively, it is reasonable to suggest that lexis plays a significant role in shaping people's impressions of others. For example, the tendency to use formal lexis may give the impression that someone is rather aloof or pompous, informal lexis that someone is 'down to earth'. However, research undertaken to examine the relationship between lexis and personality or character is patchy. (2014:182-3)

Addressing the ‘patchiness’ of research into characterisation and lexis, Culpeper builds on the work of Rimmon-Kenan (1983: 59-70) and Pfister (1991: 124-6, 183-95) to set out a comprehensive list of textual features which are important in the creation of fictional characters, both linguistic (such as lexical richness) and non-linguistic (for example, appearance and voice quality) which all contribute to characterisation. Under the heading ‘lexis’ Culpeper (2014) lists five features, which are summarised in the following sections (3.3.3 – 3.3.7).

3.3.3 Germanic versus Latinate lexis

This method explores the relationship between lexis and characterisation by considering the etymological origin of the words used by a character. The use of Latinate words compared to Germanic words is another of Culpeper’s textual clues about characterisation which is concerned with lexical choice (2014:183). Culpeper discusses the relationship between characterisation and the etymological origin of the words they use, and argues that different etymologies correlate with different stylistic dimensions:

> The more common words of English, particularly the words of speech, tend to be Germanic in origin, whereas Latin words tend to be rare and appear more often in written language.

> Germanic words are more likely to be used in informal, private contexts, whereas Latin words are the words of formal, public occasions. Germanic words tend to be simple, often words of one syllable, whereas Latin words are usually polysyllabic (2014:183).

Culpeper explains that Germanic words tend to be used for concrete items such as “wood”, “earth” or “house”, compared to words of Latin origin, which are often used for abstract concepts. To illustrate this concept, Culpeper compared the language used in a speech by the Nurse in *Romeo and* Juliet to a speech by Lady Capulet and found that Lady Capulet used predominantly Latinate words, whilst the Nurse used none, and instead relied heavily on words of Germanic origin.

3.3.4 Lexical Richness

This lexical feature is only analysed briefly because lexical richness, also known as type-token ratio, has been considered already in relation to Chapter 5. Discussing the relevance of lexical richness to fictional characterisation, Culpeper states that:

> The richness or diversity of lexis within a person's or character's speech can suggest certain characteristics. Some research has been undertaken in this area. The conclusion seems to be that ‘Generally, lower diversity results in receiver judgements of lower communicator competence, lower socio-economic status, and higher anxiety' (Bradac 1982: 107; see also Bradac 1990: 396-7). (2001:188)

This suggests that a varied vocabulary can be indicative of a character who is perceived to be intelligent and of a higher socio-economic status. In an analysis of *Romeo and Juliet*, Culpeper found significant differences in the lexical density of the Nurse compared to Capulet, and also between Mercutio and Capulet, and Mercutio and the Nurse (2014:188-189). This suggests lexical density is a feature which varies when authors write dialogue for different characters.

3.3.5 Surge Features

This dimension is one of the dimensions of lexical characterisation listed by Culpeper, but is discussed only briefly because it does not form part of my analysis. Culpeper defines surge features as being concerned with “personal affect” (p.190), which includes feelings, emotions, moods and attitudes. Citing Taavitsainen (1999), Culpeper explains that these outbursts of emotions, which can include exclamation, swearing and pragmatic particles, are called surge features. Since exclamations and pragmatic particles are explored in depth in Chapter 7, this feature was not included in the three character studies in Chapter 6.

3.3.6 Social markers: Terms of address and second person pronouns

Culpeper argues that "terms of address, including vocatives and pronouns, can be an important means of signalling social information" (2014:193). Present-day terms of address can be endearments (e.g. *darling, love*), family terms (e.g. *mummy*), familiarisers (e.g. *mate*), first names, title and surname and honorifics (e.g. *madam*). Culpeper then compares this to Elizabethan terms of address, (e.g. *my Lord*), and also the distinction between *you* and *thou*, illustrating links to characterisation with examples from Shakespeare. As Culpeper himself notes, this summary does not fully describe all the nuances of terms of address, and does not take account of the usages of these terms (p.194). This category is potentially very interesting for a forensic analysis, and greetings have been explored as potential markers of authorship. For example, Wright (2013) compares greetings and farewells in the Enron email corpus, and finds certain forms of address in email openings, such as "buddy", were "either entirely individuating of one of the traders or were shared between two or more traders, but used far more consistently by one" (2013:21). Whilst this could be a promising area for investigation, it was discounted because of the audio-only medium for which these scripts were written. In *The Archers,* characters tend to greet each other by name to identify themselves for the benefit of the audience, so the variety of terms of address is more restricted, and any findings would be highly specific to the genre of audio drama.

3.3.7 Keywords

Culpeper's final dimension, in his discussion of lexis and characterisation, is keywords. Culpeper introduces the *Keywords* function in Mike Scott's *WordSmithTools* (1999), which carries out a statistical comparison of the words in a corpus, compared to the words in a bigger reference corpus to look for unusually frequent words. Demonstrating the technique, Culpeper discusses Juliet's most key keyword of "if" and proposing that this reflects Juliet's state of anxiety throughout the play. Culpeper (2014:199) urges caution with a keyword analysis, as the process can generate 'meaningless' results. Culpeper's analysis included words which had at least four occurrences. He explains his methodology, stating that he:

> examined the function and context of each instance of a keyword, in order to validate and account for the results. This step is not required by Enkvist's definition, but was necessary, since not all keywords reflected character (some keywords, for example, arose as a result of a particular context). (2014:200)

McIntyre (2010) uses a keyword analysis to investigate how dialogue shaped characterisation in Quentin Tarantino's *Reservoir Dogs*. Using a corpus stylistic approach McIntyre was able to, "isolate distinctive features of character dialogue" (2010:163). He found that:

> The analysis of keywords, key semantic domains and n-grams indicates some of the differences of character among the criminals in *Reservoir Dogs*, and suggests which particular aspects of their speech work as characterization clues. (2010:180)

As with Culpeper's lexical density analysis of *Romeo and Juliet*, this approach from the field of stylistics is premised on the idea that writers adapt their language to represent the different voices of distinctive characters within a fictional work. These alterations produce statistically different results for different characters. The ability to create diverse characters, whose voices are measurably distinct from each other, and whose voices can change in response to plot developments, can be seen as testament to the creative skill of the playwright, but presents a difficulty for the forensic linguist if stylometric techniques cannot distinguish between inter-author variation and inter-character variation.

Of the "salient dimensions of lexical variation" (2014:183) outlined by Culpeper, the three methods of analysis I used in this study are Germanic versus Latinate lexis, lexical richness and a keyword analysis. The reasons why Surge Features and Social Markers were excluded is discussed above. A full methodology for these three studies is set out in Chapter 4.

3.4 Character Study 2: Jazzer's Dialect

The second character study explores the use of dialect. In *The Archers,* there are a number of characters who are not from Borsetshire, the fictional region where the drama is set, including Ruth Archer, a Geordie farmer, Ian, an Irish chef, and Jazzer, a Scottish milkman. In Chapter 6, I examine the linguistic features which produce Jazzer's dialect, to explore how the six scriptwriters write dialogue for characters who have a different accent and dialect from their own. The following literature review is necessarily quite lengthy because it reviews scholarship on the use of dialect in drama, and sociolinguistic theory about indexing dialect, but also provides specific detail on features of the Glaswegian dialect, which is needed for an analysis of the way in which Jazzer's dialect is realised by the scriptwriters.

3.4.1 Dialect in Drama

Dialect is a very notable type of language variety and its portrayal in fiction has been studied separately. Hodson (2014) describes dialect as "a variety of English which is associated with a particular region and/or social class." She adds that "the representation of dialects in both film and literature primarily means the representation of different spoken, rather than written varieties" (2014:1). Dialect can be described at three linguistic levels: pronunciation, vocabulary and grammar (as described in Culpeper (2014), and Hodson (2014), for example). Culpeper explains:

> Traditionally, the dialects that have received most attention are regional (the dialect spoken by the people of a particular geographical area) and social (the dialect spoken by the people of a particular social group). The term dialect refers to a variety of language characterised in terms of pronunciation, grammar and lexis; the term 'accent' can be understood to refer to a sub-set of dialect in that it refers to a variety of language characterised in terms of pronunciation only. (2014:166)

Choosing to write a character's lines using dialect, whether using any or all of these linguistic levels, is significant part of creating that character's identity. Short observes:

> The majority of English literature is written in Standard English, which thus counts as the norm. Characters speaking non-standard dialects in novels or plays stand out from the rest, and if a poet chooses to write in a non-standard form this often counts as a socio-political act of some kind. (1998:7)

As Short argues, the use of non-standard dialogue is revealing about characterisation. Writers using dialect may incorporate the grammatical and lexical features of that dialect in the text, and can also choose to represent pronunciation orthographically in a script: for example, to write “my” as “me” if that is how it would generally be pronounced by a particular character.

Hodson notes that in addition to the background factors which “govern” a person’s variety of English, “the context within which he or she speaks and the purpose of the speech will also influence the variety used” (2014:3). Researchers have explored the ways that speakers’ dialects are not rigidly fixed, but can be exaggerated for dramatic, often comic, purposes, for example (Braber, 2017) and (Coupland, 2001). Coupland, discussing the way radio broadcasters can adopt elements of Welsh dialect into their on-air persona, argues that, “invoking the idea of stylization in relation to dialect implies seeing dialect as performance rather than behaviour, and (like all sociolinguistic styling) as social practice rather than as variation” (2001:348). Applying this position to the analysis of a fictional character, it is possible to argue that a character (in this case, “Jazzer”) is able – through the scriptwriters – to perform certain aspects of his Glaswegian identity: Jazzer does not have to have a single ‘level’ or ‘type’ of Glaswegian speech patterning, and can ‘style-shift’ unilaterally within a conversation, as Hodson describes, using *Small Island* an example (2014:9).

Dialect in fiction is not neutral. Short asserts that dialect is a “socio-political” choice. Montini and Ranzato (2021) refer to the “the ideological trap that the transcription of nonstandardness may represent” (2021:1). Stockwell also adopts this position, through the lens of cognitive stylistics:

> The representation of accent and dialect in literature has a long history, both as characterisation and narratorialisation. Characters’ direct speech (and sometimes direct thought) can be presented in selected dialectal phrases or with non-standard spellings to indicate their speech patterns. Since we know that accent and dialect invoke schematic social stereotypes about those speech groups, that character accumulates some of those traits in the mind of a reader. (2020:362)

This argument suggests that choosing which dialect to use for a character is not a random decision: the choice to make Jazzer as a Glaswegian character in itself signals information to the audience about his characterisation. Braber argues that Glasgow, which has historically had high figures for unemployment and poverty, has been stigmatised:

> As the concept of a ‘Scottish’ identity has been shown to be very important to Scottish people, the existence of a strong sense of community in Glasgow is not altogether surprising. Although heavily stigmatised as a city by outsiders, its inhabitants have retained a strong sense of belonging. The stigmatization has led to Glaswegian being branded as ‘slovenly’ and ‘degenerate’ (Andersson and Trudgill, 1990), and previous research (Braber and Butterfint, 2008; Braber, 2009) has shown that Glaswegian is seen as unattractive, even by many of its speakers. (2017:268)

By the time the scripts that form my dataset were written, Jazzer was an established character within the show, so it is not relevant to speculate whether there was a decision to introduce a womanising, “jack-the-lad” character, who was later cast as Glaswegian; or whether there was a decision to introduce a Glaswegian character, whose working-class and somewhat feckless qualities were either part of the genesis of his character, or were fixed at a later stage in the production process. Regardless of his inception, there is an association between Jazzer’s characterisation and the Glaswegian stereotype of working class “degenerate”. Exploring the decision to introduce a Scottish character who speaks in dialect cannot form part of this analysis, because the decisions were taken years before the earliest scripts in the data were written. However, the ongoing linguistic realisation of the character Jazzer remains relevant, and is the subject of my analysis.

An additional note about the appearance of dialect in fiction is that speeches written in dialect often contain an amount of speech written in Standard English. As Stockwell (2020) and others, (including Kozloff, 2000 and Hodson, 2014) have noted, attempting a full phonetic transcription could become unreadable. Stockwell argues that:

> Writers pick out the most obvious features of a dialectal variety, using them as headers to instantiate a reader’s schema of that dialect and speech community. These features are almost always what Labov (1972, 2001) called stereotypes and markers, rather than the indicators which are largely much less salient signs of a particular accent, known mainly to expert self-aware speakers and sociolinguistic experts. (2020:363)

In drama scripts, this is slightly different, because the writer is not simply giving the reader a flavour of the dialect for them to imagine when reading the book: instead, the lines will be heard via the actor. This gives scriptwriters an additional choice: some may use eye-dialect to guide the actors’ pronunciation; others may avoid it, assuming that the actor will find such guidance unnecessary. My analysis reviews the various ways in which the scriptwriters create Jazzer’s dialect, exploring the different aspects of dialect that the writers focus on, and measuring the use of dialectal items by each writer.

3.4.2 Writing "otherness"

Ruzich and Blake's (2015) analysis of the 2009 novel *The Help* explores the way that white writer Kathryn Stockett created African-American characters. *The Help* is a story about African-American women employed in white households in 1960s Mississippi. The novel is told through the voices of three narrators: a young white woman and two older African-American maids. To evaluate the way that Stockett creates African-American identities through two of her protagonists, Ruzich and Blake compared the first 500 words from each new character in the book, analysing any linguistic features which were not written in Standard American English, or which were considered particular to an American dialect. They looked for three types of linguistic features:

- Eye dialect (gonna, that kind a thing)
- Vocabulary features (ain't, yonder, chilluns)
- Phrase or sentence-level grammatical features (that old woman eat two butter beans and say she full)

They found that white characters used dialect markers once in every 100 words, while black characters used them once in every ten words on average, and Skeeter, the white protagonist, only once in 285 words. They observed that the white characters would presumably have heavy Southern accents, yet their dialogue is portrayed in Standard American English. They argue that this disparity depicts the black women as outsiders. In terms of authorship analysis, this suggests that writers can be subjective in what they hear as being "other": in *The Help*, Stockett portrays black women in a way that distances them, whilst remaining 'deaf' to the way in which white, Southern accents and dialects would also differ from Standard American English. The authors argue that Stockett's creation of fictional characters is highly subjective:

> The language of Stockett's novel (or of any novel) performs not the author's own voice and racial identity, but her constructed understandings of others' identities. Additionally, readers of the novel also participate in the construction of those fictional identities, bringing their own language use and experiences to bear on their understandings of the text's language and its message about race, class, history and culture. (2015:537-8)

This relates to Coupland's observation that there has been a tendency to perceive RP as a "supra-dialect", the standard form, from which all others deviate. He argues that actually all dialect is relational and subjective (2004:288). Portrayal of dialogue and other identities does not start from a neutral position but can be heavily influenced by the writer's own perspective, and perceptions of sociolinguistic 'norms'.

3.4.3 Indexing Dialect

Ways in which individuals and groups can perform sociolinguistic identity has been discussed in 2.6.1. Here, I reflect specifically on the use of dialect in fiction. As Stockwell and Short observed, using dialect in literature can be a socio-political act. Similarly, sociolinguists have argued that speakers can deliberately use, or exaggerate the strength of a dialect to form an in-group membership. Douglas argues that, "Although much usage of Scottish English linguistic features is covert (i.e. speakers do not realize it marks them as Scots), there is also a strong tradition of overt usage with people deliberately and knowingly choosing to use Scots linguistic features, often as a way of asserting their Scottish identity (Aitken 1979, 1984b)" (Douglas, 2020:23).

The following political tweet and its predominantly critical responses provides a (real-life) example of a speaker attempting to use dialect to index a certain kind of persona; in this case, Humza Yousaf, of the Scottish National Party, was tweeting during an election campaign visit to Pollok, a town near Glasgow. Yousaf's tweet read:

> Braw day for it! Thank you to SNP activists right across Glasgow Pollok for giving up time on such a gorgeous day to get our positive message through the letterboxes! #BothVotesSNP" (@HumzaYousaf, twitter.com, 03/04/2021).

The candidate's use of the overtly Scottish word, "braw" drew overwhelmingly negative responses specifically criticising its use as being a parody of Scottish English. This political tweet and sample of the responses encapsulates the way a linguistic strategy to invoke a certain type of persona – in this case to create in-group membership with the voters of Pollok – can miss the mark with its audience. In a drama script this could be the writer indexing a certain dialect, as Stockwell suggests, to invoke schematic social stereotypes about a character. Also, the writer could show through dialogue that it is the character's choice within the fictional world, to either emphasise or downplay their dialect in a given context.

3.4.4 English in Scotland

This next section summarises extremely briefly the varieties of English language which exist in Scotland, before focusing specifically on the Glaswegian dialect. Douglas (2020) divides the languages of Scotland into two main strands: firstly, Gaelic, a Celtic language (and therefore outside the scope of Douglas' study), and Scottish English, which is described as "the distinctive localised variety of British English native to Scotland." Scottish English is then divided in two further strands: a variety Douglas terms Scots (SC); and, second, Scottish Standard English (ScSE), which was "the result of contact with the standardized form of English English during the eighteenth century" (2020:18).

Douglas outlines the political and historical background to the development of Scottish English, which encompasses regional and social varieties of Scottish on a linguistic continuum, "ranging from Scots (sometimes called Broad Scots or Scots dialect) at one end to ScSE at the other" (2020:21). Code-switching and shifting between these different varieties of dialect is common, with Douglas arguing that:

> Individuals, taking account of external factors such as context of situation, education, and social class can move along the continuum in either direction, but some people will inevitably have a stronger attraction to one pole than the other. (2020:21)

This observation aligns with Coupland (2001), Hodson (2014) and Braber (2017) who all assert that dialect can be performed, and over- or underplayed according to context. Fuller discussions of the history and varieties of Scottish English are discussed in Aitken (1979), Macafee (1992), Hagan (2002), Braber and Butterfint (2008) among others.

Despite the co-existence of multiple languages in Scotland, the language of the Scottish education system is Scottish Standard English, so when Scots write in English the Scottish English terms are generally used rather than the Scots versions (for example “could not” rather than “couldne”). Douglas explains how this has created a disjoint between written and spoken forms of the language:

> The written and spoken varieties are not as closely entwined as one might think; for example, much more Scots is spoken than is written, and few Scots are practised writers (or even readers) of SC. Literary Scots bears little resemblance to the spoken Scots one hears, and it is a curious anomaly that those few individuals who do write in Scots are usually highly educated and/ or middle-class – the very people one would least expect to hear using Scots in speech. (2020:23)

This is relevant for both stylistic and forensic linguistic analysis. When writers wish to convey a Scottish dialect, the lack of standardised written form for many Scots features which are normally encountered aurally rather than in written form (for example words such as “doesne” or “isnae”) means that spelling is likely to vary between writers, which could potentially be used to discriminate between different writers. In Chapter 5, the quantitative analysis of vocabulary found that non-standard spellings of certain words, including pragmatic noise and dialect items varied between authors, and some tokens were distinct to individual authors. Douglas’ observation also links to an important point about dialect and indexicality: the use of Scots dialect and grammar can be indicative of social class, but because of the ironic anomaly Douglas discusses, it is entirely possible that non-Scottish people (in this case, the scriptwriters) access their ideas of Scots dialect through fiction, from sociolinguistic groups who do not themselves use that language in their day-to-day lives.

3.4.5 The Glaswegian Dialect

Braber (2017) describes Glaswegian as “a distinct, often stigmatised variety and one which holds many stereotypes both for those in the city and outside its boundaries” (2017:269). She observes that, as is the case with all speech communities, there are multiple linguistic varieties of Glaswegian. Macaulay’s notable sociolinguistic study of varieties of Glaswegian spoken by different social groups is one such example of this. Stuart-Smith (1999) defines two varieties of Glaswegian. These are Glasgow Standard English (GSE), the Glaswegian form of Scottish Standard English, spoken by most middle-class speakers and Glasgow vernacular (GV), mostly used by working-class speakers, which has its own distinctive slang, and also incorporates Irish English influences. Braber points out that Glaswegian speakers can and do move along the linguistic continuum of Glaswegian, which runs from 'broad' Scots to Standard Scottish English. She argues that:

> Certain varieties on this continuum are more stigmatised than others (and these can be correlated with social class, e.g., the greatest stigma appears to be attached to the varieties more usually found in the lower socio-economic groups). Speakers can move along the continuum depending on formality and situational context. (2017:268)

As with all dialects, Glaswegian is subject to change, for example, Stuart-Smith et al.'s, *Talkin' Jockney*, (2007) comments on the growing influence of London on Glaswegian. The character Jazzer left Glasgow in around 2000 and has had limited contact with his homeland since then, so it is plausible that his particular variety of Glaswegian would be reflective of late twentieth-century Glasgow, rather than more recent influences. It is of course also likely that spending a significant length of time among non-Scottish people (ten years by the time the earliest of these scripts is written) would have softened and reduced the strength of his dialect. It is impossible to incorporate the multiple varieties of Glaswegian here, but a brief summary of some key features of Glaswegian follow, using the three levels of dialectical language outlined by Culpeper and by Hodson as a basis: pronunciation, grammar and lexis.

3.4.6 Pronunciation

Pronunciation is perhaps the least relevant of the three levels of language because the analysis is concerned with written scripts rather than spoken recordings. Some of the writers use eye dialect (for example 'wi'" for "with"), but others do not. Whilst this could be used as a marker to determine which scriptwriter wrote a particular scene, on a deeper level it does not necessarily tell us much about *how* the different authors are hearing and then reproducing a Glaswegian identity. Writers may avoid eye dialect simply because there is no need to offer pronunciation guides when it is already established that the script will be performed by a Glaswegian-born actor. It could well be the case that if they were writing the dialogue for a drama where casting does not take place until after the script is finished, the writers might choose to use eye dialect, or that they might chose to use eye dialect in a novel. Even so, looking at writers’ renditions of pronunciation is useful because it is informative about which words the writers hear as being “in dialect”, as Stockwell (2021) discussed, when examining dialect as a form of social deixis.

3.4.7 Grammar

Eunson (2019) outlines the recent resurgence in interest in teaching Scots as a language in schools, arguing that, “Scots language has for so long been seen as either something of the past, or as a dialect of English, or as being simply “wrong” or “bad” or “slang” – or many other derogatory terms that led to Scots being marginalised from both education and wider society” (*TES*, 2019). As part of this recent interest in teaching Scots in schools, Education Scotland has published a feature list of Scots grammar. The key features are produced in Table 2, with comparison to their English equivalents. Following a humanising approach to characterisation, it is worth noting that Jazzer would have finished his education and moved to Scotland before this resurgence of interest in Scots was introduced.

Table 2: Scots Grammar compared to English

(adapted from FeaturesOfScotsJan2017.pdf (education.gov.scot))

Grammatical Feature	Scots	English
Forming Negatives	negatives are formed by adding –*nae* or –*na* to auxiliary verbs. This varies between different regions. E.g. *cannae /canna; dinnae /dinna; didnae / didna; havenae / havena / hinnae / hinna; michtnae / michtna*. Four possibilities in Scotland: (1 ') *He isnae coming He's no coming* [ScE] *He isn't coming He's not coming* [=Standard Scottish English] (Pust, 1998)	In English, *not* would be used, usually contracted to, for example: *can’t you*; *haven’t you*; *won’t you?*
Forming Negatives in Questions	In questions, Scots forms negatives with *no*. In North	*Not* would be used, usually contracted, e.g. “*can’t you*;

	Eastern Scots, *nae* would be used. E.g. *Can ye no gie's a haund? Have ye no seen the film?*	*haven't you*; *won't you*?"
Forming negatives after contractions	Scots forms negatives with *no* after contractions. In North Eastern Scots, *nae* is used. Examples are: *She's no weel. I'm nae fussy.*	In English, *not* would be used: *I'm not fussy* or a different construction would be used: *She's unwell/she's ill*.
Present Participles	Present participles are formed by adding *–in*. In verbs that end in *–le*, the ending is *–lin*. E.g. *bletherin, greetin, hingin, tummlin.* There is no need for an apostrophe at the end of the word in Scots.	The present participle tends to end in *ing*.
Past Tense of Weak Verbs	The past tense of weak verbs is formed in Scots by adding *–it* or *–t*. In verbs that end in *–le*, the ending is *–elt*. E.g. *cleekit, gruppit, hingit, lowpit, blethert, gaithert, kent, scunnert*. Weak past tenses are formed for verbs which do not change their stem for forming the past participle.	The past tense of weak verbs tends to end in *–ed*
Past tense of strong verbs	Scots strong verbs change their vowels to form the past tense. E.g.: *buy* > *bocht*; *drive* > *drave* or *dreeve* (NE); *fecht* > *focht*.	There are verbs which do this in English too: *run* > *ran*.
Demonstrative pronouns	Three different demonstrative pronouns, depending on how far away the thing being 'pointed at' is: *this* (nearby *that* (middle distance) *thon* or *yon* (far away). *Yon* can reference something not present in time or space. *Thir* (nearby – English *these*) *Thae* (middle distance – English *those*) *Thon* or *yon* (far away, including out of sight)	Only *this/these* and *that/those* are commonly used. *Yonder* used to be used more commonly in English, in the way that *thon/yon* are used in Scots now.
Plurals	Some Scots nouns have distinctive, irregular plurals. Examples are: *coo* > *kye*; *ee* > *een*; *shae* > *shuin* or *sheen*. Most Scots nouns have plurals formed by adding an '*s*', as in English.	There are fewer examples of irregular nouns in English, but they do exist: *ox* > *oxen*.

Definite articles and possessive pronouns	Scots uses the definite article and possessive pronouns in distinctive ways. For example: *I'm comin doon wi the cauld*; *She's gaun tae the scuil*; *I'm awa tae ma bed*; *That's for yer Christmas*.	English tends to use the indefinite article: *a cold*; or miss the article altogether: *going to school*. Similarly, the possessive pronoun is usually not present in English: *I am going to bed*.
Double modals	In some areas of Scotland, Scots uses double modals. For example: *I used tae cud dae that*; *Ye'll no can see her the day*; *We micht can get a bus* Most dialects of Scots follow the English example.	Standard English uses one modal verb only: *I used to be able to*; *You cannot see her*.
Northern subject rule	Some Scots speakers follow the Northern subject rule, e.g.: *My feet's gey sair.* This rule states that where the subject is a noun or a personal pronoun not next to the verb, the third person singular verb is used, regardless of person.	In English this usage is seen as bad grammar (although it is often used in speech).

Double Modals are a relatively rare Scots grammatical feature. Morin et al. write, "DMs [double modals] are notorious but paradoxically elusive features found in restricted varieties of Southern American English, and even more rarely in some dialects of English in Scotland, England, and Ireland" (2020). Glaswegian and Scottish grammar is described in further detail in Aitken (1979), Müller (2010, 2011), Douglas (2006, 2019) and Hagan (2002).

3.4.8 Lexis

Müller (2010) observes that, "Glasgow has undergone significant loss of distinctive traditional lexis and few words unique to that area remain in current circulation" (2010:156). One of the reasons it is problematic to specify what constitutes Scottish lexis, and subsequently Glaswegian lexis, is that so much vocabulary is shared by Scottish English and the rest of the United Kingdom. Tulloch (1997:378) adopts the same position as the Scottish National Dictionary, by defining Scots vocabulary as Scots words which are not shared with, or have a different semantic meaning from, 'English English'. Müller explains that Tulloch "also points to the colloquial register as the strongest source of

Scots elements, such as vocabulary items, and points out that this will necessarily ensure its strong connections with colloquial and slang English" (2011:152). Müller argues that:

> Tulloch's identification of the colloquial register as a source of distinctively Scots vocabulary helps to explain the major obstacle in identifying Glasgow lexis. Early Scots researchers did not prioritise the Glaswegian variety as an important subject for study because of its hybrid nature and the inclusion of non-Scots colloquialisms and slang. Due to the influence of large numbers of immigrants, and their languages which do not have roots in historical Scots, Glaswegian was considered a corrupt variety. (2011:152)

Although researchers such as Macafee have researched the Glaswegian dialect, defining a Glaswegian lexicon is problematic. As Macafee argues, defining Scots dialect words based on being used exclusively in Scotland, "reinforces the impression that the shared vocabulary belongs to Standard English whereas Scots consists only of what is uniquely Scots' (1992:51). Faced with an incomplete lexicon of Glaswegian speech, Corbett (cited in Müller) suggests using quasi-academic and literary points of reference such as *The Patter* by Munro (p.56).

This literature review on the use of dialect in fiction demonstrates some of the ways in which representing dialect is not a neutral process, but is ideologically loaded, and may be revealing about the writers' attitude towards an accent, and what they perceive to be 'other'. The literature review on Glaswegian dialect has identified a number of features which can be analysed to compare writers' tendencies, when writing in dialect.

3.5 Character Study 3: Lynda's (im)politeness

The third character study analyses the (im)politeness of Lynda Snell to compare how the six writers portray Lynda's politeness strategies. This is another approach to addressing the second of my sub-questions, exploring the writers' ability to recognise prominent character features – in this case, Lynda's assertive nature – and to use include these linguistic traits in the dialogue they write. 3.5.1 sets out Brown and Levinson's politeness theory, with reference to other key works and methods of analysing (im)politeness in drama, and 3.5.2 discusses selected analyses of (im)politeness in fiction.

3.5.1 Politeness Theory

Brown and Levinson (1987) outline their theory of linguistic politeness in English, which remains the major work on the subject, and McIntyre and Bousfield (2017) argue that a face-based model has generally been the most favoured for analysing fiction. Culpeper's work on (im)politeness (in particular Culpeper, 2011) is also highly influential and is discussed here.

Brown and Levinson (1987) draw on a number of influential theories to outline their politeness theory, including Austin (1955, 1962), Goffman, (1967)) and Grice (1975). At the centre of their theory is the notion of 'face' which they describe as follows:

> Something that is emotionally invested, and that can be lost, maintained, or enhanced, and must be constantly attended to in interaction. In general, people cooperate (and assume each other's cooperation) in maintaining face in interaction, such cooperation being based on the mutual vulnerability of face. (1987:311)

They divide "face" into two related aspects: positive and negative, and argue that negative face is the want of every 'competent adult member' that 'his' actions be unimpeded by others. This includes the "basic claim to territories, personal preserves, rights to non-distraction" (1987:312). It is concerned with the "politeness of non-imposition" and utilises formal politeness. They define positive face as **"**the want of every member that his wants be desirable to at least some others" (1987:312). This includes the idea that others will approve of the speaker's self-image.

Any action which negatively affects a hearer's 'face' is described as a Face-Threatening Act (FTA) and politeness strategies have developed to mitigate against this. Brown and Levinson argue that a speaker (S) will use different approaches, depending on whether the hearer's (H) positive or negative face is threatened. They describe more fully acts which primarily threaten H's negative face-wants (1987:312), but these include: orders and requests, suggestions, advice, remindings, offers, promises, compliments and expressions of strong negative emotions towards H. The acts which threaten H's positive face-want, "by indicating (potentially) that the speaker does not care about the addressee's feelings, wants etc." include: expressions of disapproval, criticism, contempt or complaints; challenges, boasting, raising divisive topics, non-cooperation, and inappropriate use of address terms and other status-marked identifications. As Brown and Levinson note, "There is an overlap in this classification of FTAs, because some FTAs intrinsically threaten both negative and positive face (e.g., complaints interruptions, threats, strong expressions of emotion, requests for personal information)" (1987:314). Further, they distinguish between acts that primarily threaten H's face, and acts that threaten S's face, such as acceptance of H's thanks, responses to H's faux pas, apologies, and self-humiliation.

Brown and Gilman (1989:173) use three variables to measure the weightiness of an FTA: power (P), distance (D) and extremity (R). McIntyre and Bousfield observe that "their assumption is that since politeness theory makes predictions about the level of politeness to be expected relative to the weightiness of each variable, isolating variables makes it possible to determine the effect of individual variables on politeness" (2017). Although a formula is used to assess the weightiness of the FTA, it is obviously impossible to compare the relative impact of distance or power in a quantitative scale: for example, "open the window!" cannot be valued as *x* times less polite than "please open the window". Yet even without a measurable scale, it is useful to analyse interactions using the framework of power, distance and extremity to evaluate the character's politeness strategies. Brown and Levinson argue that, "In the context of the mutual vulnerability of face, any rational agent will seek to avoid these face-threatening acts, or will employ certain strategies to minimize the threat" (1987:315).

Distance between Speaker and Hearer, and weightiness of the request also have a considerable bearing on whether or not a situation requires mitigating facework. For example, in the data, there are scenes where Lynda is recruiting volunteers to perform in the village Christmas pantomime. As the self-appointed director she is often heard enlisting other village residents to join the cast. In some cases, the Hearer is willing and has a track record of performing, so the weightiness of the FTA is relatively small; in other cases, the hearer is very busy or otherwise reluctant and has been coerced into auditioning, so the context increases the weightiness of the FTA. These situations can be further nuanced. For example, asking somebody to the play the lead role, rather than a member of the chorus, is a weightier request, in the sense that Lynda is asking them to accept more work. However, being offered the lead role in a play is inherently flattering, arguably reducing the weightiness of the FTA, whereas someone might justifiably feel put upon if they were asked to commit to rehearsals only to perform a very minor role, or might feel that their acting skills were unappreciated.

In 1996, Culpeper outlined "an impoliteness framework which is parallel but opposite to Brown and Levinson's (1987) theory of politeness" (1996:349), and examined the ways in which speakers create social disruption using strategies which are oriented to attacking the 'face' of their hearers. Culpeper proposes an open-ended list of impoliteness output strategies (1996:357-8), divided into positive and negative strategies. Positive impoliteness output strategies include: ignoring the other person, or failing to acknowledge their presence, excluding them from an activity, being disinterested or unsympathetic, using inappropriate identity markers, and seeking disagreement. Negative impoliteness strategies included condescension, scorn or ridicule, emphasising one's own relative power, and being contemptuous. Culpeper's later revisions of (im)politeness (2005, 2011, 2013) emphasise that strategies and face-threatening acts can be performing more than one strategy at a time, something which he discusses when he explicitly abandons Brown and Levinson's (1987) distinction between positive and negative face. He argues that impoliteness features can be attacking positive and negative face simultaneously: "A particular problem we inherited from Brown and Levinson (1987), and one that is becoming increasingly well-known, is the distinction between positive and negative face" (2003:1576). In conversation with Dynel, Culpeper has advocated the importance of focusing on contexts rather than taxonomies:

> Identifying given strategies in a piece of data doesn't mean that it is, therefore, impolite, because these strategies are always subject to the context they are in. (2013:165)

Since (im)politeness strategies can overlap and are heavily context dependent, it seems appropriate to analyse them qualitatively.

Locher (2005) explores politeness at a discursive level to question whether different writers adopt different strategies of politeness. Moving away from exploring politeness and (im)politeness as binary opposites, Locher instead argues for 'relational work', described as something, which:

> Comprises negatively marked behavior (im)politeness/rudeness), positively marked behavior (politeness), as well as non-marked, politic behavior which is merely appropriate to the interaction in question and not polite as such". (2006:49)

Locher explains her departure from Brown and Levinson's facework by arguing that "Relational work is described as 'the 'work' individuals invest in negotiating relationships with others' (Locher and Watts, 2005:10) and language is seen as one of its crucial means of communication" (2006:250). In Locher's sense, relational work is not so much about facework and its mitigation, but rather "it is understood to cover the entire spectrum of behavior, from rude and impolite, via normal, appropriate and unmarked, to marked and polite" (2005:250). This position echoes Culpeper's argument that features can be performing multiple positive and negative face strategies simultaneously, and is able to consider social cooperation more broadly.

3.5.2 (Im)politeness Studies in Fiction

Characterisation has been analysed using theories of (im)politeness to explore fiction, including Simpson (1989), who uses Brown and Levinson's politeness strategies as a framework to analyse characterisation in Ionesco's *The Lesson*. Further studies include Culpeper (1998), who analyses the Colonel's dialogue in the film *Scent of a Woman*; and McIntyre and Bousfield (2018) who analyse the language of US Marine Corps recruit-training in the 1987 film *Full Metal Jacket*. These studies show ways in which the analysis of (im)politeness in dramatic dialogue can reveal information about characterisation, and can be used as a replicable methodology to track changes in characterisation.

McIntyre and Bousfield (2017) argue that drama provides a rich resource for studying (im)politeness:

> The stylistic effects that we encounter in fiction (e.g. conflict, dramatic tension, plot development, humour, etc.) are often created by narrators and characters violating aspects of interaction. Such violations can be revealing of how processes of interaction work and these insights can be useful to pragmaticians in reassessing and revising pragmatic concepts and frameworks for analysis. (2017:759)

One such example of violated norms is found in Ionesco's *The Lesson*. Simpson (1989) analyses the politeness strategies, using Brown and Levinson's framework. He demonstrates how the play's two characters, The Professor and The Pupil swap roles as the play progresses, and the Professor's numerous initial hedging strategies are replaced by a series of non-redressive FTAs as the play builds to its murderous climax.

Simpson, like Culpeper, advocates the importance of context in analysing politeness. He argues that strategies of politeness, "are not arbitrarily chosen by speakers in interaction. On the contrary, their choice is constrained by important contextual features, such as the relative power of the speakers, the social distance of the speakers, and what the speakers happen to be negotiating at the time" (1989:171). For these reasons, Lynda Snell's dialogue is analysed by exploring the full scenes, including stage directions, so that the relational work (or lack of) can be explored in context, taking into account the power dynamics between Lynda and the other speakers. I will also consider the constraints which Simpson refers to as "the higher level of literary organization" of the interaction between the writer and audience, by considering ways in which politeness strategies operate not just at the level of character to character, but also as a message from playwright to audience.

3.6 Analytical Methods for Chapter 7: The Functions of *Oh*

This section on analytical methods used in the three main studies relates to the final study (Chapter 7): a pragmatic analysis of the token *oh.* This chapter explores the third of my sub-questions, which asks whether higher-level pragmatic features provide a base for authorship analysis in cases of linguistic identity disguise (see 2.5.1 for a summary of Grant and MacLeod's position on levels of language analysis). This next section relates this idea of 'higher-level' analysis to the use of pragmatic markers, in particular, pragmatic noise.

3.6.1 Introduction

Understanding meaning is a central concern of pragmatics (Culpeper and Haugh, 2014:1). Pragmatics covers a wide range of interests, including a focus on the intentions and interpretations of utterances. Culpeper and Kytö write:

> Like sociolinguistics, the field of pragmatics generally takes the view that language is a societal phenomenon, and also emphasises use, uses and contexts. The focus is typically on:
>
> - The utterance
> - The speaker's intentions
> - The hearer's interpretation of the speaker's utterance and the intentions behind it
> - The social interaction between the speaker and the hearer.
>
> (2010:8)

My study focuses on pragmatic noise, which is particularly relevant to spoken data, and, increasingly to written data that incorporates speech-like elements, such as social media communication. First I discuss selected literature on pragmatics in the context of forensic linguistic analysis. Then I review literature on *pragmatic markers*, before focusing on a subset of this category, *pragmatic noise*, which is the feature analysed in Chapter 7. My final study is an analysis of the meaning of *oh* in context, and therefore the final part of this literature review discusses research into the various functions of *oh*. In Chapter 4, I describe how the sub-corpora were compiled for the analysis and I set out how the token *oh* was codified into separate categories.

3.6.2 Interjections, Inserts, Discourse Markers and Pragmatic Markers

The following section provides a very brief review of some of the main literature on pragmatic markers and their counterparts. Pragmatic markers occur frequently in spoken language. Archer et al. describe these as a notable feature of dialogue:

> We pepper our conversation with 'smallwords' (or pragmatic markers) such as *well, you know, I mean*. At first sight they seem to mean very little. However, they play an important role in making our speech coherent and in establishing or maintaining our relations with interlocutors in conversation. (2012:74)

Similarly, Ameka (1992) discusses interjections, calling them: "those little words, or 'non-words', which can constitute utterances by themselves" (1992:101). He describes them as conventionalised vocal gestures which express a speaker's mental state, and likens them to Goffman's (1981) response cries. There are a multitude of ways that these 'smallwords' or 'nonwords' are classified. Brinton (1996) lists over twenty alternative terms, including connective, continuer and discourse particle (1996:29). There have been detailed studies on these 'smallwords' or 'nonwords', such as Schiffrin's influential work on Discourse Markers (1987), in which she uses the term to describe words such as "oh", "well", and, "but", or, "so", "because", and "y'know". Aijmer (2002) used the term Discourse Particles, Ameka (1992a, 1992b and 2006) used Interjections, and Biber et al. (1999) described them as Inserts. Archer et al. discuss these overlaps in terminology:

> Some use the term pragmatic markers, as we have done, some, the terms discourse markers, discourse particle, connective, filler, etc.. However, following Fraser (1996), we take 'pragmatic marker' to be an umbrella term encompassing a large number of related pragmatic phenomena with an 'insert' function, and 'discourse marker' to be a term which expresses a relation between utterances such as elaboration, contrast, or inference. (2012:76)

Although there is a multitude of labels for these 'smallwords', the focus here is on the types of meaning they convey, rather than their categorisation.

Brinton notes the lack of consensus on the definition of a pragmatic marker, but explains why it is her preferred term: "The term marker is preferable to either word or particle since it can encompass single-word items such as "so", as well as phrases such as "you see" (1996:29). Brinton points out that different definitions of pragmatic markers in the literature seem to bear very little resemblance to each other, noting that different researchers place differing emphases on the marker, such as Schourup's (1985:3) exploration of pragmatic markers as a response signal, or Fraser's focus (1998, 1990) on the sequential discourse relationship. What is relevant is the communicative purpose, and whether or not there are observable differences between different writers in their use. Because I carry out a pragmatic analysis of these words, and am concerned with the meaning and functions of *oh*, I follow Archer et al. in using the term *pragmatic markers* as an umbrella term.

3.6.3 Conveying Meaning in Pragmatic Markers

Scholars have debated whether or not pragmatic markers convey meaning. Brinton describes them as 'grammatically optional and semantically empty" (1996:35), and Schiffrin argues that they operate at a discourse level, and are not tied to any particular sentence structure (1987:37). Fischer critiques Schiffrin's vagueness:

> An important problem is presented by Schiffrin's insistence that discourse particles do not have meaning. Obviously, *meaning*, is reserved, for her, to 'ideational meaning' (but see also Redeker 1991:1162). (2000:279)

Fischer discusses at length whether pragmatic markers are polysemous, or should be treated as homonyms (2003). The multiple possible meanings of pragmatic markers mean speakers have many more options about *how* they will use a word such as "ah" or "well", compared to other word classes, for example nouns or verbs. Its meaning is interpersonal or textual rather than contributing to the "content" of the sentence or turn (Archer *et al.* 2012:74). Fischer's argument that meaning in pragmatic markers has a meaning beyond ideational meaning is echoed in Culpeper and Kytö (2010) who discuss the importance of pragmatic markers in conveying interpersonal meaning:

> In semantic-pragmatic terms, pragmatic markers have in common the fact that they have little or no propositional meaning but tell us about the pragmatic relationships between a speaker, their message(s) and its context. (2010:361)

Likewise Ameka argues that pragmatic markers, "encode speaker attitudes and communicative intentions and are context-bound" (1992:107). Archer also argues that pragmatic markers are important for conveying meaning in interactions. She writes:

> *Well, I mean* and *you know* have pragmatic meaning (e.g. interpersonal or textual meaning) and do not contribute to the content. Although it is difficult to say what they mean, they obviously play an important part in making the conversation coherent. If we 'dismantle' the utterance from any of these extras, the message gives an abrupt or brusque impression. (2012:74)

Archer's position that discourse and pragmatic markers play an important part in conveying meaning is an important theoretical position for my analysis: these interpersonal and contextual meanings can be used to investigate variation in meaning of the same token in different contexts. Aijmer discusses the flexibility of pragmatic meanings for words such as *oh*. She writes:

> Discourse particles are different from ordinary words in the language because of the large number of pragmatic values that they can be associated with. Nevertheless speakers are not troubled by this multifunctionality but they seem to know what a particle means and be able to use it appropriately in different contexts. The problem is how we can account for their multifunctionality without multiplying the meanings of the particle. (2002:3)

This "large number of pragmatic values" is interesting for authorship attribution, and is discussed more fully, below, because it allows a comparison between the different meanings that each writer uses for a single lexical token. My research adopts the position that pragmatic markers carry communicative meaning, and, like Ameka, I take the position that the meaning conveyed in pragmatic markers is "context-bound". Accordingly, the use of *oh* will be studied in context, to determine meaning, and to compare variation in communicative purpose.

3.6.4 Pragmatic Noise

This next section describes and discusses *pragmatic noise*, a subcategory of pragmatic markers. As defined by Culpeper and Kytö (2010), pragmatic noise refers to tokens such as *Ah*, *Oh*, *Hum* in their single and / or reduplicated forms (e.g. *Ha Ha Ha*). Culpeper and Kytö outline six key criteria for inclusion in the category of pragmatic noise items. In their definition, pragmatic noise items:

> (a) do not have related words which are homonyms in other word classes (thus ruling out many interjections, such as, for example, HEAVENS, which may be an interjection or a noun.
>
> (b) do not participate in traditional sentence constructions (the majority appear parenthetically in initial position)
>
> (c) are morphologically simple (always uninflected, very rarely compounded)
>
> (d) do not have propositional or referential meanings but pragmatic or discoursal meanings (e.g. expressing the speaker's surprise, disgust, rejection of what the previous speaker said, and so on)
>
> (e) have less arbitrary meanings compared with most words (they are sound symbolic to a degree)
>
> (f) may not have the typical phonological structure expected of a word form (e.g. PSHAW, assuming the initial <p> was pronounced, or Present-day MHM).
>
> (2010:199-200)

The first requirement (a) is close to Ameka's category of primary interjections. Ameka distinguishes between "*primary* interjections, that is they are not used otherwise; and other words which come to be used as interjections by virtue of their notional semantics. These may be considered *secondary* interjections" (1992:105); for example, the difference between, "Ouch!" and "Heavens" respectively.

Culpeper and Kytö describe the lexical items included in this group of pragmatic noise as "central members" of Biber's group, "Inserts". In my thesis, I follow Culpeper and Kytö's position of avoiding the term interjection, "because of its narrower focus and bias towards standard written grammar" (2010:203-204). Situating pragmatic noise within the broader theoretical field, they write:

> Pragmatic noise items have much in common with pragmatic / discourse markers. It is no surprise, then, that interjections have been discussed within pragmatics, most notably in a special issue of the *Journal of Pragmatics* (1992). … It is in this general sense that we use the term 'pragmatic' in the label pragmatic noise. Pragmatic noise can be seen as a subgroup of either pragmatic (or discourse) markers. (2010:204)

As Culpeper and Kytö (2010:200), and Archer et al. (2012:120) point out, pragmatic noise features in naturally occurring dialogue, but also features regularly in play-texts, making it an apt pragmatic feature through which to analyse my data.

3.6.5 Meaning in Pragmatic Noise: *Oh*

This next section reviews some key studies into *oh* and its range of pragmatic uses, including Heritage (1984), Schiffrin (1987), Aijmer (1987), and Macaulay (2005). Aijmer observes that *oh* is a frequently used marker, stating, "in almost any conversation between two or more participants speakers use a lot of *oh* and to a lesser extent *ah*" (1987:61). Different researchers, from a number of decades, have, as expected, focused on different dimensions of *oh*. Heritage's 1984 paper looks mostly at sequencing of *oh*, and how it fits a three-part structure of question, answer, and *oh* response. Aijmer's 1987 paper is a very detailed taxonomy on the different pragmatic functions of *oh* in the London-Lund corpus. Macaulay (2005) approaches the question from a sociolinguistic perspective investigating its varied use. Section 4.6. gives a detailed account of how *oh* was codified in the analysis, and 7.2 reviews different categories of the functions of *oh* as found in my data. This next section discusses some of the functions of *oh* which have been identified in research.

Heritage (1984) provides perhaps the first in-depth investigation into the uses of *oh* (Fox Tree and Schrock, 1999:281). He emphasises the function of *oh* as a change-of-state token, writing that, "Evidence from the placement of the particle in a range of conversational sequences shows that the particle is used to propose that its producer has undergone some kind of change in his or her locally current state of knowledge, information, orientation or awareness" (1984:299). He breaks down the functions of *oh* into a number of categories. The first of these is "informings", where he contrasts *oh* as information receipt to other options such as "yes" and "mm":

> It is proposed that *oh* specifically functions as an information receipt that is regularly used as a means of proposing that the talk to which it responds is, or has been, informative to the recipient. Such a proposal is not accomplished by objects such as "yes" or "mm hm," which avoid treating prior talk as informative." (1984:307)

Another type of informing, is question-elicited informing. Heritage states that: "Just as "oh" receipts regularly occur in the environment of informings that are, in various ways, initiated by the informant, so they also regularly occur in response to informings that are elicited by questions" (194:307). This links to another of Heritage's categories, Understanding Checks, where *oh* can be used as a display of understanding where "recipients may wish to show that prior talk has been adequately descriptive and / or that they have competently understood its import" (1984:321).

Heritage addresses the subject of *oh* as a complete turn, or *oh* as a turn-initial particle, writing that, ""Oh" is systematically weaker than an "oh" plus inquiry or "oh" plus newsmark receipt in that (1) it fails to invite the informant / news announcer to tell more and (2) in projecting additional turn components, it may invite the announcer to await them by withholding from further talk." (1984:329). In fiction, conversational norms are often violated for dramatic effect (McIntyre and Bousfield, 2017), so this use of *oh* as a means of halting the conversation or failing to invite the speaker to continue, could be revealing about a character's way of speaking, but could also be revealing of authorial style, if it is a frequently used dramatic device to increase friction within a scene.

Aijmer (1987) investigates different functions of the word *oh* in the London-Lund corpus, using a 'bottom up' approach of extracting every instance of the word *oh* and categorising thirteen different ways that *oh* and *ah* could be used, including: surprise (which can be an interruption from the hearer to speaker, or can be the speaker interrupting themselves (1987:63); in answer to a question; conventionalized phrases: *oh thank you, oh I beg your pardon* (1987:80); and as a qualification, giving the example of "*Oh well*" being used to concede something reluctantly (1987:78). There are overlaps between Aijmer's categories, which creates issues with coding occurrences of *oh* into discrete categories. For example, Aijmer discusses the use of *oh* as a sign of frustration and annoyance, but also lists a separate category, surprise. Attempting to code occurrences of *oh* as either frustrated and annoyed or as surprised would be a subjective exercise, which could lead to a method which is not replicable in the way that is required for forensic authorship analysis. These overlaps are not problematic for Aijmer's purposes, but in quantitative approaches it would be entirely possible for an instance of *oh* to fall in both categories, and for an analyst to decide whether an *oh* was more annoyed than frustrated, or vice versa, is too subjective to be replicable.

Similarly, emotionally defined occurrences of *oh* such as disappointment, annoyance and frustration (1987:66) are listed separately, but would be difficult to code as separate emotions in a consistent way. For Aijmer's purposes this overlap is not an issue, but for a coding exercise these overlaps become problematic. For example, evaluation and endorsement could have numerous occurrences which fit both categories.

Aijmer also discusses the effect of medium on the frequency of *oh,* arguing that, "The obligation to show the cognitive effects of an utterance may be stronger in some genres of spoken English than others. A comparison with other genres of spoken English in LLC showed for instance that *oh* and *ah* were more frequent in telephone conversation (density: 0,68%) than in face-to-face conversation (0,42%)." (1987:80-81). These findings could suggest that the frequency of *oh* will be higher in a radio drama because it is a non-visual medium, just like a telephone call is (or at least, was, at Aijmer's time of writing). This suggests that verbal acknowledgements and continuers, as realised by *oh* and *ah,* are used more frequently because non-verbal communication, such as nodding and smiling, is not available to the participants.

Another influential study on the uses of *oh* is Schiffrin (1987). Schiffrin provides a detailed analysis of the functions of *oh*. She divides *oh* into groups, including: *oh* in question/answer/acknowledgement sequence, *oh* and shifts in subjective orientation, and also the use of *oh* as a backchannel (i.e. a single word response of *oh*, before the speaker continues with their turn, so *oh* does not cause a change in speaker.

Macaulay, in his sociolinguistic study of variation in discourse along age, gender and social class divisions, explores the variation in the uses of *oh*. His division of *oh* into its various functions, as found in his data, is one of the simplest breakdowns, using just five main categories. They are:

- Acknowledgement
- Agreement (often enthusiastic agreement, e.g. "Oh yes")
- Emotion (as part of a phrase, "oh dear", or before a strong statement of opinion, e.g. "Oh he's a wee arsehole")
- Quoted dialogue ("And I went, 'Oh I need to bring my pyjamas then?')
- Questions: to introduce questions, usually asking for confirmation or elaboration.

(Macaulay 2005:58-59)

Macaulay's fourth point is echoed in Furkó and Abuczki's analysis of the functional spectrum of pragmatic markers in news and celebrity interviews. They describe how *oh* is used as a 'ventriloquism' technique, where a speaker uses it to indicate that they are about to voice the words of another person" (2014:56). They illustrate their point with this example: "And then they say, ***oh*** it doesn't need to come in to effect for eighteen months or two years" (2014:56, emphasis added). Again, it is easy to see how these categories could, in practice, overlap. For example, "Oh dear" expresses emotion, but is also an acknowledgement of new information.

Culpeper and Kytö's (2010) analysis of speech-like language, including comedy playscripts and courtroom trials in Early Modern English outlines different functions of *oh*. The authors carry out a corpus study of features of pragmatic noise. *Oh* is divided into a number of different functions: emotive expressive, cognitive and conative. A number of functions of *oh* are discussed, including: to convey distress (the emotive expressive function); to express moments of surprise; sudden realisation; and frustration (2010:239). Discussing the differences in functions of *oh* and *ah* they found that *oh* was often used as a preface to an answer, and as a politeness strategy" (2010:241). In their historical corpus, it was found that *oh* often collocated negatively, such as "O, Fie!", whereas "Ah" was more likely to have a positive collocation (2010:241). This is not a distinction that necessarily holds today, and may have been affected by diachronic change. They also noted that *oh* mostly reinforced an affirmative answer to a yes/no question, and was far less frequently used to reinforce a negative answer (2010:42). Observations such as this could be useful for noting a marked use of *oh* by an author. The authors found that "in some cases the answer-preface signals surprise that the question was asked." It could be that this surprise is genuine, or could be an affectation to make a point, for example, a teacher's mock surprise response to some missing homework.

Culpeper and Kytö also note that *oh* can be used as a politeness strategy, as in the example below, where the speaker uses "O" as a hedge, before refusing to answer the question:

Aim.	And pray, Sir, what is your true Profession?
Gib.	O, Sir, you must excuse me – upon my Word, Sir, I don't think it safe to tell you.
	(Drama/Farquhar, The Beaux Strategem, 1707:27-8 (2010:243)

One observation arising from this historical usage of *oh* as a preface is its frequency as part of a conventionalised phrase, for example, "O, Sir". Culpeper and Kytö's example is drawn from Early Modern English, but they compare it to the contemporary usage of *oh* in conventionalised phrases, such as "oh thank you" and "oh sorry", noting the similarity in pragmatic function in Present Day English. Scholarship on the functions of *oh*, reviewed here, has informed the way in which *oh* was coded in Chapter 7. The detailed coding for *oh* and an explanation of how the categories were reached is set out in Chapter 4.

3.6.6 Pragmatics Markers in Drama Scripts

Before analysing the use of pragmatic markers such as *oh* in drama scripts, it is important to consider how this feature might differ in scripted data, compared to the way it appears in naturally-occurring conversation. Brinton observes that, "Pragmatic markers are predominantly a feature of oral rather than written discourse. The appearance of pragmatic markers is a result of the informality of oral discourse and the grammatical "fragmentation" caused by the lack of planning time, which makes the use of pragmatic markers expedient (1996:33). She adds that the feature is not restricted to oral texts though. Using data where the illusion of spokenness is common might allow useful links to be drawn with computer mediated discourse: much authorship attribution work uses computer mediated discourse as a medium, and in this medium, deliberately using features associated with spokenness is commonplace (Crystal, 2006).

Discussing the types of texts which contain features of oral language, Culpeper and Kytö propose three categories of "speech related" texts:

- *speech-like*, e.g. Personal correspondence
- *speech-based*, e.g. Trial proceedings
- *speech-purposed*, e.g. Plays."

(2010:17)

The data used here are clearly from the third of these categories: the scripts are written first, with the purpose of being spoken aloud. Furthermore, the sub-genre of the play script as a soap opera dictates that "everyday" language is used, in contrast to many of the rhetorically florid sixteenth century texts analysed by Culpeper and Kytö.

This categorisation provides a helpful distinction between the ways we may encounter 'spoken' language in written form. In the first category, informal correspondence may adopt a number of linguistic features traditionally associated with oral language. In the second, spoken language is transcribed, and in their third category, a text is written with the express purpose of being spoken aloud. Discussing the use of pragmatic noise in "speech-purposed" drama scripts, Culpeper and Kytö write:

> Part of the reason why they appear at all may be the wish to (re)create an illusion of spokenness, but an equally important reason relates to their role in signalling meanings in interaction. As we shall demonstrate, writers use them to signal a participant's thoughts and feelings about something – most often the previous participant's thoughts and feelings about something – most often the previous participant's discourse. Clearly, this is an important means of displaying character relations and characterisation generally in Play-texts. (2010:200)

Arguably there will be differences in the uses of pragmatic noise between these three categories. Unless aiming for a certain style of verbatim drama, scripts will almost certainly contain fewer instances of self-repair than spoken, unscripted language, as Quaglio (2009) observed, in his analysis of dialogue in the TV sitcom *Friends.* Topic management will also operate differently, since one authorial voice is controlling all the voices involved in the conversation. To use drama scripts as an example of how pragmatic noise is used in naturally-occurring conversation would be problematic. Instead, the aim is to analyse the variation in use of pragmatic markers in one text type, to explore whether different authors use pragmatic markers to convey different pragmatic meanings in drama scripts, and whether this varies between characters or remains constant regardless of who is speaking.

If successful at discriminating between authors, the same methodology could be used to investigate authorship in other data types, such as chatroom transcripts.

3.6.7 Audience Design

The influence of audience is, of course, significant in play-texts, because the drama exists for the benefit of the watching and/or listening audience (See also 2.4.6 and 2.7.3). This issue has been raised within the field of Pragmatics as well as in theatrical and literary studies. Culpeper and Haugh discuss the difference between first-order and second-order perspectives:

> A first-order perspective is that of the participants themselves, the ones who are using language to mean and do things. A second-order perspective is that of the analysis, including ourselves, the writers of this book, and you the readers. (2014:11)

This analysis includes second-order perspective because the audience is not simply an overhearer, but is the intended recipient of the information. In line with the work of drama theorists such as Pfister (1991) and Wallis (1998), Culpeper and Kytö explain how this triangular relationship works with regards to pragmatic noise:

> At the author-audience discourse level, *all* pragmatic noise items are pragmatic markers. They are all *authorial* pragmatic markers, signalling to the audience the attitudes and intentions of characters and how character-talk should be taken. Items such as OUCH or laughter, produced as relatively spontaneous reactions, cannot be dismissed as unconscious non-strategic items (i.e. not pragmatic markers), since they have been put there on purpose by the author. Similarly, hesitators cannot be dismissed as normal non-fluency. It is only at the character-character discourse level that *some* pragmatic noise items will also – to varying degrees – be *speaker* pragmatic markers. (2010:221)

This perspective, also discussed in Richardson (2010) and McIntyre (2015b) suggests caution should be exercised before taking any findings on the use of pragmatic markers and applying them directly to other speech-related or conversational genres.

Culpeper and Kytö discuss the importance of pragmatic noise in conveying spokenness. Citing Koskenniemi (1962:73) they write:

> Firstly, pragmatic noise has an intimate association with both spokenness and interaction. It has been suggested that interjections are associated with realistic dialogue … The argument must be even stronger for pragmatic noise, since 'noises' have a fundamental connection with spontaneous vocal reactions. (2010:200)

Following this observation, it is possible – or even likely – that the writers use *oh* with a secondary simultaneous function – to convey spokenness – along with whichever 'primary' function is being used, for example, surprise or confirmation. Any findings in these data have a parallel in many other data forms, such as social media discourse, but should be applied with caution to other modes of communication, such as actual spoken conversation, where creating an illusion of spokenness is, of course, not an aim; although creating a sense of informality, may be, and may use similar linguistic strategies. However, in all modes of communication there can still be a performative element to *oh*, for instance to convey a feeling of surprise or disappointment about information which is actually not news to the participant, or to communicate that a reply has received only minimal consideration.

3.6.8 The Prosody of *Oh*

Drama scripts are, in a sense, incomplete documents. Unlike a novel, which is the finished form received by its audience, a play-text is intended to be passed through the necessary intermediary process of production and performance to reach its audience. In this way, there are further layers of interpretation by actors, director and designers. Discussing the impact of actors on meaning, Culpeper and Kytö write, "Pragmatic noise items are particularly sensitive to variation in pitch, vowel length, loudness, and voice quality – all of which convey nuances of meaning" (2010:206). Discussing Interjections, Norrick notes that, "intonation can play an important role in the interpretation of interjections. For instance, *oh* with a rising contour certainly fulfils different discourse functions from *oh* with a falling contour" (2009). Sometimes stage directions will specify a character's emotion, but often the actor will interpret the emotion and will use their own performance to convey an emotional or pragmatic interpretation of a lexical item.

Fox Tree states that "discourse markers are solutions to problems of spontaneous talk" (2015:64), for example when used to convey hesitation or in repair. Exploring any differences between spontaneous speech and spontaneous writing, Discourse Markers were divided depending on whether they were (1) attitudinal, (2) tailored, (3) temporally sensitive, or (4) cohesive. Results showed that “although they vary in frequency in spoken versus written domains, discourse markers are used similarly across domains, but with particular communicative functions that make them non-interchangeable” (2015:64). This supports the idea that any findings from a play-text, which purports to present spontaneous speech, could, cautiously, be applicable to other domains, such as online communications.

3.6.9 Pragmatic Markers as Style Markers

Brinton’s description of pragmatic markers as “grammatically optional” and “semantically empty” (1996:35) suggests two features which are of interest for authorship attribution: as a “grammatically optional” feature, an author has more freedom of choice about whether or not to use pragmatic markers, and some syntactic freedom about where to place them if used. This logically increases the chances of individual variation. Secondly, the “semantically empty” nature of the marker allows for the possibility that different authors will apply different meanings to the same token. This gives each writer greater scope to choose the pragmatic function of the marker, which gives a higher chance of variation. Two important questions are: does an individual writer alter their use of pragmatic markers for different characters? And secondly, do the different scriptwriters (either instinctively or deliberately) use pragmatic markers in different ways for different characters? For authorship analysts, this second point is interesting, because this semantic and structural freedom allows authors to make individual decisions about where, how and whether to use pragmatic noise; increasing the possibilities of inter-author variation.

Two further benefits of analysing *oh* are: firstly, being unmarked (Larner, 2014), and being frequent (McMenamin, 2020). In my data, pragmatic markers meet both of these conditions. Culpeper and Kytö discuss the fact that spoken language is non-clausal and does not follow the rules of written grammar, pointing out that, “Pragmatic markers dominate this non-clausal material, and thus deserve our attention on frequency grounds alone” (2010:362). Aijmer, too, comments on the frequency of discourse particles in spoken interaction:

> The frequency of discourse particles sets them apart from other words in the language. Altenberg (1990:185) found, for instance, on the basis of a 50,000 word sample from the London-Lund Corpus, that 'discourse items' (also including greetings, thanks, apologies) accounted for 9.4% of all word-class tokens, and in fact, constituted the fourth largest word-class only outranked by verbs, pronouns and nouns but outranking the basic grammatical categories preposition, adverb, determiner, conjunction and adjective. (2002:2-3)

These figures show that pragmatic markers meet McMenamin's desired attribute of frequency. These combined qualities of pragmatic markers as frequently occurring, and able to convey a variety of communicative purposes, makes them highly promising features for authorship analysis.

3.7 Conclusion

This chapter has drawn together some of the selected academic literature which informs my analytical methods. Since each of my analytical chapters deliberately explore linguistic identity disguise using a range of methods, the topics that have been reviewed here are necessarily wide-ranging. The topics discussed provide essential background information to the following chapter, in which I set out my methodology.

4. Methodology and Data

4.1 Introduction

This is the second chapter concerned with Methodology. Chapter 3 reviewed the relevant literature for each analytical method, and in the present chapter I describe the data used, and set out my methodology for each analysis. First, I describe my overall dataset. Then, for each of the three main analyses, I describe the sub-corpora and methodology used, explaining the software used and the rationales behind decisions. The second of these analyses ("Three Character Studies") is divided into three sub-sections, one each for each of the characters, and the first of these ("The Lexis of Jim Lloyd") further sub-divides into three separate analyses of Jim's language. The chapter is structured so that the data and methodology for each analysis are discussed together to avoid separating the descriptions of corpora from the analyses for which they were prepared. At the end of this chapter, Table 9 provides a summary of all the corpora used in my thesis.

4.2 Data

I was provided with the digital copies of the "Script As Broadcast" for 1440 episodes *The Archers* by the programme's production office at BBC Birmingham. Each 15-minute episode script is a separate Word document with its own title page (see Appendix 1 for an example). The "studio script" is the version of the script which has been edited by the script editors and formatted ready for the studio recording, and includes technical and production information for the cast and crew. The "Script As Broadcast" is a digitally annotated copy of this studio script, documenting the final version of the programme as broadcast. It marks any cuts or alterations to the dialogue made in the studio recording or during post-production.

The scripts were broadcast between 2010 and 2017, and were written by six scriptwriters who gave written permission for their scripts to be analysed for my thesis. All six writers were established members of the scriptwriting team by the time the first scripts in the data were written, so I was comparing six experienced writers, rather than a mix of established and new. All six wrote regularly for the show between 2010 and 2015, with five of the six writers continuing to write regularly for the show until 2017 and beyond. Table 3 shows the number of scripts per writer. This corpus of 1440 studio scripts is referred to as the *Archers*-Complete Corpus. The writers have been anonymised, but are referred to by the same number throughout the thesis: for example, Writer 1 always refers to the same person. The scriptwriting team is made up of around 12 freelance writers, who are not necessarily commissioned for equal numbers of episodes per year. For example, in one year, one writer wrote two weeks' of episodes, while another writer was commissioned to write seven weeks' of episodes.

Table 3: Number of scripts in the *Archers* Complete Corpus

Writer 1	Writer 2	Writer 3	Writer 4	Writer 5	Writer 6
288	228	216	258	228	222

4.2.1 Preparing the Data: Removing "Script As Broadcast" Annotations

In studio, it is standard practice for directors to omit to record lines if the estimated programme length will be longer than the programme's scheduled transmission slot, or to cut lines during the editing process if the recorded programme is too long. If a line is not recorded, or is cut during post-production, it is marked with strikethroughs. All "Script As Broadcast" annotations were manually removed to return the scripts more closely to the writers' versions. For example:

JOE Much the same as last year, then.

~~DAVID One or two new things - Josh is going to film it, and we're still hoping Pip might make a contribution when she's finished her revision.~~

JOE Eddie and me would be more than happy to offer you a helping hand.

(Writer 5, 2010, Studio Script with "Script As Broadcast" annotations)

These deletions were reinstated to keep the data as close as possible to the writers' final scripts.

It is also common for actors to make minor alterations to the lines, as in the italicised “Mm, yeah” in this extract:

LEWIS   That’s the third, isn’t it?

NIGEL   *Mm, yeah.* Can you get in touch with your people and tell them they’ll all be needed on that day.

(Writer 4, 2010, Studio Script with “Script As Broadcast” annotations)

In this example, the actor has added the pragmatic noise item, “Mm” followed by an affirmation, “yeah”, perhaps to sound more conversational, or if a script is ‘running short’, to add some fillers, to increase the recording’s length. These changes are marked in blue in the studio scripts, but are represented here in italics.

In a third example, this change (italicised) was made in studio.

RUTH   Um... it’s a bit early for ~~the twins~~ *Lily and Freddie*, isn’t it?

ELIZ   Sorry?

(Writer 4, 2010, Studio Script with “Script As Broadcast” annotations)

In this instance, “the twins” was reinstated and “Lily and Freddie” deleted. It has been established in the fictional world of Ambridge that Elizabeth prefers her children to be referred to by their individual names, rather than as “the twins”, so presumably this change was made in studio for continuity reasons (even if the scriptwriter’s logic was that other characters may not necessarily adhere to Elizabeth’s preference.) All such additions and alterations were removed.

Scripts in the *Archers*-Complete Corpus were then saved as plain text files, so that they could be opened in Notepad++. Each script has a front cover which gives production details including the writers’ and directors’ names, studio recording dates and times, and the programme number. I then created the *Archers*-Writer corpora. To do this, I wrote regular expressions (Regex) and used the *Find and Replace* function in Notepad++ to remove all the front cover text, so the remaining text consisted only of the scenes within each episode. (All Regex coding can be found on the Quantitative Results spreadsheet in Appendix 2a). Once this was completed, I created six folders: one for each writer. In each folder, I compiled the text from the studio scripts to create a document of all the dialogue written by one author during one year (for example, Writer 1, 2010). This process was repeated for all the years of data, and then for the remaining five writers.

From these *Archers*-Writer corpora, a number of sub-corpora were created. These corpora were compiled and prepared in different ways, depending on the method of analysis for each study. These sub-corpora are now described, followed by the related methodology for each study.

4.3 Data for Chapter 5: the 20-Character Corpus

The first of the three studies is a quantitative analysis of structural level features, including a word-n-gram-based test, and three basic stylometric tests. To prepare the data for this study, I used scripts by all six writers from 2010-2012, the first three years of each *Archers*-Writer Corpus.

The decision to select three years' worth of data was an iterative process: using longer periods of data meant that characters fluctuated in frequency over the period; some characters were written out of the show, and new characters were introduced, and some had major storylines producing large amounts of data, whilst others appeared less frequently. To find the twenty most-frequently speaking characters, I removed all of the stage directions from each *Archers*-writer corpus, because my focus in the quantitative study was to analyse variation in dialogue between writers and characters, so retained only the dialogue. The stage directions were removed by writing Regex code to remove all text within parentheses. (To avoid confusion in studio, parentheses in *The Archers* scripts are only used for stage directions, and not for dialogue). After cleaning the data, only the speaker name (showing which character says a line) and the dialogue itself remained. Then I combined the six Writer corpora for the years 2010-2012 into one single file.

I then imported the data into RStudio, and adapted the code in Jockers and Thalken (2020:82-84) to return a frequency list of the number of lines spoken by each character. The same twenty characters were studied for all six writers. It should be noted that these are not necessarily the twenty-most frequently speaking characters in the programme as a whole, because approximately only half of the programme's writers were included in my study. These characters are listed in Table 4, with brief information about each character, taken from www.bbc.co.uk/archers (where more details can be found for each of these characters, as well as for characters not in this list of twenty.) The direct quotes in this table are taken directly from the relevant character profiles on *The Archers* website. This table provides some contextual information about the characters and their relationships within the fictional universe which may be useful background information for when I discuss my results in Chapters 5-7.

Table 4: Character Summaries

Character (in order of descending frequency)	**"Biographical" details: occupation, close family and year of birth.**
David Archer	Farmer, married to Ruth Archer. Born 1958
Pat Archer	Farmer, married to Tony Archer; cousin of David Archer. Born 1952
Ruth Archer	Farmer, married to David Archer. Born 1968
Jennifer Aldridge	Housewife. Wife of Brian Aldridge. Born 1945
Lilian Bellamy	Property developer, "bonne vivante". Sister of Jennifer Aldridge and Tony Archer. Born 1947
Tom Archer	Manager at Bridge Farm. Son of Pat and Tony Archer. Brother of Helen Archer. Born 1981
Brian Aldridge	Farmer and Landowner. Husband of Jennifer Aldridge. Born 1943
Kenton Archer	Brother of David Archer. Pub co-manager. Born 1958
Elizabeth Pargetter	Elizabeth Pargetter (née Archer). Stately home owner / manager. Born 1967
Helen Archer	Shop manager / cheesemaker. Daughter of Pat and Tony Archer; sister of Tom Archer. Born 1979

Lynda Snell	B&B owner, and Director of the village pantomime. Born 1947
Pip Archer	Farmer. Daughter of Ruth and David Archer. Born 1993
Tony Archer	Farmer. Husband of Pat Archer; sister of Jennifer Aldridge and Lilian Bellamy. Born 1951
Jill Archer	Retired Farmer's wife. Mother of David, Kenton, Elizabeth and Shula. Born 1930
Susan Carter	Manager of village shop; dairy worker. Born 1963
Fallon Rogers	Daughter of the village pub's landlady. Born 1985
Jim Lloyd	Retired Professor. Date of birth not specified, based on dialogue, around 1940
Eddie Grundy	Various "money-making wheezes". Born 1951
Brenda Tucker	Works in property. Girlfriend / ex-girlfriend of Tom Archer. Born 1981
Jazzer McCreary	Pigman and Milkman. D.o.b. not specified, but based on contextual information early 1980s

For each of the six scriptwriters, twenty corpora were produced, one for each of the twenty most frequently speaking characters. Again, this was done using Regex with the *Find and Replace* function in Noteapd++ to find all lines of the script beginning with a particular character name. Once these had been separated into 120 separate corpora, the scripts were cleaned again, to remove the name of the speaker which begins each line of dialogue. This was also done writing a Regex code for the *Find and Replace* function. What remained in each sup-corpus was a plain-text file with a single character's dialogue, written by a single author. These corpora were then checked manually, and any remaining character idents or stage directions were individually removed. Finally, all three years for each sub-corpus were merged into a single file for each writer, creating an individual character file for each scriptwriter with data spanning three years' worth of broadcast material. For each writer, I also created an "Other" corpus, which was the dialogue of all the remaining characters combined. The total word count for the 20-Character corpus is 760,486 words (excluding the "Other" characters), and an overview of these 120 sub-corpora can be found in the spreadsheet in Appendix 2a, which gives details of the word count, turn length, and lexical richness of each character within each corpus.

For the word-n-gram-based test, a separate corpus was made, derived from the 20-Character corpus. Firstly I created six All-Character corpora, one for each individual writer. Each All-Character corpus consisted of all the lines of dialogue, spoken by all characters (including the "Other" characters) for the years 2010-2012. Each of the six writer corpora was standardised to 138,826 words, to match the lowest word count of the six corpora. This allowed for a normalised comparison. Secondly, I created 10 separate character corpora for each writer. From each writer's 20-Character corpus, ten smaller corpora were created for each writer. Each of the 10-Character corpora comprises one character's dialogue written by an individual writer in the 2010-2012 period. The same ten characters were used for all six writers, selected by word count, using the ten characters with the highest minimum word count of any one writer. Each individual character corpus for each writer was cut to the length of the smallest individual writer-character corpus, which was 2804 words. This allowed for a comparison of raw frequencies, but did have the disadvantage of substantially reducing the amount of data available for analysis.

4.4 Methodology for Chapter 5: A Quantitative Exploration of Linguistic Identity Disguise

This quantitative stylistic analysis addresses the first of my sub-questions, asking to what extent do quantitative, structural-level analyses identify character style rather than authorial style? Four separate tests were carried out using the 20-character corpus as a basis. Firstly, a word-n-gram-based analysis was carried out for each of the authors on 10 characters. Then, three separate tests were carried out on all 20 characters from the 20-Character corpus. These were: Average Word Length, Average Sentence Length and Vocabulary Richness.

4.4.1 Word-n-gram Methodology

Using AntConc 4.0.11, a 2-gram word-level analysis was carried out on Writer 1's All-character corpus, returning all bi-grams with a minimum frequency of 20 occurrences in that writer's All-character corpus. The bi-gram results showed, in descending frequency, how many times each bi-gram was used by Writer 1, down to the threshold of 20 occurrences. This was repeated for the remaining five writers to compare how many writers used each particular bi-gram. Bi-grams unique to individual writers are presented in Table 10 in Chapter 5, and the full results are in Appendix 2c. The bi-grams were colour-coded, using Excel's conditional formatting. Bi-grams which were used by four or more authors were coded red; bi-grams used by three or four writers were coloured orange; those used by two writers were in a green font, and any bi-grams used by only one writer were in green font with a green fill. The threshold of a minimum of 20 occurrences created a way of finding which bi-grams were used repeatedly by only one or two authors, and were thus potentially indicative of authorial style. A weakness of this threshold is that a bi-gram could be identified as unique to one writer, because it was used 20 times by that author, yet could occur 19 times in another writer's All-character corpus, but would not be counted. The results were then compared in a table, which showed how many times each bi-gram or tri-gram occurred. From this table, bi-grams which were used by only one or two writers selected for further investigation. The bi-grams were selected subjectively, if they seemed suggestive of interesting features of characterisation. The 10-character corpora for each author were then examined, making concordance plots of selected bi-grams and tri-grams to explore whether the n-grams which were used by only one or two of the six writers were associated with particular characters (i.e. potentially indicative of character style) or whether they were used across a number of characters (i.e. potentially indicative of authorial style). The figures for concordance plots for use of n-grams by individual writers are presented as tables, rather than plots for two reasons. Firstly, the figures for individual character use are generally low, so the concordance plot often had only one or two results per writer, and was more clearly displayed in a table. Secondly, the ordering of the characters does not affect the interpretation of the results, compared to, for example, an n-gram analysis which tracks the change in frequency of a particular n-gram throughout a novel.

This whole process was then repeated using tri-grams. The reason for carrying out this analysis was to demonstrate, in an explicable way, if certain n-grams were used persistently by a writer, regardless of which character's voice they were writing (which could therefore be viewed as authorial style), or if they used certain words, or collocations of words, for particular characters (which could be seen as character style).

4.4.2 Average Word Length and Average Turn Length

The next two tests carried out were Average Word Length and Average Turn Length. Along with Vocabulary Richness, these were categories of textual measurement reviewed in Grieve (2007) and are some of the most frequently used variables for authorship analysis. The eight types of textual measurement Grieve identifies are:

(i) Word length
(ii) Sentence length
(iii) Vocabulary richness
(iv) Grapheme frequency
(v) Word frequency
(vi) Punctuation mark frequency
(vii) Collocation frequency
(viii) Character level n-gram frequency
(Grieve, 2007)

From the eight types of measurement discussed by Grieve, three common style markers were analysed: Word Length, Sentence Length and Vocabulary Richness. These are the first three categories reviewed in Grieve (2007) and are some of the most frequently used variables for authorship analysis. My reason for selecting these is that, of the eight categories analysed in Grieve (2007), Word Length, Sentence Length and Vocabulary Richness are the variables which correlated with sociolinguistic profiles, where a higher measurement has been posited as more prestigious and is associated with higher social class and education (Culpeper, 2014; Nini, 2014). These selected features are, however, not the features with the highest attribution success rate in Grieve's analysis. Grieve's results showed that the most effective markers for a quantitative analysis were word and punctuation mark profiles, and character-level n-grams. Discussing punctuation mark profiles, Grieve observes:

> Overall, the frequency of individual punctuation marks is therefore one of the most potent quantitative indicators of authorship, despite the fact that this measurement has rarely been analyzed in attribution studies. Punctuation mark frequency is probably a good indicator of authorship because there is so much opportunity for variation in usage (2007:262)

Although punctuation performed extremely well in Grieve's analysis, it is excluded from my analysis, for two reasons. Firstly, during the production process, this is the feature most likely to be altered while the script is being edited and formatted for studio: punctuation changes are not specifically audible, so need lower levels of authorisation than is required to change lines of dialogue. Secondly, the nature of the data is written in order to be spoken – it is intended to be heard rather than read, so the thesis focuses on those parts of language which the intended audience can directly hear, rather than a textual feature that could not be inferred with any accuracy by a listener.

Character-level n-grams were also excluded. When evaluating character level n-grams, Grieve concludes:

> The n-gram algorithms are some of the most accurate techniques tested in this study. The most accurate n-gram algorithms are those based on the frequency of sequences of two and three characters: the 2- and 3-gram algorithms can distinguish between two possible authors with 94% accuracy, and can distinguish successfully between up to ten possible authors. (2007:262)

Despite the accuracy of character-level n-grams in determining authorship, this feature was not analysed here, because, unlike a word-n-gram-based analysis, it lacks a link to any particular trait of characterisation. A particular result cannot be said to correlate with any particular characteristic of the fictional personae, in the same way that a higher type-token ratio might suggest higher levels of intelligence or education.

Following Grieve, Average Word Length was "calculated by dividing the total number of digits and graphemes in a text by the total number of words" (2007:252). A spreadsheet was created recording the number of words spoken by each individual character from the 20-Character Corpus, for each of the six writers (Appendix 2a). Words were counted using a Regex code and the *Find All* function in Notepad++. I followed Grieve's definitions of terminology as follows:

> A *character* is an indivisible textual unit, including graphemes, digits, punctuation marks, and whitespaces; a *grapheme* is a letter of the alphabet; a *word* is a continuous string of graphemes and/or digits; a *sentence* is a continuous string of characters, excluding question marks, exclamation marks, newlines and nonabbreviatory periods (2007:252).

Next, the number of digits and graphemes for each individual character for each writer was calculated, also using Regex code and Notepad++'s *Find All* function. Following Grieve, I then calculated the Average Word Length for each character, as written by each of the six writers by dividing the total number of digits and graphemes in each individual corpus by the total of number of words for each. Full results are presented in Appendix 2a. Using SPSS Statistics Version 26, a plot of marginal means was created, and the results were viewed descriptively (discussed in Chapter 5).

This process was repeated to measure Average Turn Length. The second category in Grieve's article is Sentence Length measurement. Grieve measures this in four ways: average sentence length; sentence length in characters; average sentence length in characters, and fourthly, sentence length distribution in characters. Again, one of these measurements – Average Sentence Length in words – is used as a representative measure to explore this category, and is "calculated by dividing the total number of words in a text by the total number of sentences" (2007:252). In Grieve's *Telegraph* corpus, the text uses standard sentence structure. In *The Archers*, the dialogue emulates spoken language and is written to be spoken, so there were frequently turns which were written not using standard sentence structure, unlike Grieve's corpus of *Telegraph* columns.

Dividing the data by turns provided a much clearer, less subjective way of dividing it into units. In a drama script, the simplest way to delineate an utterance is as turn-length. Number of turns was the unit of frequency, rather than number of words, on the basis that each turn is the core 'unit' of speech, because in most analyses of conversational interaction, including, most obviously, Conversation Analysis, speaker turn is the basic unit of analysis (Sidnell, 2010). Herman (1995) argues that:

> The linguistic units of analysis appropriate to dialogue as interactional speech are utterances. The sentence is an abstract entity in linguistics, defined in relation to particular grammars, and not in absolute terms. Utterances bring back into reckoning the contextual factors which are abstracted away by grammatical sentences. (1995:13)

As Herman argues, turns can be identified more accurately, because spoken language (or language that has been written to emulate speech) does not follow clearly marked sentence structures, as these examples, below, from my data demonstrate. Even within an individual character corpus from a single writer, there is inconsistency about what would constitute a sentence. This can be seen in the following three, separate, lines of dialogue, all written by the same author:

ELIZABETH: Yes... it wouldn t be easy.

(Writer 3, 2011)

ELIZABETH: Yes... I was just looking at the spires on the cathedral. Amazing.

(Writer 3, 2011)

ELIZABETH: Yes. At least, the children are in there now.

(Writer 3, 2011)

In the first line, the lower case ‘I’ in “it” suggests the author views this as a single sentence. In the third example, the capital “A” makes clear that it is a two-sentence line. The second line would be a subjective decision because the “I” would have been capitalised whether or not it was a sentence initial position.

In other cases, sentences were split across separate turns, with an interjection by another speaker, which makes it problematic to analyse sentence length, as in this example:

LILIAN Please. (KENTON POURS TONIC IN) And what makes it so infuriating..

KENTON (PUTS GLASS IN FRONT OF LILIAN) There we go.

LILIAN ..is that he's been right all along.

(Writer 1, 2011)

Using Turn length as a definition makes the process clearer, less subjective and far less time-consuming, which is therefore more practical when dealing with large amounts of data. The word counts from the Average Word Length test were used, and following Grieve, I divided the total number of words by the number of turns. Following Grieve (2007), Average Turn Length was calculated by "dividing the total number of words in a text by the total number of sentences" (2007:252). The number of turns was found simply by checking the line count on Notepad++, where each turn of dialogue appears on a single line. Again, these results were imported into SPSS to create a plot of marginal means.

4.4.3 Type-token Ratio

A Type-token ratio (TTR) test was used to explore the category of Vocabulary Richness. The type-token ratio is carried out using the formula "V/N", where "N is the total number of words in a text (i.e. word tokens)" and "V is the total number of vocabulary items in a text (i.e. word types)" (2007:252). As Grieve cautions, "the Type–Token Ratio is known to be very sensitive to text-length— as a text gets longer, new word-types are introduced at a slower rate" (2007:253). For this reason, the individual character corpus with the lowest word count was identified. This was Jazzer's dialogue, written by Writer 2, which was 695 words. To perform the test on corpora of equal sizes, only the first 695 words of data were used for all of the TTR tests. The TTR was carried out in RStudio, version 4.0.3, adapting the code in Jockers and Thalken (2020:82-84). To test data with higher word counts, a second TTR analysis was carried out using only the top 14 characters, standardised to match the lowest word count of these 14 characters, which was 2366 words. This enabled me to test longer sections of data for each character to compare results. Results for these quantitative analyses are presented and discussed in Chapter 5.

4.5 Data and Methodology for Chapter 6: Three Character Studies

4.5.1 Introduction

Chapter 6 studies prominent linguistic features associated with the dialogue of three distinctive characters within *The Archers.* The first of these three character studies, exploring the lexis of Jim Lloyd, sub-divides into a further three short analyses, taken from Culpeper's dimensions of lexical characterisation (discussed more fully in Chapter 2). The corpora I used are described here, followed by an explanation of the methodology used for each of these studies.

4.5.2 Lexical Richness: Data and Methodology

Lexical Richness was the first analysis carried out. This feature was analysed only briefly, because type-token ratio has been considered in more detail in Chapter 5. Using the *Archers*-Complete Corpus, the text of Jim's dialogue (cleaned of stage directions) broadcast between 2010 and 2015 were copied into six new files, one for each writer. This is the "Jim" corpus, shown in Table 5.

Table 5: Tokens per Writer in "Jim Corpus"

Writer 1	Writer 2	Writer 3	Writer 4	Writer 5	Writer 6
9953	2911	9796	5486	9171	4733

Using the same method as in 4.4.3 the Type-token ratio was calculated. However, rather than using the whole corpus to calculate the type-token ratio, I used the first 500 words and the last 500 words of each Jim corpus. The reason for this was because Jim was a relatively new character in 2010, so this test looked for any differences between Jim's vocabulary richness as a new character, compared to his vocabulary richness as an established character. The decision to use the first 500 and the last 500 words was based on the distribution of the data throughout the five years. Whilst the number of words per year in the Jim Corpus was not equal, using the first and the last 500 words allowed for a comparison between 2010-2011 and 2014-2015 for all writers. Increasing the word count in each TTR analysis beyond that meant that for Writer 2, whose Jim corpus had the fewest words at 2911, would then have included dialogue from 2012-2013, thereby losing the distinction between earlier data and later data. Although the most recent data are from 2017, for four out of six writers, 2015 was the latest year in which Jim appeared in their corpus, so 2015 was used for all writers to make a closer comparison. The results were compared to look for patterns of variation between the earlier and later data.

4.5.3 Keyword Analysis: Data and Methodology

For this keyword analysis of Jim's vocabulary I compiled all of Jim's lines written by all six scriptwriters combined. This was the "Jim Target Corpus". Next a "Jim Reference Corpus" was created, by copying and pasting the cleaned dialogue of all the other characters, but excluding Jim, from the same period, written by all six scriptwriters, into one single document. AntConc 3.5.9 was used to carry out a keyword analysis, to identify keywords in Jim's dialogue (as written by the six writers combined), compared to the reference corpus of the remaining characters. The purpose of this was to generate a list of keywords which are particularly associated with Jim's amalgamated character as portrayed throughout the programme. I then carried out a keyword analysis for each individual writer's Jim lines, using the same Jim Reference corpus, to generate a keyword list for Jim as written by each of the six individual writers. Each writer's keyword list was compared in turn to the original keyword list of Jim's keywords, as written by all six writers combined. The purpose of this was to explore whether any of Jim's keywords could be associated with any particular writer, or whether a number of writers linguistically gravitated towards each other, using similar keywords to Jim's dialogue.

These keyword lists generated were used to identify and compare words which conveyed information about character style, in contrast to those words which arose through situation. These results were used to identify notable tokens for further, qualitative exploration, rather than to produce a statistical analysis.

4.5.4 Germanic Versus Latinate Lexis

The final analysis of Jim's lexis considers whether Jim's vocabulary consists of Latinate or Germanic words. Following Culpeper (2014), I use Latin as a proxy for formality. Whilst other etymological roots, such as Ancient Greek words, could also be indicative of formality, Culpeper argues that using Latinate words is a way of gauging the formality of a text. Deciding which words are formal or prestigious is subjective, but using the method of selecting words of Latin origin allows for a systematic and replicable method, although the interpretation of results is more subjective. In *The Archers*, Jim is one of the most highly-educated characters, so a relatively high frequency of Latinate words might be expected in his speech. However, this method is not without problems: for example, the word "mobile" is used by the character, Eddie, in reference to his phone. The word "mobile" is Latin in origin, but is such a standard way of describing a phone that it cannot be viewed as a prestigious or formal.

For each of the six writers, the first 500 words in the corpus were taken, and compared to 500 words from 2015 to compare the relative levels of Latinate words. As for the lexical richness analysis in 6.2.1, using 500 words was chosen because it allowed for a comparison of 'early' Jim and 'late' Jim, to look for diachronic change in character style. Increasing the word count analysed would have meant, for the writers with shorter Jim corpora, that the distinction between early and late would be lost, because data from the middle years would have been included in both the early and the late analysis.

To determine which words are of Latin origin, the methodology of DeForest and Johnson (2001) was used. DeForest and Johnson measured the density of Latinate words in Jane Austen's novels. They write:

> On the advice of statisticians, we excluded proper nouns, titles (such as Mr. or Captain) and common words. This last group comprises words that serve grammatical functions: auxiliary verbs, prepositions, pronouns, and articles—the so-called 'function words'. These are the most common words, and there are no alternatives for them. They are predominantly Anglo-Saxon. (2001:390)

As with DeForest and Johnson's study, Latin words which entered English vocabulary via French were included. They justify the decision:

> Readers cannot be expected to distinguish between Latinate words taken from the French and Latinate words taken directly from Latin. Even Samuel Johnson did not make the attempt. In the preface to his Dictionary he apologized: 'Of many words it is difficult to say whether they were immediately received from the Latin or the French ... It has perhaps sometimes happened that I have mentioned only the Latin, when the word was borrowed from the French'" (1977:281). (2001:392)

All vocabulary coded as Latinate has had its etymological source verified in the Oxford English Dictionary.

To compare these results to other characters of different sociolinguistic backgrounds, similar character corpora were created for two other characters: David Archer, a college-educated farm owner, and then Eddie Grundy, an uneducated character. David Archer and his wife, Ruth, are portrayed as the "Everyman" characters at the centre of the soap. The Grundys are portrayed as uneducated agricultural workers who just about scrape by, relying on money-making wheezes and a random assortment of minimum wage jobs. These individual character corpora were also limited to the first 500 words, to use a standardised length across the analyses. Based on their educational levels, Jim would be expected to have use the highest number of Latinate words, followed by David, then Eddie the fewest.

4.5.5 Jazzer's Dialect: Data

The second character study is of Jazzer's dialect, to explore how the scriptwriters write characters with a different dialect to their own. Jazzer was selected for detailed analysis because he is a long-standing Scottish character in a predominantly English drama written by non-Scottish scriptwriters and produced by a mostly English production team. A sub-corpus was compiled of all of Jazzer's lines for each of the writers during the seven-year period of available scripts (2010-2017). Using Regex, and the *Find and Replace* function on Notepad++, all of Jazzer's lines were extracted and compiled into six author corpora, forming the Jazzer corpus. Only Jazzer's lines were analysed, not the speech of his interlocutors. Although this approach meant excluding an analysis of whether Jazzer adapts or moderates his speech depending on context, it allowed for a cleaner comparison of his speech by the different scriptwriters. Once certain lexical or grammatically notable features had been identified, I referred back to the original scripts to analyse the dialogue in the context of its original setting with the other characters.

4.5.6 Jazzer's Dialect: Methodology

To examine the way in which the writers wrote a character whose dialect was different from their own, my analysis followed similar methodology used in Ruzich and Blake's analysis of *The Help*. Whilst Ruzich and Blake's analysis considered different characters written by the same author, my analysis considered the same character, as written by different authors. The shortest of these individual files from the Jazzer corpus was 1938 words. In the first instance, a quantitative analysis of each corpus was carried out: this was identified at the levels of pronunciation, lexis and grammar, using the key features of Glaswegian identified in 3.4. The aim was not to run a statistical analysis, but simply to compare the frequency of dialectical features for each writer. Since a complete list of Scots, Scottish English and Glaswegian features would be impossible, there were further subjective decisions when this was compiled: when features occurred in the text which did not read as Standard English, these features were investigated using existing dialectology research to identify whether or not the non-standard English features were associated with Scottish dialects.

Secondly, the same analysis was carried out on the first 1938 words of dialogue of a *River City script* for a comparison. Only one episode script was available online, which necessitated using dialogue from whichever characters happened to appear in this episode, rather than using only those who matched the sociolinguistic profile of Jazzer, a working-class male in his late twenties / early thirties. *River City* is a soap opera produced in Glasgow, so is a close match for text-type, although the main audience is different: *River City* is available online throughout the UK, but is broadcast in Scotland only, so the primary audience is in Scotland. After my initial quantitative analysis, a qualitative intra-author analysis of each writer's corpus was carried out, to explore how consistently they wrote Jazzer's dialogue (in some cases over a seven-year period); to identify any linguistic features which were realised inconsistently, or were unconvincing in their realisation of the Glaswegian dialect.

Firstly, the number of dialectical features in a standardised length of dialogue by each writer was counted. These were divided into grammatical, pronunciation and lexical. The aim of the quantitative analysis was to get an overview of the relative levels of dialect features used by the authors. Whether these features occurred in the lexis, grammar or 'eye dialect' to denote pronunciation was of secondary interest.

The word "no" was coded as a lexical dialect feature when it was used in the Scots sense of "not", as in, "I'm no gonna do that." One of Tulloch's criteria for including an item as Scottish lexis was that it was used in a different way from Standard English use, which is the case when "no" is used to mean "not". Also, the enclitic ending, "-nae" was counted as lexical. As discussed in Pust's article on Scottish negation, "I cannae see it", there are a number of options for negating in Scottish English, so using words such as "hadnae" becomes an authorial vocabulary choice: it is not simply the Scottish form of a contraction, or a guide to pronunciation.

Instances such as "I'm away to find my wrench" were coded as grammar, because all the words in the sentence are used in the same way as in Standard English, but the absence of a verb such as "going" marks it out as being non-Standard English, and the use of a present tense "I'm" to describe a future action of going away is a Scottish feature. There were also occurrences of non-standard grammar such as "my heart's been broke" which were included as grammatical examples of dialect. It could be argued that this might also occur in non-standard English dialogue: for example, the working-class Grundy family often deviate from standard English grammar. Such examples were included because they are reflective of Jazzer's speech patterns, even if they are not exclusive to Jazzer.

Instances of eye dialect, such as "pint o' semi" were counted as examples of pronunciation. Further, there were some features of eye dialect, denoting pronunciation which were not specifically associated with Scottish English. "Gonna" was used by Jazzer, but was also used by many English characters, such as the Grundy and Carter families, It was still included in this analysis because it shows the extent to which the writers conceive Jazzer's dialect as an "other", non-standard dialect. As Stockwell (2021) argues, writing in dialect such as "sed" to denote a poor character demonstrates social deixis because this is largely how most people would pronounce the word. There were also occurrences of a dropped final 'g', such as "buzzin, me". Again, these were included because they demonstrate the extent to which Jazzer is perceived as "other".

On analysing the data, relatively few dialect words were used, and they tended to be the same small number of tokens. It would be impossible to pre-select a clearly defined lexicon of Glaswegian speech, since there is an enormous overlap between the two languages. As such, the lexical analysis was subjective and included any lexical items which I (an English-born researcher) deemed to be Scottish. These were then checked against the Scottish National Dictionary before being included. Following this process meant there was a high accuracy that tokens were not incorrectly added (i.e. that only Scottish words, as defined by the Scottish National Dictionary, were included); however, the weakness in the method is that Scottish words may not have been recognised as such, and therefore been incorrectly excluded from the count. The results for this analysis are presented and discussed in 6.3.

4.5.7 (Im)politeness of Lynda Snell: Data

Using existing literature on politeness (Brown and Levinson (1987) and Culpeper (1996, 2005, 2011), and relational work (Locher, 2006, Locher and Watts, 2005), I analysed the politeness strategies used by Lynda Snell as director of the village pantomime. I compiled the Lynda-corpus from the *Archers*-Complete corpus. Scenes with Lynda in were identified using a search function on Microsoft Word. Each scene was inspected, and any scenes which contained Lynda directing a show were copied into a new document. A separate file was created for each writer. Using whole scenes, including stage directions, enabled me to analyse Lynda's dialogue in context, considering factors such as relative power, social distance, and plot development.

The plot-based requirement for inclusion in the Lynda corpus was quite tightly defined (Lynda directing a show), so it was unsurprising that there were unequal amounts of data for each writer. For example, in 2010, Writer 2 wrote scenes where Lynda was directing a show, but did not write any more episodes after that which included Lynda directing a show. However, as this was a qualitative analysis, the differences in size of data were not problematic.

4.5.8 (Im)politeness of Lynda Snell: Methodology

I analysed scenes where Lynda is carrying out the same activity, directing a show, because (im)politeness is strongly linked with interaction (Bousfield, 2008, Culpeper, 2001, among others), and is heavily context-dependent (Culpeper, 2013). Therefore, using scenes in which Lynda is carrying out the same activity allows a comparison of closer text-type, which is a well-established principle in authorship attribution (e.g. McMenamin 1993, Grieve, 2007). This also relates back to Ohmann (1964), who conceives style as "a way of doing it" (p.426). By keeping the "it" the same (directing a show), I was able to focus on the "way" Lynda set out to achieve her interactional goals, although as Ohmann cautioned, separating form from content in literature is complicated (1964:427). Even with this measure of control, the nature of the data means that the exact circumstances of each directing scene will never be entirely comparable: it would be expected that each year, the village show storyline has a different plot and tone from previous years.

Once the corpus was created, I analysed it qualitatively, drawing on the (im)politeness literature discussed in 3.5. I did not formally code the data into different strategies, for example hedges, or apologies. One of the issues of coding politeness and impoliteness features is the necessary subjectivity of coding. Culpeper (2009) warns that, "The most heinous crime when performing an analysis of impoliteness strategies, or politeness for that matter, is to simply count them up on the assumption that if the strategy is there, it necessarily is performing impoliteness". Whilst the selection of only Lynda's scenes in the rehearsal controlled the context to an extent, there is, of course, still a great deal of variation between the different scenes. Rather than attempting to force a codified framework on the analysis, the results and discussion in Chapter 6 explore how the different scriptwriters carry out facework – or in Locher's terminology, *relational work* – and some comparison of the differing approaches. The study focuses on the directives Lynda gives to her cast, in particular her use of imperatives. The results for this analysis are presented and discussed in 6.4.

4.6 Data and Methodology for Chapter 7: The Functions of *oh*

4.6.1 The Functions of *oh*: Data

My third study focuses on the multifunctional meanings of *oh*, to explore the different functions it has in the text, and to explore whether these functions vary between writers and between characters. A corpus was created containing duologues between two characters, Helen and Rob. This was taken from the *Archers*-Complete corpus, which contains stage directions, and also the character names, showing which character is speaking each line of dialogue. The Helen and Rob corpus was created by using Regex, and the *Find All* function on Notepad++ to extract all lines spoken by Helen and Rob. Additionally two smaller corpora were created, both containing duologues between two couples. These were the 'Lilian and Paul' corpus, and the 'Elizabeth and Roy' corpus. These two smaller corpora were taken from the same time period.

The “Script As Broadcast” annotations had already been removed. It is possible that some of uses of *oh* could have been added by producers at the script edit stage, which would not be marked. However, there is no reason to believe that *oh*, or other pragmatic noise items, would be a particular area of focus during script editing, which tends to concentrate on plot continuity and dramatic impact, rather than pragmatic markers. If some instances of *oh* were added by the production team, the high frequency of *oh* overall reduces the impact of any occasional interference with the data.

For authorship attribution, comparing closer text-types is preferable (Grieve, 2007), so using three corpora where the couples are having affairs allows a comparison of character and authorial variation, rather than a comparison between different situations and text-types. This was why three sets of duologues between couples was used, rather than the whole *Archers*-Complete corpus. Within the context of comparing conversations between couples, choosing three couples in new, and clandestine relationships uses closer text types than comparing the conversations between, for instance, a very new relationship and a decades-long marriage. Of the three sets of duologues, the Elizabeth and Roy corpus has a number of work-related conversations because the characters began their affair at work, so differences in language might be partly attributable to a variation in conversational setting, rather than an inter-character variation. The reason for choosing couples who are having affairs is also a practical one: because the couples are being secretive, there are multiple scenes where the couples are on their own and speaking only to each other. There is of course, a difference in emotional and dramatic situations between the scenes, as would be expected in any drama, but the use of scenes with only two characters, speaking privately to each other, aims to reduce these other factors as far as possible.

Scenes were selected only if they were a duologue between Helen and Rob, with the exception of the presence of Helen’s toddler son, Henry. In *The Archers*, the voices of young children are usually kept to a minimum, so Henry often has a few lines to establish his presence at the beginning of a scene, or is present in the scene, but without any dialogue, as in this example:

```
4. INT. BLOSSOM HILL COTTAGE. LIVING ROOM. 11.20A.M.
```

```
OFF: HENRY IS IN THE HALL, PLAYING WITH HIS LEGO.
OCCASIONAL PLAY NOISES THROUGHOUT. HELEN IS SITTING
ON SOFA, UNREAD BOOK ON HER LAP. ROB HAS JUST COME
IN).

HELEN     He’s not in the way out there, is he?

ROB       Henry? No, no, no. He’s quite happy,
          playing.
```

(Writer 1, 2015)

Henry’s lines, or any use of *oh* spoken only to Henry was deleted. This was done manually by reading through the corpora. Of the three corpora, the Helen and Rob corpus is significantly larger. The total words count for each corpus, by writer, is shown in Table 6 below.

Table 6: Corpora word count

	Total words	**Helen & Rob**	**Lilian & Paul**	**Elizabeth & Roy**
Writer 1	15118	9925	2547	2646
Writer 2	9368	5788	1370	2210
Writer 3	12807	8559	3570	678
Writer 4	13247	7270	3021	3428
Writer 5	12643	5663	6367	613
Writer 6	15709	7437	4251	4021

The variation in corpus size between writers does not reflect authorial choice, because writers are issued with storylines documents stating which storylines they will write.

4.6.2 The Functions of *oh*: Methodology

All occurrences of *oh* were extracted from the three corpora, using Regex to extract every line with the word *oh* in. In addition to every occurrence of *oh*, variant spellings such as *ohhh*, were also included in the count, since the pronunciation is the same. These were identified by reading through the complete corpora and extracting them manually. *Ooh* or any extended version (e.g. *oooh*) was excluded because its pronunciation, rather than just its prosody, is different. As in previous studies, I referred back to the *Archers*-Complete corpus when contextual detail was required.

Several tests were carried out. Firstly I carried out a simple, quantitative comparison of the frequency of *oh* to see if this discriminated between writers. Secondly I broke this down by character, to analyse whether writers varied in the use of this token, depending on which character they were writing. This was firstly done on the Helen and Rob corpus, and then repeated for Lilian and Paul, and then for Elizabeth and Roy. Having reviewed these results, I then coded the data by function, to attribute a different function for the use of *oh* to each occurrence in the corpora, to explore whether it was possible to discriminate between more pairs of authors, when taking into account the function for which *oh* was being used, rather than simply by counting its frequency. This was to address the third of my sub-questions, investigating whether higher-level methods of analysis, such as pragmatics, are better able to discriminate between authors than structural level features.

Using the existing academic literature on *oh* (discussed in 3.6), and observations from the data, the different pragmatic functions of *oh* were codified into six broad main categories, to analyse differences in usage by writer and by character. The breadth of approaches to analysing *oh* and the multitude of functions in the literature attributed to *oh* made this a complex task. As Aijmer (2013:1) observes:

> Research on pragmatic markers has avalanched in recent years and pragmatic markers have been promoted to a major area in pragmatics as shown by the large number of approaches devoted to the topic. The approaches are synchronic and diachronic, formal and informal. (2013:1)

From the literature reviewed in Chapter 3, there are differences in the approaches: Macaulay from a variationist sociolinguistic study, Heritage placing emphasis on sequencing and turn-taking; Schiffrin on discoursal structuring, and Aijmer using a corpus approach to analyse the functions of *oh*. For authorship analysis, being able to code the functions into discrete functions is desirable. Macaulay's categorisation of *oh* was initially used as a framework, because the variationist approach he used necessitated discrete categories which was helpful for coding purposes. I initially coded the data according to Macaulay's five categories, but when attempting to code the data, I found that this did not adequately account for all occurrences in the data, so I retained Macaulay's categories of Acknowledgement, Agreement and Emotion, but also included 'Surprise' and 'Downplayer'. These final categories were reached as the result of a 'bottom-up' approach. I followed Aijmer's (1987) methodology, where I studied the data, identifying individual uses, and then looked for similarities which could be grouped to form the additional categories. These supplemented the original categories.

Initially I used the following five categories for the functions of *oh*.

1. Emotion / vocative, e.g. "Oh, Rob", "oh dear"
2. Agreement, e.g. "oh yes"
3. Surprise: linguistic and non-linguistically initiated.
4. Acknowledgement / back channel *oh* / receipt of day-to-day information
5. Downplayer / casualness / spokenness

Macaulay's category of 'quoted dialogue' did not arise in the data, so was discounted.

There is, of course, an element of subjectivity in this process, so using the test-retest method, I checked the reliability of my coding. Randomly selecting Writer 3, I coded all 61 instances of *oh* in Writer 3's corpus. I repeated this exercise 24 hours later to test the reliability of my coding. I coded 48 of the occurrences the same but 13 differently, giving a reliability rate of 79%. It is likely that with two separate coders, or a longer timeframe between the test and re-test the score would be even lower. I compared the results to see which categories had been coded differently on the second test than the first. This comparison of differently coded occurrences is shown in Table 7, below.

Table 7: Inconsistencies found in 1st Test-retest reliability results

First coding	Second coding
3 Surprise	4 Acknowledgement
3 Surprise	4 Acknowledgement
3 Surprise	4 Acknowledgement
3 Surprise	1 Emotion / vocative
1 Emotion / vocative	3 Surprise
1 Emotion / vocative	5 Downplayer / spokenness
5 Downplayer / spokenness	1 Emotion / vocative
5 Downplayer / spokenness	1 Emotion / vocative
3 Surprise	1 Emotion / vocative
1 Emotion / vocative	3 Surprise
3 Surprise	4 Acknowledgement
3 Surprise	1 Emotion / vocative
5 Downplayer / spokenness	1 Emotion / vocative

As can be seen, Category 3 "surprise" and Category 4 "emotion" were the most frequently inconsistently re-coded. There were four instances of surprise and acknowledgement being confused. To improve the reliability of this coding, I made some adjustments: Category 3, 'Surprise', was coded specifically on the sense of an interruption – either thought or presence. If the information itself was surprising but the structure of the delivery was not, it was coded as acknowledgement, to avoid the inconsistency arising from decisions about whether a response crossed an invisible threshold from acknowledgement to surprise.

To reduce confusion between Category 4 (emotion) and Category 5 (spokenness), the definition for Category 1 was tightened to include only conventionalised phrases ("oh no!" "oh gosh"), or instances of *oh* plus a name, relating to the vocative function described by Culpeper and Kytö as originally coming from the separate word "O". It also included set phrases, such as "oh no" and "oh my goodness". In almost all cases, the *oh* acts as a lesser-stressed grace note, adding emphasis to the second word in the phrase. With the exception of “oh my goodness” and “oh dear”, the oh could be deleted and the sense of the statement retained. I carried out a second test-re-test exercise 48 hours later, in which I found that 50 out of 61 occurrences of *oh* were coded the same way (giving a reliability rate of 81%). By tightening the criteria for Category 3, this improved test-retest reliability between Categories 1 and 3, although it did not significantly improve the overall reliability rate. Those occurrences of *oh* which were coded differently are shown in Table 8:

Table 8: Inconsistencies found in 2nd Test-retest reliability results

First coding	Second coding
4 Acknowledgement	3 Surprise
5 Spokenness	3 Surprise
5 Spokenness	3 Surprise
5 Spokenness	3 Surprise
5 Spokenness	3 Surprise
5 Spokenness	3 Surprise
4 Acknowledgement	1 Vocative / conventionalised
5 Spokenness	1 Vocative / conventionalised
4 Acknowledgement	1 Vocative / conventionalised
4 Acknowledgement	3 Surprise
2 Agreement	5 Spokenness

The remaining most frequent inconsistency was caused between Category 3 (surprise) and Category 5 (spokenness), followed by inconsistencies between Category 4 (emotion / vocative address) and acknowledgement. To tighten these criteria further, I decided that an *oh* which followed a non-linguistic event (e.g. another character arrives in the scene) was not necessarily coded as “3. Surprise” following Aijmer’s sense of interruption to the train of thought, unless there were an exclamatory tone (often written with an exclamation mark.”). Anything in which the speaker “interrupts” themselves with a new thought retained its coding as “3. Surprise”. Secondly, a prioritisation scheme was used (to improve consistency of coding, rather than suggesting certain uses of *oh* are more significant than others). If an item potentially belonged to the group “1. Conventionalised phrase / vocative”, this was selected first. Then, if a group could belong to the category “2. Agreement”, this was selected next, and so on. Finally, *oh no* was only coded as “1. Conventionalised phrase / vocative” if it could be substituted for another empathetic opinion, e.g. *oh my goodness*, and the line retain its sense. Or, to put it another way, *oh no* was only included in Category 1 if it was an expression of empathy or shared horror / disgust / fear or similar, but not if it was an answer to a yes/no question.

Using these tightened criteria, I tested the data sample a third time (48 hours apart), achieving a retest reliability rate of 89%. Using these finalised criteria for coding, I then carried out a test-re-test on Writer 1’s corpus, in case part of the improved reliability rate was due to remembering my previous coding choices for the Writer 3 data. When I carried out a test-re-test on the data for Writer 1, 24 hours apart, a reliability rate of 92% was achieved.

Following the test re-test reliability process, the final categories used are as follows:

1 Conventionalised phrase / Vocative
2 Agreement
3 Surprise (in the sense of interruption)
4 Acknowledgement
5 Downplayer / spokenness / hesitation
6 Continuer / topicaliser.

The results from these tests are presented and discussed in Chapter 7, where I also provide examples of these different functions of *oh* using examples from the data.

4.7 Summary of Corpora

Table 9 summarises the individual corpora used in each of the studies which have been discussed in this chapter. All corpora are in digital format, and all corpora feature all six scriptwriters.

Table 9: Summary of Corpora

Name	Description	Years	Relevant Chapter
Archers Complete	1440 individual episode scripts. “Studio Script” version (all amendments to reflect broadcast version of episode manually removed).	2010-2017	All
Archers Writer corpora	For each writer, there is a folder with one document per year, containing all lines of dialogue spoken by all characters, written by that writer.	2010-2017	All
20-Character corpora	120 individual character corpora of the 20-most frequently speaking characters (combined) for the six writers. Character’s dialogue only, cleaned of stage directions and speaker name. For simplicity, this is called the 20-Character corpus, but also contains a file of “Other” characters for each writer, containing all dialogue of speakers outside the 20-most frequent speakers.	2010-2012	Chapter 5
All-character corpora	Six corpora (one for each writer) containing all lines of dialogue from all characters in the 20-Character corpora, including the “Other” characters. Standardised to 138,826 words for all writers to match the lowest word count.	2010-2017	Chapter 5, Chapter 7
10-Character corpora	The first 2804 words for each of the 10 most frequently-speaking characters from the 20-Character Corpus	2010-2015	Chapter 5
Jim corpus	All of Jim’s lines, divided into six sub-corpora; one for each writer.	2010-2015	Chapter 6
Jim Target corpus	All of Jim’s lines as written by all six writers combined.	2010-2015	Chapter 6
Jim Reference corpus	All of the characters in the data (taken from the 20-Character corpus) except for Jim	2010-2015	Chapter 6
Jim_500 corpus	Jim’s first and final 500 words taken from the Jim Corpus	2010-2015	Chapter 6
David_500 corpus	David’s first 500 words, taken from the 20-Character corpus	2010-2015	Chapter 6
Eddie_500 corpus	Eddie’s first 500 words, taken from the 20-Character corpus	2010-2015	Chapter 6
Jazzer corpus	Jazzer only lines (not standardised for length, but only the first 1938 words were used in the quantitative analysis)	2010-2017	Chapter 6
Lynda corpus	Whole scenes incl. stage directions where Lynda is directing the village show.	2010-2017	Chapter 6
Helen and Rob	Whole scenes featuring only Helen and Rob	2010-2017	Chapter 7
Lilian and Paul	Whole scenes featuring only Lilian and Paul	2010-2017	Chapter 7
Elizabeth and Roy	Whole scenes featuring only Elizabeth and Roy	2010-2017	Chapter 7

5. A Quantitative Exploration of Linguistic Identity Disguise

5.1 Introduction

This chapter investigates the first of my research sub-questions, which asks, to what extent do quantitative, structural-level analyses identify character style, rather than authorial style? As discussed in 2.4.6, the division between *authorial style* and *character style* is something of a complicated distinction because the two are not separate entities: in drama, authorial style is discernible *through* character style. As defined in 2.4.6 I use *authorial style* to mean those linguistic features which are consistently used by an author regardless of which character they are writing, and by *character style*, I mean those linguistic features which a writer uses only for certain characters. In terms of my superordinate research aim, I use the quantitative analyses to investigate whether consistent features of authorial style are found across multiple characters. This may be suggestive of the linguistic leakage found by Grant and MacLeod in their data (2020:78), where the writer has not suppressed linguistic features associated with their authorial style.

First, I conduct a word-n-gram-based test to identify word n-grams which are used by only one or two of the authors. N-grams which I judged to be relevant to character style were investigated further. Using concordance plots on AntConc, these are then investigated by character, to discover whether those n-grams are used by a range of characters, or only by certain characters. If the n-grams are evenly distributed across characters, these n-grams could be suggestive of authorial style: if the n-grams are clustered in the dialogue of one, or only a small number of characters, these n-grams are likely to be suggestive of character style. This analysis is followed by three separate tests: Average Word Length, Average Turn Length and Type-token ratio, which investigate whether the scriptwriters' use of structural-level features, such as word length, vary depending on which character's voice they are writing.

The aim of the quantitative analysis is not to attempt an exhaustive comparison of how all known style markers perform in cases of adversarial or fictional writing. Such an aim would be impossible: in 1998, Rudman estimated that over 1000 style markers had been proposed in authorship analysis tasks. Reflecting on this in 2012, he wrote, "With Suguru Ishizaki and David Kaufer's Docuscope inventory of over forty-one million language strings, and other additions, the total number of style markers has become moot – approaching infinity, which Richard Forsyth and others claim" (2012:267-8). Instead, this chapter explores some of the common style markers used in authorship attribution cases to explore whether measurable features of language which occur in all writing, are altered, depending on which character's voice is being written.

5.2 Word-n-gram-based Results

As discussed in Chapter 4, a word-n-gram-based test was carried out for the six writers on the 20-Character corpus, analysing bi-grams and then tri-grams. The bi-grams results for each writer in turn are discussed here, followed by the tri-grams results for each writer.

5.2.1 Bi-grams Results

Table 10 shows the bi-grams which were used only by one or two of the six writers, 20 times or more in the data. The full bi-gram results are in Appendix 2c.

Table 10: Bigrams used by only one writer

Writer 1	Writer 2	Writer 3	Writer 4	Writer 5	Writer 6
o k	ice cream	the car	lots of	i'm glad	it's fine
a real	of tea	him and	at home	the baby	look i
ah ha	the echo	any of	they can	a baby	yeah yeah
before i	has been	it seems	it looks	the funeral	i s'pose
in for	bridge farm	i dunno	the place	much more	yes all
one or	e coli	me oh	better get	i like	it now
told him	you remember	love you	he's a	lot to	just want
this year	anyway i	the weekend	i'll just	see it	the business
much better	the thing	can we	of thing	to them	now that
a bad	and he's	grey gables	thinking of	the family	suppose i
before you	i don't	have had	with them	been so	just got
in on	oh that	ok i	is this	didn't think	just need
too bad	out to	and as	so it	even if	the website
how long	just to	they're not	some time	glad you	were going

not too	oh look	you yeah	the office	oh mum	me what
say anything	oh my	are going	uh huh	that but	not as
at a	in and	to eat	how was	yes so	right so
before we	oh don't	you've done	that sort	i hadn't	spoken to
each other	the board	and all	that's nice	much of	does it
lot more	what it	to meet	we all	said it	er yes
with us	and well	well we	at it	she might	only just
back on	told her	yeah so	be happy	you it	the garden
in touch	and she's	good to	he had	course it	well don't
a moment	but they	that's ok	looking at	doesn't want	you been
a second	no idea	you're doing	on earth	i understand	you later
on his	some kind	about your	out for	it does	you up
we got	to buy	are then	planning to	not very	but no
you're sure	what have	but he's	after the	made a	get out
as the	about what	have it	anything else	maybe you	it's okay
better than	afternoon oh	having to	haven't you	see him	might as
for us	full of	i love	show you	them i	not what
he says	give it	i love	so it’s	to i	I'm only
her own	he just	in ambridge	when you’re	don’t see	just saying
i'm all	her a	interested in		him he	no you're
let's get	how it	it like		i never	on i
ready to	my goodness	it's no		is i	speak to
take the	not yet	parish council		long time	you two
up i	said you	the parish		very happy	how you
we've been	tell her	a job		yes she	in your
what if	what it's	all day		you’d be	okay i
back from	you that	as much		at lower	really well
chance of	a call	be fine		be doing	the money
i came	a really	bye bye		but she’s	ah well
if the	and there	don't suppose		might not	as though
if there's	at this	i'll go		she won’t	have i
was quite	dad and	i'm fine		didn’t have	is everything
way to	do for	no and		do something	it were
we want	he does	no thanks		find out	I'm still
a quick	i went	right it's		for all	lovely to
and dad	just had	right now		good heavens	the cows
anything to	on Friday	see that		isn’t he	well no
coming in	said he	so that's		just don’t	the cows
for my	thing i	thanks i		she does	you look
going through	and when	you out		she hasn’t	you look
i go	believe it			she’ll be	you look
in his	environmental health			there i	and you’re
instead of	i’ll give			you would	around the
not bad	managed to			a start	He might
starting to	morning oh			asked me	he's been

there's nothing	that and			said she	i i
	the market			they’ll be	if that's
	well and			thought of	leave you
	went to			to check	now you
	when the			us to	oh sorry
				wish i	telling me
				you of	that I'm
				you that’s	think we
				about him	we might
					yes okay
					you never
					you wanted
					you're the
					you about

Writer 1

Two bi-grams used by only Writer 1 were “ah ha” and “o k”. These bi-grams occur as a result of Writer 1’s tendency to write “okay” as “o.k” and “ah ha” as “ah-ha”, rather than “aha” or other variant spellings, so are arguably not strictly bi-grams, but were identified as such based on the token settings used in the analysis. “o k”, should arguably be counted as a single word, and not be counted as a bi-gram, but since the purpose of a carrying out this n-gram analysis was to reduce some of the manual selection process, it is still included in this discussion because it was returned in the results as a bi-gram. In Writer 1’s combined character corpus, “ah ha” featured 25.9 times per 100,000 words. In the individual character corpora, it was used five times by Kenton (178.3 per 100,000 words) and only once each by Brian, Helen and Ruth. This does seem to be suggestive of Writer 1 using this bi-gram for Kenton’s dialogue. However, Kenton’s use of “ah ha” only accounts for six times out of the 36 uses in Writer 1’s all-Character corpus, and the majority of uses are by other characters.

“o k” is likewise used by multiple characters. Writer 1 uses it 177 times in the main corpus, and a total of 37 times in the ten-character corpus, spoken by 8 out of the 10 characters. Tom uses it 10 times, Helen 7 times, David and Pip 5 times each, Ruth twice, and Brian and Elizabeth only once. As with “ah ha”, some characters use “o k” more frequently, but again, it is used across characters, not as a binary feature for some characters and not others. These results suggest that “ah ha” and “o k” are features of authorial style: they are used only by Writer 1, and their usage across multiple characters suggests a degree of linguistic leakage, where Writer 1's linguistic traits are identifiable. Within Writer 1's corpus, both bi-grams seem to be used more by certain characters than others (for example, Kenton's usage of "ah ha"), suggesting a degree of conscious or instinctive manipulation as well. The writer has increased the use of "o k" and "ah ha" for certain writers, but not suppressed its usage in others.

Three bi-grams which appear 20 times or more in Writer 1's corpus are "before i" (25.93 times in 100,000 words), "before we" (19.45 in 100,000 words) and "before you" (18.01 in 100,000 words). "Before" does not appear as part of a bi-gram in any of the other five writers' results. A concordance plot was produced using the ten-character corpora for Writer 1 of "before i" which showed that it was used twice by Tom, and once each by Brian and David. The All-character corpus for Writer 1 has 36 instances of "before i", spoken by a total of 33 different characters. This even distribution across characters strongly suggests that the bi-gram is not being used to create certain characters' styles.

Another bi-gram unique to Writer 1 results is "a real". This is interesting because "real" is an adjective with countless alternatives in many contexts, and can be used with both positive connotations (for example, "Let's be bad and have a real indulge this afternoon") and negative ("“A real pain in the neck." Writer 1, 2010). "A real" is used 42 times in Writer 1's All Character corpus (30.25 in 100,000 words), as shown in Table 11.

Table 11: Frequency of "a real" per writer

Writer	Frequency of "a real"	Occurrences per 100,000 words
Writer 1	42	30.25
Writer 2	9	6.48
Writer 3	7	5.04
Writer 4	17	12.25
Writer 5	10	7.20
Writer 6	6	4.32

Writer 1 uses “a real” more frequently than the other writers; over twice as frequently as Writer 4, who is the next nearest writer. In Writer 1’s ten-character corpora, “a real” occurs seven times, distributed evenly between seven characters. These figures are too low for a statistical analysis. On a close, qualitative reading, it does seem that Writer 1 uses “a real” as an intensifier, as in these examples:

> “Kirsty was making **a real** fuss of him”
>
> “Families are **a real** pain in the neck.” (Writer 1)

In phrases with a similar meaning, other writers have used different adjectives. For example, Writer 2 uses, “That man is **such a** pain in the neck” (spoken by Brian), and Writer 3 uses the phrase “**a world of** difference” (spoken by David). A possible explanation for Writer 1’s higher use of the bi-gram “a real” may be that the other writers use “a real” in a more restricted sense, to imply something is genuine, as in this example by Writer 4: “And I had some fights with him like **a real** dad, too” (Fallon’s dialogue). In contrast, Writer 1 seems to use “a real” for a wider range of purposes.

Writer 2

Four of the bi-grams which are unique to Writer 2 are content-based, rather than bi-grams which can be used to draw inferences about authorial or character style. These are: *ice cream*, *the echo, bridge farm* and *e coli.* All four bi-grams relate to a storyline about e-coli found in the ice cream at Bridge Farm, and the family’s concern about bad publicity in the local newspaper, *The Echo.* Potentially of interest is the bi-gram “has been”. As shown in Table 12, this is used 35 times by Writer 2 (25.21 per 100,000 words), which is more than double the use of the next closest writer, Writer 4.

Table 12: Frequency of “has been” by writer

Writer	**Raw Frequency of “has been”**	**Occurrences per 100,000 words**
Writer 1	5	3.60
Writer 2	35	25.21
Writer 3	8	5.76
Writer 4	15	10.80
Writer 5	12	8.64
Writer 6	8	5.76

This could be indicative of a more formal register, as in these examples:

> There **has been** a certain … lack of consensus (Brian)
>
> Mr Pickering **has been** full of praise (Lynda)
>
> This year **has been** particularly difficult (Kenton).

In each of these examples, the writer could have used a contracted form (e.g. "There**'s been** a certain"), rather than "has been". The use of "has been" features in Brian and Lynda's corpora, both of whom have higher word-length results and higher turn-length results, and have higher social status in the fictional world. The use of "has been" (as opposed to "'s been") could be indicative of the linguistic way in which Writer 2 creates formality for these two characters. However, the figures here are too low for a meaningful statistical interpretation. Further, Writer 2 uses "has been" across multiple characters, so again, it does not seem to be strongly associated with particular characters.

Writer 3

A potentially interesting bi-gram for Writer 3 is "try and". In Writer 3's corpus, this bi-gram occurs 20 times. In the individual character corpora, it occurs only three times, used once each by David, Kenton and Pat. Throughout Writer 3's All-character corpus, "try and" is used across a range of characters. A common alternative to "try and" is "try to", and a search on all six writers' usage of "try to" shows that Writer 3 is a relatively low user of "try to" (Table 13).

Table 13 "try and" and "try to" by Writer

Writer	"Try and" raw frequency	"Try and" times per 100,000 words	"Try to" raw frequency	"Try to" times per 100,000 words
Writer 1	7	5.04	13	9.36
Writer 2	16	11.53	4	2.88
Writer 3	20	14.41	9	6.48
Writer 4	10	7.20	21	15.13
Writer 5	18	12.97	20	14.41
Writer 6	12	8.64	21	15.13

It may be that Writer 3's relatively high usage of "try and" is caused by a preference for using "try and" instead of "try to". This marks a difference from Writer 1 and Writer 4, who use "try to" twice as often as "try and". Using "try and" more frequently than "try to" is not unique to Writer 3. Writer 2 also uses "try and" more frequently than "try to", with "try and" occurring 11.53 times per 100,000 words compared to "try to" which occurs 2.88 times per 100,000 words. The threshold for inclusion was 20 occurrences, but repeating the search with no minimum frequency, "try and" occurs 18 times in Writer 5's corpus and 16 in Writer 2's corpus. Therefore, although it shows as a unique occurrence (Appendix 2c) where the minimum range was set at 20, on further investigation, the frequency was not significantly different from Writer 5 and Writer 2's usage. Whilst the overall figures are too low to draw a conclusion, the distribution of "try and" throughout numerous characters in Writer 2's corpus suggests, again, that "try and" may be a feature of the writer's authorial style, rather than a feature which is adapted, either increased or decreased, for specific characters' dialogue.

Writer 4

One of bi-grams unique to Writer 4 is "uh huh". In the 10-character corpora, "uh huh" is used once by Ruth, once by Kenton and three times by David. In all six All-character corpora, "uh huh" is used 23 times by Writer 4. In Writer 4's All-character corpus, "uh huh" was used by sixteen different characters, suggesting it is a feature of authorial style, rather than a particular character's idiolect. A weakness of "uh huh" as an individuating bi-gram is that, along with Writer 1's use of "ah ha" and "o k", the bi-gram "uh huh" is arguably a feature of spelling preference, rather than lexical choice. The scriptwriters are writing in the knowledge that the audience will hear their words, rather than read them, so it may be the case that in a different medium, for example, a novel, the same writer would vary the spelling for different characters, to convey different characterisations.

Writer 5

Writer 5 uses the bi-gram "the baby" 61 times (Table 14).

Table 14: Frequency of “the baby”

Writer	Raw Frequency of “the baby”	Occurrences per 100,000 words of “the baby”
Writer 1	7	5.04
Writer 2	9	6.48
Writer 3	17	12.25
Writer 4	13	9.36
Writer 5	61	43.94
Writer 6	7	5.04

In Writer 5’s 10-character corpus, “the baby” is used four times by Helen, three times by Pat and once by Lynda. In the fictional universe, this is unsurprising because Helen is pregnant during the period from which the data are taken, and Pat is Helen’s mother, so the two characters often discuss her pregnancy, and later in the data, another pregnancy is discussed (mostly by characters who are not in the top ten most frequently speaking corpora). Writer 5’s use of “the baby” seems potentially individuating, if “the baby” is habitually used instead of alternatives such as “baby” (without an article), or common nicknames such as “the bump”. Alternatively, it could simply be Writer 5 was given a set of storylines about Helen’s pregnancy. To explore this further, the single word, “baby”, was searched in the six All-character corpora (Table 15).

Table 15: Frequency of “baby”

Writer	Raw Frequency of “baby”	Occurrences per 100,000 words of “baby”
Writer 1	20	14.41
Writer 2	27	19.45
Writer 3	42	30.25
Writer 4	60	43.22
Writer 5	151	108.77
Writer 6	29	20.89

The results show that Writer 5 has characters using the word “baby” to a much greater extent (140 times, compared to 54 for the next closest writer (Writer 4), suggesting that this n-gram is indicative of content, rather than a potentially individuating use of “the” co-occurring with “baby”. It could be argued that the high usage of “baby” (with or without the determiner) is a possible feature of authorial style, if Writer 5 has a tendency to select topics from a domestic sphere, which would account for the higher-than-average occurrence of “the baby”. However, it is not possible to conclude with any certainty what proportion of occurrences of “baby” were mandated by storylines documents, and what proportion were Writer 5’s own inclusion of more domestic conversations, which might be indicative of authorial topic choice.

Writer 6

A bi-gram used heavily by Writer 6, and also by Writer 3 is “yeah yeah”. Only Writer 3 and Writer 6 are over the minimum threshold of 20 occurrences (as shown in Table 16).

Table 16: Frequency of “yeah yeah” by Writer

Writer	Raw Frequency of “yeah yeah”	Occurrences per 100,000 words
Writer 1	13	9.36
Writer 2	7	5.04
Writer 3	33	23.77
Writer 4	10	7.20
Writer 5	4	2.88
Writer 6	41	29.53

When this was investigated further using concordance plots, both writers seemed to use “yeah yeah” heavily for one particular character, which is suggestive of character style. Interestingly though, the clustering of “yeah yeah” was for a different character: for Writer 3 it was Tom, while for Writer 6 it was Pip (Table 17).

Table 17: Frequency of "yeah yeah" by Character for Writers 3 and 6

	Writer 3		**Writer 6**	
Character	**Raw frequency**	**Normalised per 1000 words**	**Raw frequency**	**Normalised per 1000 words**
Pip	1	0.36	5	1.78
David	1	0.36	2	0.71
Tom	4	1.43	2	0.71
Helen	1	0.36	1	0.36
Kenton	1	0.36	1	0.36
Ruth	0	0	1	0.36
Pat	1	0.36	0	0

Pip is a teenage character, which could account for the higher use of "yeah yeah", expressing the fraught family relations in the storylines at the time these scripts were written. These results could tentatively suggest that writers do use certain collocations of words with certain characters, and have their own idiosyncratic ways of linguistically styling each character, as demonstrated by Writer 3 using "yeah yeah" more frequently for Tom, whilst Writer 6 used it more frequently for Pip. However, as with other results, once the search is divided by writer and then by character, the figures are too low to make this interpretation with any certainty.

Another bi-gram which occurs frequently in Writer 6's data is "er yes" (Table 18), which Writer 6 uses nearly four times as frequently as the next nearest writer. In Writer 6's 10-character corpora, "er yes" is used once by David and once by Elizabeth. Throughout Writer 6's All-character corpus, "er yes" is used by 18 different characters, so it does not seem to be used as part of any particular character's style.

Table 18: Frequency of "er yes" by Writer

Writer	**Raw Frequency of "er yes"**	**Occurrences of "er yes" per 100,000 words**
Writer 1	6	4.32
Writer 2	1	0.72
Writer 3	2	1.44
Writer 4	7	5.04
Writer 5	0	0.00
Writer 6	25	18.01

5.2.2 Tri-grams Results

The process was repeated with tri-grams. Each writer’s All-character corpus was analysed using AntConc to identify the most frequently used word-tri-grams, in descending order, down to a minimum frequency of 20 occurrences (Appendix 2d). Following the same process as for bi-grams, the tri-grams were presented in a table (Appendix 2d) which was colour-coded to identify those tri-grams which were used by five or six of the writers (red font), three or four of the writers (orange font), by two of the writers (green font) or by only one writer (green font with green fill). The tri-grams used by only one or two of the writers are presented in Table 19.

Table 19: Tri-grams used by 1-2 Writers

Tri-grams used by only one writer are asterisked(*)
(Figures are raw frequencies as the corpora were a standardised length. Min. Freq. = 20)

Writer 1	Freq	Writer 2	Freq	Writer 3	Freq	Writer 4	Freq	Writer 5	Freq	Writer 6	Freq
one or two*	34	there you go	38	got to go	31	you know the	24	I'm glad you*	25	no I know	65
a lot more	23	some kind of*	23	you are then*	21	that sort of	22	of you to*	24	come on then*	33
that's why I	21	how did you	20			we've got a	21	would be a	23	just going to	30
for a few*	20							thank you for*	22	there you go	28
I told him	20							I think that's*	21	don't need to	25
We'll have to	20							that was a*	21	like i say	25
								a long time*	20	i just want	24
								about it I*	20	i just need*	22
										what about you*	22
										you don't need	22
										just need to	21
										do you need*	20
										do you	20

										reckon *	
										you could have*	20
										you no no*	20

Potentially individuating tri-grams were selected and, using the 10-Character corpora for each writer, concordance plots were examined. The tri-gram “no i know” was used 65 times by Writer 6, which was nearly twice as frequently as the next nearest tri-gram (Writer 2, “there you go”). However, as happened with bi-grams, each tri-gram did not feature frequently enough in the 10-Character corpora enough to allow a statistical interpretation. The results of the concordance plot are shown below (Table 20). As can be seen, “no i know” was used by five of the ten characters, most frequently for Helen, with 7 occurrences. Even with this initially higher figure of 65 occurrences, the individual frequencies by writer and then by character are too low to draw any statistical inference. The 10-character corpus only captures 19 of the 65 occurrences: the remaining 46 occurrences either fell outside of the truncated corpora, or were used by other characters in the All-character corpus. There does not seem to be strong literary interpretation as to why “no i know” might be used by Brian, David, Elizabeth, Helen and Tom, but not the other five characters.

Table 20: Writer 6: “no i know” by character

Character	**Raw frequency**	**Normalised occurrences per 100,000 words**
Brian	2	71.13
David	3	10.7
Elizabeth	4	14.27
Helen	7	24.96
Kenton	0	0
Lynda	0	0
Pat	0	0
Pip	0	0
Ruth	0	0
Tom	3	10.7

The next token analysed by character was Writer 2's "there you go", but of the 38 occurrences in the All-character corpus, only two were in the 10-character corpus, both spoken by Pat, so again, it was impossible to draw any inference from this result.

The third most frequently occurring tri-gram from Table 18 was "one or two", which appears 34 times in Writer 1's All-character corpus, notably more than for the other five writers. Broken down into occurrence by character, the phrase occurs three times for Lynda, twice for Brian, once for Helen and once for Pip. Exploring this further by searching the term in the original All-character corpus, "one or two" is used by 24 different characters, which likewise shows that the writer is not using this tri-gram for specific characters.

Writer 6 has the highest number of tri-grams used by only one or two writers, but a similar issue occurs when analysing these tri-grams in the 10-character corpora. For example, "I just want" occurs 24 times in the All-character corpus (17.29 occurrences per 100,000 words), but when analysed in the 10-character corpora, there were six hits (two for Ruth, and one each for David, Helen, Pip and Tom).

5.2.3 N-grams Conclusion

These findings for word bi-grams suggest potential features of individual authorial style, in the bi-grams and tri-grams used repeatedly (with a minimum threshold of 20) by only one or two of the six writers. Trying to draw a statistical inference by writer and then by character is problematic, because even with a moderately large initial dataset, the results are too low to explore inter-character variation by author.

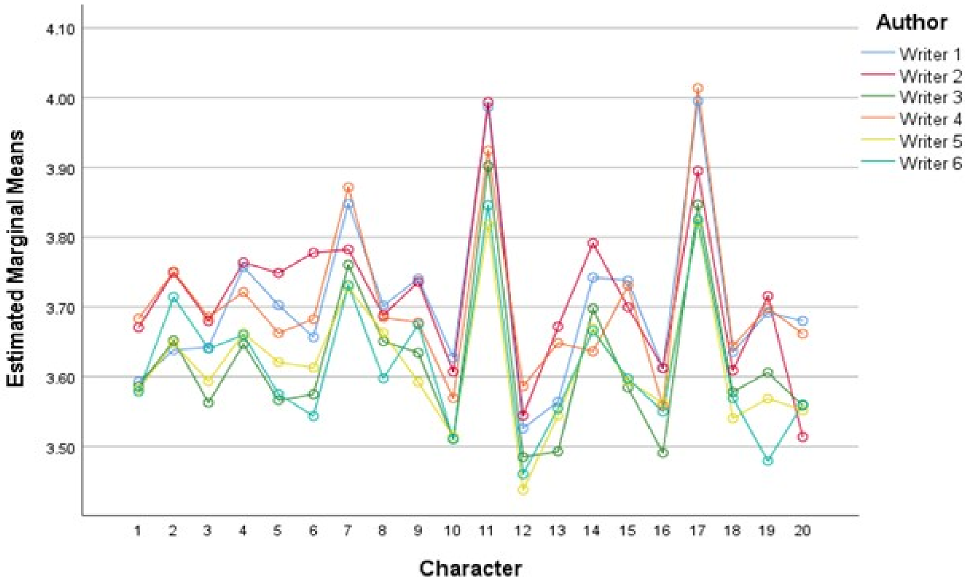
In this test, the n-gram analysis elicits some individuating n-grams for individual authors but the corpora are not big enough to demonstrate whether within individual author results there are discriminating n-grams by character. It was, however, possible to identify n-grams which were associated more with authorial style, because they were distributed evenly across characters, for example "one or two", which is used by Writer 1 across 28 different characters, suggesting that it can be associated with the authorial style of Writer 1, rather than any particular character: in the context of my research aims, this could be seen as a lack of suppression of authorial style, because the tri-gram is used regardless of which character is being written.

There were some exceptions, for example Writer 6's use of "yeah yeah" which was mostly used by the then-teenage character, Pip, and Writer 3's usage of "yeah yeah" by the character Tom. However, "yeah yeah" was also used by other characters. To use Grant and MacLeod's (2020) terminology, this is suggestive of linguistic leakage, that a writer has used a particular n-gram ("yeah yeah") in creating the voice of one character, but has not suppressed this usage when writing a different character. The results, however, are too low for this to be anything more than suggestive of linking "yeah yeah" to Writer 6's creation of the character Pip (and similarly, to Writer 3's creation of Tom).

This n-gram analysis does identify potential terms for qualitative exploration, for example the different distributions of "a real" for Writer 1. From these results, there does not seem to be any evidence that the adjective is used differently for different characters, but carrying out the n-gram test is effective in identifying tokens worthy of further exploration, for example, "a real", where a detailed exploration of its semantic and pragmatic usage could be revealing about authorial style.

5.3 Average Word Length Results

As set out int the Methodology in Chapter 4, the 20-Character corpus was analysed for Average Word Length for each character, as written by each scriptwriter. An ANOVA would have enabled a comparison of variance between the writers. Pre-conditions of an ANOVA include normally distributed data, with roughly equal variance across samples, with an even number of cases in each group (Grant et al. 2017:128) It was not possible to re-structure the data to address these issues, so an ANOVA was not carried out. However, the plot of estimated marginal means provides a useful basis for interpreting the statistics. Each line represents a scriptwriter, and along the horizontal axis are the top 20 most frequently-speaking characters. It might be expected that the characters who speak most frequently would be more consistent, since these are the more firmly established characters. For this reason, the characters along the horizontal axis are ordered in decreasing frequency of turns in the three-year corpus.



Character Key

1. David	2. Pat	3. Ruth	4. Jennifer	5. Lilian
6. Tom	7. Brian	8. Kenton	9. Elizabeth	10. Helen
11. Lynda	12. Pip	13. Tony	14. Jill	15. Susan
16. Fallon	17. Jim	18. Eddie	19. Brenda	20. Jazzer

Figure 2: Average Word Length: Plot of Estimated Marginal Means

Overall, it seems that the writers do adapt their average word length for each of the characters, and show a fair tendency to follow a broadly similar pattern of increases and decreases for the twenty characters. The similarities in the increases and decreases in average word length suggest a deliberate attempt to create varied linguistic styles for each character.

Writer 2’s characters frequently have the highest average word length out of all the writers, with a notable exception in Character 20 (Jazzer). Writer 1 and Writer 4 also tend to write dialogue with higher average word length, regardless of which character’s voice they are writing. Writers 3, 5 and 6 all tend to be the lower of the six writers, again regardless of which character they are writing.

Very roughly, the six writers tend to follow the same pattern of increases and decreases for each character, with a couple of notable exceptions – the characters Lilian and Tom have a relatively broad range between highest average word length and lowest, and the character Brenda has three writers clustered closely, two in the middle and one (Writer 6) much lower. For Brenda, the difference between highest average word length and lowest is 0.236, so there is a relatively large split in the different average word lengths of Brenda's dialogue, compared to other characters, for example David (0.105), Pat (0.112) or Ruth (0.117).

These results suggest that the writers do alter their writing style – or at least their average word length – for different characters. Certain characters have notable spikes in the average word length. They are Character 7 (Brian), Character 11 (Lynda) and Character 17 (Jim). All six writers have average word lengths higher than 3.8 letters for both Lynda and Jim, and two out of the six writers (Writer 1 and Writer 4), have average word lengths over 3.8 for Brian. For Lynda and Jim, even the lowest average word length score of the six remains higher than all other characters, with the exception of Writer 1 and Writer 4's word length for Brian.

For these three characters, the range between highest to lowest averages are: Brian (0.144), Lynda (0.176), and Jim (0.1921). Although the results for Brian, Lynda and Jim show clear spikes, reflecting three characters (a wealthy landowner, a well-to-do older lady, and a retired professor respectively), and all six writers increase the average word length when writing these characters, there is still a range between the highest and lowest averages for these writers which is similar to the range between other characters where the average word length seems a less pronounced feature of their characterisation.

Nini (2014) associates higher average word length with being male, older, educated and of a higher social class, which is reflected in the inclusion of Brian (older, wealthy, higher social class), Jim (retired university professor) and to an extent Lynda – who is portrayed as snobby and well-to-do, although not university educated (however as a female character born in 1947, this is arguably a generational circumstance and would not be the case for a younger woman from a similar social background). Lynda is portrayed as very proper, exemplified in lines such as this:

> "Ingenious? It's downright deceitful! Not only is it a travesty of what Lent is supposed to be about, it's depriving a worthy charity of their contribution!"
>
> (Writer 3, 2011)

Jim, as a retired university professor is often portrayed using markedly long phrases and grammatical constructions, as in this example:

> "And spouting some spurious mumbo-jumbo ostensibly designed to ensure the germination of mistletoe."
>
> (Writer 4, 2011)

Jim often uses formal language, and sometimes slips into Latin, as in this example:

> "Joe, I did that research for you **uberima fides**."
>
> (Writer 4, 2010, emphasis added).

This is immediately explained in Jim's next line:

> "In utmost good faith, because I thought you were genuinely interested."
>
> (Writer 4, 2010)

In this example, Jim is speaking to Joe Grundy, a character described in *The Telegraph* as a "Thomas Hardy rustic" and "the work-shy patriarch of the downtrodden Grundy family" (*Telegraph,* Obituaries 2019). In naturally occurring conversation, one might expect Jim's dialogue outside of work-related settings to show linguistic accommodation and not use higher register forms and Latinate phrases. The purpose of these linguistic quirks is to create the picture of a History scholar, rather than to portray naturally occurring conversation, which might well include linguistic accommodation. Arguably there is a greater burden of expectation of consistency for a fictional character than would be expected in real life. This is especially true in a radio drama, where the audience has no visual information, and is relying wholly on dialogue to identify characters.

Lynda and Jim, the two characters with the highest average word length are both portrayed in a somewhat caricatured manner – Jim the stuffy, retired Professor, and Lynda, the fiercely snobby village matriarch. Other characters in these analyses are similarly 'broad' in their conception, for instance Jazzer, the amorous Scottish milkman, and Eddie Grundy, described on the programme's own website as "the king of daft schemes". From these twenty characters, two of the more caricatured ones, Lynda and Jim (and to an extent, Brian), seem to manifest this exaggerated nature through a measurable textual feature, as seen in the sharp increase in their average word length, but this is not to suggest that textual variables are the only way that exaggerated characters (such as Jazzer and Eddie) might become apparent.

Johnson and Wright (2017) point out that some authors are more distinctive than others, which Grant (2020) compares to physical differences:

> Just as with person description, where identifying a 195cm tall man with a flaming red beard is easier than discriminating a more non-descript character, so too distinguishing two authors who have relatively non-descript writing styles will always be a harder task than if one of them has unusual style markers. (Grant 2020:572)

This seems to be the case here, that a number of characters are painted with broader brush-strokes (Lynda, Jim, Eddie, Jazzer), but it is only for two of these four where it is done in a way that is observable in basic stylometric tests, suggesting – very roughly – that the other two characters are distinctive through behaviour, rather than speech patterns.

Contractions might have been one possible explanation for the higher average word length. In this analysis, apostrophes were treated as word boundaries so a word such as "they're" would be treated as two words: "they" and "re". Characters who avoid contractions would be likely to have a higher average word length because of the high frequency with which the elongated forms (in this case "they are") would appear, since most are in the highly-occurring word group of function words. However, for all six writers, both Jim and Lynda frequently use contractions such as "It's" and "you're". From a qualitative exploration of this possibility, it seems that the higher average word length is created through vocabulary choice of content words rather than a lack of contractions creating longer function words, as can be seen in this example:

> "So my little recitation is to be delivered to an audience of diners as they pick their teeth and eructate between courses."
>
> (Writer 2, 2011).

Jim could use 'recital' instead of the longer 'recitation' or 'burp' instead of the ostentatiously formal 'eructate'. There are many instances in the data where both Lynda and Jim deliberately select longer, more prestigious lexical choices, when shorter, simpler synonyms would suffice. Jim's lexical choices are discussed in more detail in Chapter 7.

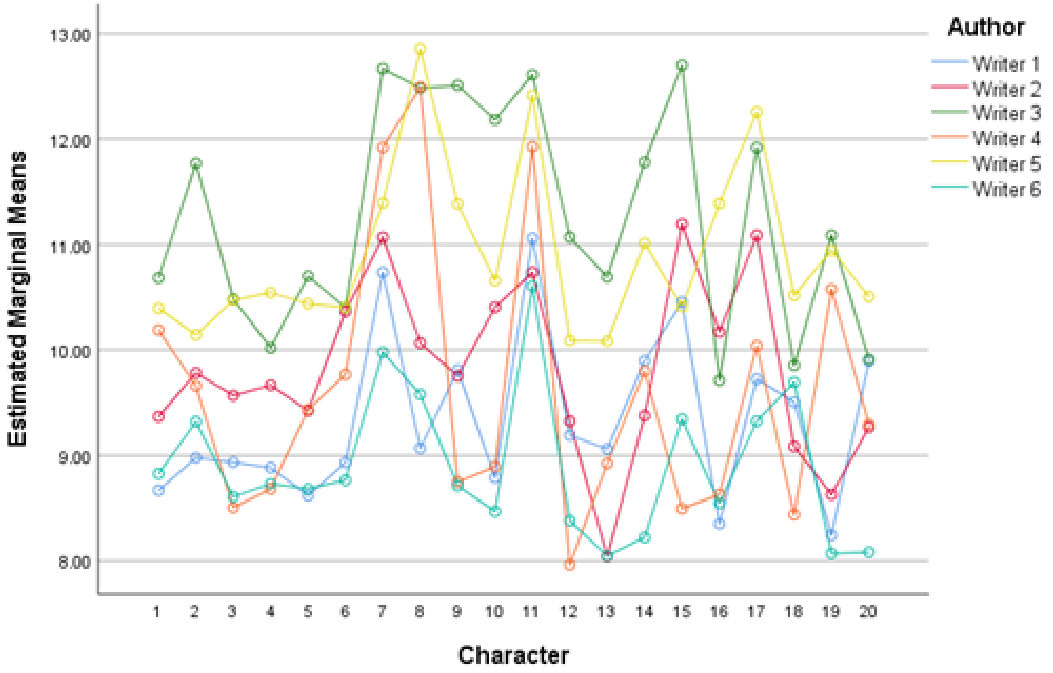
Whilst the three notably 'higher register' characters fit the profile of speakers who might be expected to have a higher average word length, there are other characters who would also fit Nini's sociolinguistic profile, but do not have notably higher average word lengths. These include characters such as Character 5 (Jennifer), who is married to Brian. She is an older character (mid to late sixties when these scripts were written). Her snobbishness and desire to always have the finest things are renowned within the village. Another character who fits this profile is Jennifer's sister Lilian, described as "the Camilla Parker Bowles of Ambridge, happiest with a stiff G&T in her hand and a highly strung hunter between her knees" (Nancy Banks-Smith, 2016).

From these results it is possible to tentatively infer that whilst some characters might be written with a higher average word length, it does not necessarily follow that all characters who fit the sociolinguistic profile will be written this way. In Jennifer's case, there are often other ways which show her high social status, for example, always referring to her "Aga" rather than her oven. Interestingly, it seems to be the referent which marks Jennifer out as being socially superior, rather than the lexical style, as in the case of Jim's ostentatiously elongated vocabulary. For others characters, it may be non-stylistic or textual features, such as Jazzer's womanising ways and Eddie's oft thwarted "get-rich-quick" schemes".

One character with a lower average word length is Pip. For the three years of the data, she is age 17-20, so it seems that she fits the profiling corollary that the younger characters will use shorter words, especially in these data, where children and adolescents do not represent the core demographic of the station's audience, and speak less frequently than older adults. For two of the writers, (Writer 5 and Writer 6), Pip's average word length is notably shorter than for their other characters. For Writer 4, Pip's average word length is low, but compared to the way Writer 4 creates other characters, it was not lower than the way the Tony (a middle-aged male farmer) and Fallon (a 20-something woman who runs a pub) are written. Their social backgrounds and educational levels are similar, but they have disparate ages and gender. The plots show that what might be a significantly lower average word length for one writer is not distinctively so for another writer (in this case Writer 4) relative to the way they write other characters, so while the average word length results show general tendencies in linguistic disguise, that the writers can and do change their use of a textual feature to write different characters, this feature alone could not successfully attribute any authors with any degree of certainty, in line with Grieve's (2007) findings and Juola's (2008) observations. This would be expected in most quantitative authorship attributions, where a number of textual measurements are analysed in combination.

5.4 Average Turn Length Results

The third test carried out was Average Turn Length. As discussed in Chapter 4, Turn Length, rather than Sentence Length, was used as the unit of analysis. The results for Average Turn Length show some similarities to results for Average Word Length, with some similarly notable 'spikes', as shown in the graph below. As for Average Word Length, the necessary conditions to run a full ANOVA were not met, so the plot of estimated marginal means was analysed instead. Again, each line represents a scriptwriter, and along the horizontal axis are the top 20 most frequently speaking characters.



Character Key

1. David 2. Pat 3. Ruth 4. Jennifer 5. Lilian
6. Tom 7. Brian 8. Kenton 9. Elizabeth 10. Helen
11. Lynda 12. Pip 13. Tony 14. Jill 15. Susan
16. Fallon 17. Jim 18. Eddie 19. Brenda 20. Jazzer

Figure 3: Average Turn Length Plot of Estimated Marginal Means

The three characters who were written with higher-than-average word lengths (Brian, Lynda and Jim) all had relatively high average turn lengths. However, many other characters also had a higher turn length. In some cases, this might be explained by the function of the character in the drama, rather than their "personality" per se. For instance, Kenton is established as the character who talks on the Public Address System (PA) at any public events in Ambridge, such as hosting quizzes at The Bull, compering village fetes, and providing commentary at sporting events. This means there are a number of scenes where he has an extended monologue, a less frequent occurrence in a soap opera than in theatre. Writer 5 has the highest average turn length for Kenton, but that writer's corpus includes a number of scenes where Kenton is performing a monologue on the PA, as in this 82-word example:

> "And it s the last chance saloon for Tom Archer. He needs a four off the last ball to beat Eddie.

> Everything to play for on the last ball of this round. And Eddie s rubbing that ball till it shines - clearly thinks he's got something special up his sleeve. Will that do it? It s got to get to the boundary, but can Will get there first and take his Dad through to the semi final? No, he can t!”
>
> (Writer 5, 2010)

Taking the first 50 turns of Kenton’s lines in the Writer 5 corpus, up to the very first occurrence of a scene where Kenton is delivering a monologue, there are 520 words, giving an average turn length of 10.4 words per turn, which would be much more similar to many of the other characters. This suggests that Writer 5 is not necessarily writing Kenton with a longer turn length, in quite the same way that the writers are increasing the turn length for Lynda and Jim. Rather it depends whether Kenton is having a duologue, or whether the storyline requires Kenton to deliver a monologue, and the presence of these longer monologues raises the mean figure for Kenton’s turn Length.

Kenton also has a high Average Turn Length in Writer 3’s corpus. This high figure seems to be partly created by a smaller number of speeches which are particularly long, for example a 136-word turn where we only hear Kenton’s side of the conversation on the phone. If the other side of the conversation had been scripted and audible, the speech would have been broken down into a number of smaller turns, thereby lowering the figure for average turn length. Whilst it is partly authorial choice to make a phone call one-sided with pauses, or to hear both speakers, it is not necessarily indicative of any character traits, or the creation of a linguistic profile. In these examples average turn length seems to be a partially situational choice, rather than a feature of characterisation through linguistic style.

Interestingly, Brenda, who divided the writers in Average Word Length, also divided the writers over Average Turn Length. Three writers write her with a relatively high turn length, and three with lower. However, it was not the same writers: Writer 4 was high for both variables, but for average word length, Writer 1 and Writer 2 were also higher, with Writers 3, 5 and 6 writing shorter words. For turn length, Writers 3, 4, and 5 gave Brenda longer turns on average. There are a number of ways to explain this. It may be that Writers 3, 4 and 5 present Brenda as a ‘chatty’ character, who speaks a lot, often multi-sentence turns, but who does not necessarily use long words. For example:

> “Five houses to try to sign up and then I think we re done. There s a nice pub up there. I ll buy you a drink.”
>
> (Writer 5, 2010)

There may also be some social and institutional reasons for differences – for example some characters are often heard in their professional capacity, e.g. Alistair the vet or Alan the vicar. In such cases we may well expect longer speeches (in the case of Alan’s sermons), or longer average word length (in the case of Alistair discussing veterinary issues) which could influence the results. In contrast, Brenda tends to be heard in social or domestic contexts; in conversation with family and friends rather than professionally. This is likely to affect the vocabulary she uses. *The Archers* is a rural, agriculturally-centred drama, so the nature of jobs held by the characters influences whether or not they might be heard on air in their professional settings: we are more likely to hear the work-related conversations of a farmer or agricultural vet than a character such as Brenda, who works in marketing. Whether or not the characters are heard in their personal or their professional surroundings could well affect features such as average turn length. Further research could focus only on comparable situations for characters, for example in the pub, or only scenes at home, or could compare register variation for a single character in different personal and professional situations. However, the dramatic nature of the data, compared to naturally-occurring conversations, means that even then there will be other influences on the text because each scene is required to progress the plotlines in some way. This will influence what the writer has to “do” with each scene, which in turn, is likely to influence the language they use.

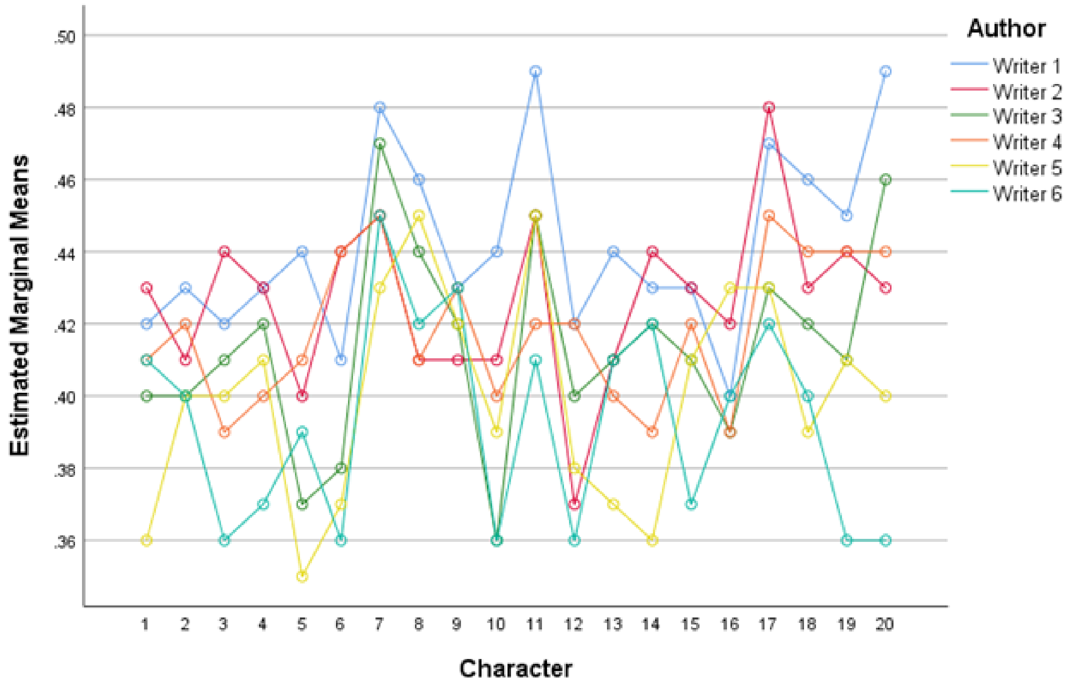
Notwithstanding these various influences, there do seem to be shared patterns of increase and decrease by the script writers for some of the characters, and in particular Lynda and Jim, suggesting that Average Turn Length is a variable which can be manipulated by writers to create different voices, although this patterning does seem to have less inter-author similarity than Average Word Length.

A possible explanation for the greater disparity in Average Turn Length compared to Average Word Length is that writers pay more attention to word choice, rather than discourse-level features such as turn-length and turn-taking, and so are more successful at mirroring each other lexically, rather than at the level of turn-length. Burton (1980) who advocates analysing dramatic dialogue at a discourse level (see 2.7.4), illustrates through her analysis of conversational structure in Pinter's *Last To Go*, how the mechanisms of humour operate at a discourse level, beyond the individual words, and that this is the level at which his distinctive "Pinteresque" humour becomes apparent. If, as Burton suggests, evidence of individual authorial style can be found at a discourse level of text analysis, this perhaps explains the greater inter-author disparity in the stylometric results for Average Turn Length.

5.5 Vocabulary Richness

5.5.1 Vocabulary Richness Results

Two tests were run for Type-Token Ratio. The first used all top 20 characters by all six writers. The smallest of these 120 corpora was 695 words, so the first 695 tokens from each corpus were used to calculate the Type-Token ratio for all the corpora, to control for turn length, following Grieve's methodology in the Vocabulary Richness test (2007:253). As discussed in the section on Turn Length, this may not be representative of all the characters' type of speech – for example the difference between hearing characters in their professional or personal contexts. As previously, the conditions for an ANOVA were not met, so the plot of marginal means was analysed.



Character Key

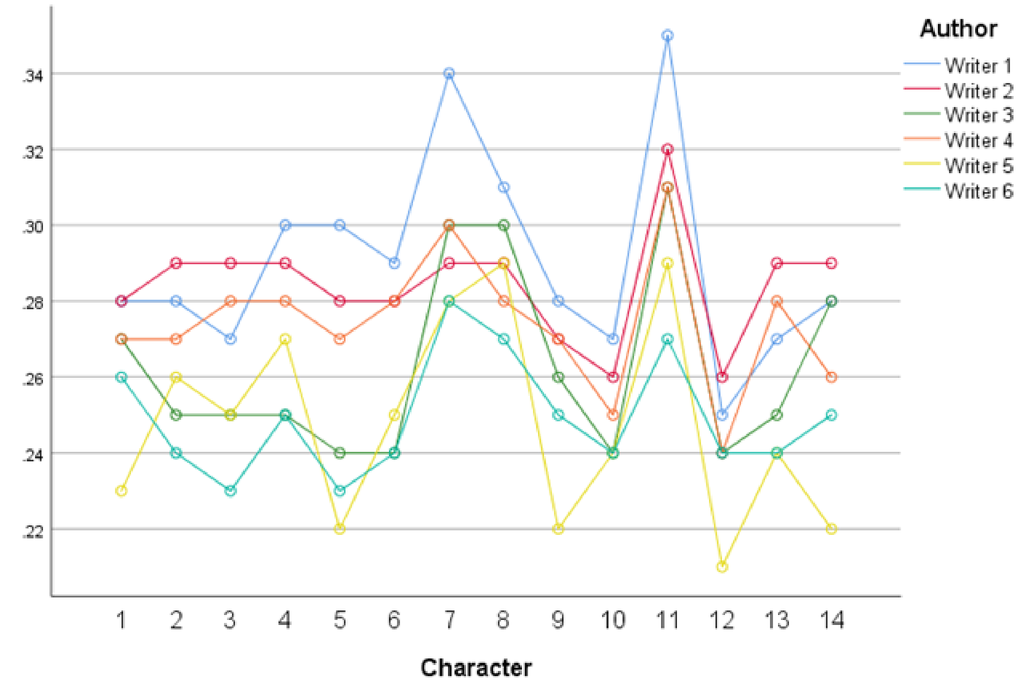
1. David
2. Pat
3. Ruth
4. Jennifer
5. Lilian
6. Tom
7. Brian
8. Kenton
9. Elizabeth
10. Helen
11. Lynda
12. Pip
13. Tony
14. Jill
15. Susan
16. Fallon
17. Jim
18. Eddie
19. Brenda
20. Jazzer

Figure 4: TTR, Plot of Estimated Marginal Means

Figure 4 shows an analysis of 20 characters, on corpora of 695 words each. As can be seen, the results are rather mixed. There are the same three peaks for Brian, Lynda and Jim, suggesting that their richer vocabulary usage indicates their higher social or educational status. As in the previous two analyses, Pip scores lower, as does Helen – who fits the same age, gender, and social class as Brenda, who does not. Brian's vocabulary richness is higher here, on average, than Lynda and Jim. One possible suggestion for this might be that, in the data, Brian is heard in a number of different settings – agricultural, corporate, domestic and social. This might increase the range of vocabulary he uses. In MacLeod and Grant's (2020) terms, he has a number of linguistic resources available for his identity construction (as farmer, parent, husband, landowner, board member), all of which increase the linguistic identity performances he engages in, which may in turn explain the increased lexical richness.

Four of the writers have a higher TTR figure for Jazzer, whereas Writer 6 is notably lower. One possible reason was that Jazzer is a Scottish character, and writers using non-standard spellings to create eye dialect could increase the variety of vocabulary. However, Writer 6 also writes Jazzer's voice using eye dialect, such as "nae" for "not", suggesting that the presence or absence of dialect cannot wholly explain Writer 6's lower lexical richness for Jazzer. For a fuller discussion of Dialect, see Chapter 6. It seems that for certain characters (e.g. Pat), the writers are quite tightly clustered together, but for other characters, they change their writing style, but not with any inter-author consistency.

A second TTR analysis was carried out using only the top 14 characters, meaning that the minimum corpus size (to which all other corpora were standardised) was 2366 words. Again, some of the results for this analysis of a much longer corpus follow a similar pattern: there are spikes for Lynda and Jim, and lower figures for Helen and Pip. These results mirror the results of Average Word Length and Average Turn Length. It is not entirely clear why the results would be lower for Helen in particular, but all the writers seem to have a similar 'dip'. Writer 5 continues to write David with a significantly lower TTR than the other five writers.



Character Key

1. David	2. Pat	3. Ruth	4. Jennifer	5. Lilian
6. Tom	7. Brian	8. Kenton	9. Elizabeth	10. Helen
11. Lynda	12. Pip	13. Tony	14. Jill	

Figure 5: TTR, Plot of Estimated Marginal Means (14 characters)

There does seem to be some intra-author consistency for this feature: Writers 3, 5 and 6 all write with slightly lower TTR figures, and Writers 1 and 2, and to a lesser extent, 4, tend to have the highest TTR throughout. These results seem to show that authors do alter their TTR, depending on which character they are writing, but that an author with a tendency towards a lower TTR figure will maintain this relatively low position regardless of which character's voice they are writing. This suggests that writers do alter the richness of their vocabulary when they create linguistic identities, but that there will be an element of linguistic leakage: those writers who tend towards a high TTR will maintain that, even when the TTR of a character is reduced.

5.5.2 Unique Vocabulary

To explore in more detail the measure of Vocabulary Richness, a corpus was compiled for each of the six writers, consisting of all twenty characters combined, for the years 2010-2012. For each of the six corpora all tokens were compiled into a table of decreasing frequency. As would be expected the most frequent words were the common function words. As the frequencies decreased, any tokens which were unique to a single writer were marked, and the unique tokens were analysed.

The first unique token (in order of decreasing frequency) is “Jaz”, which was only used by Writer 3, who used it as an abbreviation of the character name Jazzer. Other writers use “Jazz” as an abbreviation but “Jaz” seems to be consistent to Writer 3, who uses “Jazzer” 68 times, and “Jaz” 25. “Jazz” is also used 6 times (one of which is a music reference, not the character’s name.) As an identifying feature for authorship attribution, this could be useful, although it does not offer any insight into creating linguistic personae. Further, it is not anything audibly distinctive, so has no bearing on how authorial style is detectable on air.

Pragmatic Noise

The next unique token is Writer 1’s pragmatic noise item, “Pffh”, which features 14 times in Writer 1’s corpus and is unique to that writer. “Pfff” appears twice in Writer 1’s corpus and “Pffhh” also twice. Other similar examples of pragmatic noise were also unique to individual writers. These include;

Writer 2: *hmph* (4 occurrences), additionally two occurrences of ‘hmmph’.

Writer 4: *hrm* (5 occurrences)

Writer 6*: Sshh* (7 occurrences)

Writer 6: *urgh* (14 occurrences)

Pragmatic noise is an interesting part of speech because the writers are writing down phonetic versions to indicate the sound they wish the actors to make, so have the freedom to choose spellings. Writer 6 only uses the “sshh” spelling of this word, and is the only writer to use this spelling. Writer 4 is the only writer to use “hrm” and there do not seem to be equivalent uses of this word with other spellings for the other five writers. Writer 2 uses ‘hmph’ five times, and also one occurrence of ‘hmmph’. The other five writers use ‘hm’, ‘hmm’ and other versions of ‘hm’ with various added ‘m’s on the end, but Writer 2 is the only writer to add the ‘ph’ sound. No other writer uses ‘hm’ with a ‘ph’ of ‘f’ on the end (although in the ‘master’ data – which includes all speaking characters, not just the top 20), Writer 3 has one instance of ‘Humph’, but written as the standard English spelling. ‘Hmph’ seems to be unique to Writer 2, regardless of which character is speaking – the six occurrences are used by four different characters. Pragmatic noise, because of its idiosyncratic, non-standardised spellings, is an area where inter-author variation is sometimes apparent, but this is often a reflection of idiosyncratic spellings rather than vocabulary differences. Furthermore, these items do not occur frequently, so become less useful as a style marker. Pragmatic noise is explored in detail in Chapter 7.

Dialect

Jazzer is a Scottish character from Glasgow, which is reflected in his accent and dialogue. Some script writers use more dialect items (for example “wee” and “cludgie”) than others, and some use eye dialect (e.g. “cannae”, “disnae”, which leads to individualised spellings of certain words. ‘Ne’ is a word that is unique to Writer 6, and has 10 occurrences. Nine of these are a non-standard spelling of ‘not’ or ‘no’, as in, “It is ne that bad, is it? (Writer 6, 2010). There is also one occurrence of ‘ne’ in the phrase ‘je ne said quoi’, spoken by another character. Writer 4 uses ‘nae’ the equivalent to Writer 6’s ‘ne’ (there is also one occurrence of ‘nae’ by Writer 6). Writer 4 also uses (among others) didnae, disnae and dinnae; for example, “Nah, dinnae bother fannying around.” Writer 4 uses ‘didnae’ four times and ‘doesnae’ twice for Jazzer. Others write the same phrase using standard written English. Writer 5 uses isnae, wasnae and hasnae (maximum of twice each). Writer 3 and Writer 2 do not use the ‘nae’ equivalent and instead write Jazzer’s dialogue in standard written English. Jazzer’s dialect is discussed in more detail in 6.2.

Red Herrings (and crayfish)

Some vocabulary items seemed to be unique to a writer and used relatively frequently. For example, Writer 3 is the only writer to use the word 'clap' and does so six times (NB other writers use variant forms of the word, e.g. clapping). Writer 4 uses the word 'crayfish' six times. However, in such cases, there is frequently an episode-specific reason. Writer 3 has an episode where a pub talent contest is scored by a 'clap-o-meter', so the phrase gets frequent mention. Similarly, there is an episode about young boys catching crayfish and wanting to sell it to the local hotel, so again, the word is mentioned heavily in one episode and then not again. Results such as these seem initially promising but actually reveal very little about the writers' style. For other words, there seemed to be a notable distinction by one writer. For example, the word 'behoves' is used four times by Writer 4. However it is used once by Lynda, who is established as a verbose and somewhat snobby character, and the subsequent three mentions are other characters teasing her about her use of the archaic word choice.

One interesting word which is unique to Writer 3 is 'dinnertime'. In British society, the name choice of meals – dinner, supper or tea – can often be comically divisive between geographical regions and social class, as discussed in Barr's (2018) opinion column in *The Independent*. Writer 3 is the only writer to use 'dinnertime' and it is used by three different characters, Fallon, Susan and Jazzer, who are all from similar social backgrounds. However, observations such as these are not particularly illustrative about vocabulary richness as such, and it seems that the distinctiveness is more in the particular choice of word (such as dinnertime) which is revealing about the writer's sense of character. In this sense it seems more closely linked to schema theory (discussed in 2.7.2), that writers draw on the "top-down" bundles of knowledge to portray different aspects of characterisation, such as the French desserts for Jennifer Aldridge, or names of Greek and Roman gods for Jim Lloyd. More is revealed by the real-world knowledge of what the word means, than by the lexical or syntactic qualities of the word itself.

5.6 Conclusion

The word-n-gram-based test provides an explicable approach because I was able to identify bi-grams and tri-grams which were heavily associated with some writers and not others, for example “a real” for Writer 1. However, the figures were too low when analysing by writer and then by character to show strong evidence of bi-grams or tri-grams being associated with the linguistic style of any individual characters. There are suggestions in the bi-gram “yeah yeah”, which was used heavily by Writer 6, and was used more frequently in Writer 6’s Pip corpus than any other character corpus. However, it was still used by Writer 1’s other characters, tentatively suggesting some linguistic leakage by the writer.

The results for Average Word Length, Average Turn Length and Type-token ratio suggest that writers can and do alter features of their language as they change linguistic identities. However, it seems that only some of the characters are distinguishable with significantly higher or lower average word length, turn length and vocabulary richness. This is comparable to the idea of the tall man with flaming red hair, who is easily noticed in a crowd through certain individuating features (Grant, 2020:572). In this case, using certain distribution features, Lynda, Jim and to a lesser extent, Brian, are notably high, whilst two other characters, Pip and Helen are lower. With Pip’s low values, there are external reasons (mainly her age) why this would be so. As a teenager interacting, often reluctantly, with adults, her side of the conversation is shorter. Further, she is only heard in her home environment of a farm, and not, for example, in her college environment, where it might be expected that she would speak at greater length, and perhaps use longer words or have a more varied vocabulary. The results for Helen show an issue with the stylometric test used. All writers have shown the same pattern of writing Helen with a lower average word count, but this test lacked the explicability of the word-n-gram-based test, because there is nothing in the results to explain these low results for Helen, shared by all six writers.

It is extremely unlikely that any of the writers are consciously trying to increase or decrease features such as average word length, but the common spikes for certain characters suggest that this is a feature which shows inter-character variation, and is therefore susceptible to manipulation, either consciously or unconsciously. Whether or not these variation in word length in a conscious decision is an interesting question: MacLeod (2020) details the linguistic training given to undercover officers to prepare them for authorship synthesis. Writers in collaborative projects such as multi-authored dramas are not routinely given linguistic training, and nobody would suggest that scriptwriters have a statistical level knowledge of sociolinguistic features. However, part of the process of creative writing relies on observation, and as such, those features which are above conscious manipulation, may be deliberately altered to create varying characterisations. McMenamin (2002) refers to language choices as both conscious and unconscious:

> A written-language style is also defined by the individual writer's range of variation, i.e., the aggregate set of variable forms and uses of language, conditioned separately and together as a set by the conscious and unconscious choices the writer makes during the writing process (2002:26).

He further develops the discussion about what parts of writing are conscious and unconscious:

> Not enough is known about the composition to establish precisely what in writing is conscious or unconscious. The reasons for this are the difficulties associated with such studies, i.e., that every writer's level of conscious choice of forms in writing is different, and that writers demonstrate varying levels of consciousness in language production, e.g., unconscious, subconscious, semiconscious and conscious. (McMenamin, 2002:169).

Some variables have been viewed as the product of unconscious choice. Stamatatos, (2009:540) states that function words are an important discriminator of authorship partly because they are used in an unconscious manner by authors. Other choices, such as certain vocabulary choices, may be more conscious decisions. The ability of writers to manipulate average word length in broadly similar patterns suggests an element of conscious control.

In the context of my overarching research aim of exploring imitation and suppression of style, the results of the Average Word Length, Average Turn Length and Type-Token ratio tests suggest that writers are able to imitate certain characters, which is shown through the shared patterns of increases and decreases, and in particular for the characters Jim, Lynda and Brian. The marked results for these characters suggest that the writers are able to imitate features of each others' style. However, these tests do not show whether the writers are able to suppress features of their authorial style.

On a larger corpus, or on a similar-sized corpus with fewer characters, a word-n-gram-based test could make it possible to provide evidence for individuating linguistic features which the authors have not suppressed. This would be identified if there is a clear pattern of use distributed evenly among the characters. In my data, however, the raw frequencies of usage of a given n-gram divided by author and then by character were not high enough to suggest patterns of usage with any confidence.

6. Three Character Studies

6.1 Introduction

This qualitative chapter focuses on three fictional characters, and uses a different linguistic domain to analyse each. The three domains are lexical choice, dialect, and (im)politeness strategies; and the three characters are, respectively: retired History professor, Jim Lloyd; womanising Glaswegian agricultural worker Jack ("Jazzer") McCreary; and the village's self-appointed organiser-in-chief, Lynda Snell. Drawing on theory and methodology from sociolinguistics and stylistics, this study explores whether the linguistic features used to help create distinctive characters can also be used to discriminate between authors. This addresses the second of my research sub-questions, which asks whether writers are able to identify consistent intra-character features, and apply these in the dialogue they write. Firstly, I analyse Jim Lloyd's lexis, to consider which features of Jim's language choice the writers utilise, and how closely they match each other.

My second character study continues my exploration of the second of my research sub-questions. Using the feature of Jazzer's Glaswegian dialect, my analysis considers how consistently each author writes Jazzer's dialect, in terms of both intra-author consistency and inter-author consistency. Jazzer's dialogue, for all six writers, is also compared to a script from a Glasgow-based drama, *River City*, to explore which aspects of Scottish, and specifically Glaswegian, dialect the English writers draw on, and which are used less frequently, if at all. Thirdly, I examine the (im)politeness of Lynda Snell to compare the politeness strategies used by the different scriptwriters, to consider the ways in which the writers used common strategies to portray Lynda's assertive character.

Lynda, Jim and Jazzer could all be described as larger-than-life characters in the programme. The quantitative analysis (Chapter 5) showed that the scriptwriters' dialogue for Lynda and Jim had higher than average word-length, higher than average turn-length, and a higher type-token ratio, so this chapter explores ways in which the writers imitate these marked features of characterisation, and considers how closely the writers match each others' styles.

6.2 Study 1: The Lexis of Jim Lloyd

This analysis of Jim's characterisation explores the lexical features identified in Culpeper's chapter on textual clues in characterisation (2014), because it is arguably one of the most comprehensive discussions on lexis and characterisation. The relevant areas of lexical choice outlined by Culpeper are:

(i) Lexical richness

(ii) Key Word Analysis

(iii) Latinate versus Germanic words.

Each of these methods of analysis is explored and discussed. The character Jim was chosen because he provided the opportunity to describe a relatively new character. Jim first appeared in 2007, but became a regular 'resident' a couple of years later, just before the beginning of these data. It was also an opportunity to analyse further a character that all six writers had – quantitatively – written in similar ways. Jim had a higher-than-average word length and sentence length in the data, so I was interested in comparing whether Jim was always written with a high type-token ratio, or if there were fluctuations in the data as the character developed over time.

6.2.1 Lexical Richness Results

In the quantitative results in Chapter 5, the character of Jim was written with a relatively high type-token ratio by all six scriptwriters (relative to their average TTR levels for each of their top twenty characters). To analyse Jim's type-token ratio in more detail, I calculated the type-token ratio for the first 500 words of Jim's corpus, and then the compared this to the type-token ratio for the final 500 words from 2015 to explore whether the character had changed over time, from being a relatively new character, to being firmly established in the programme.

Table 21 shows the TTR results for the writers' early and late words of Jim's dialogue. The reason that the token count is higher than 500 is because 500 words were selected using Microsoft

Word, and then AntConc was used to calculate the TTR, which separates words at apostrophes (for example, “what’s” as two separate words, “what” and “s”).

Table 21: Jim Lloyd Type-token ratio, 2010 and 2015

	2010			**2015**		
	Types	Token	**TTR**	Types	Token	**TTR**
Writer 1	274	536	**0.51**	297	534	**0.56**
Writer 2	260	527	**0.49**	275	524	**0.52**
Writer 3	250	533	**0.47**	249	534	**0.47**
Writer 4	262	533	**0.49**	271	531	**0.51**
Writer 5	253	535	**0.47**	262	529	**0.50**
Writer 6	229	540	**0.42**	260	527	**0.49**

As can be seen from these results, Jim’s lexical density remains fairly consistent between 2010 and 2015. Writer 1 is very slightly higher in the 2015 corpus than 2010, and Writer 6 is marginally lower in 2010 than 2015, but broadly the lexical density is similar, and would not reliably distinguish between any of the scriptwriters. Jim’s relatively high TTR compared to other characters suggests that this is a feature which has been manipulated by the writers, and shows a distinctive character style, where all six writers have produced similar results.

The findings here, suggest that lexical density does not discriminate between the way these six writers voice Jim. Furthermore, is not necessarily a reliable authorship marker if writers are able to manipulate it to create the voice of a character, following McMenamin’s (2020) preference for ‘unconsciously’ used style markers. This is not to suggest that any of the writers were aware of, or deliberately focused on increasing Jim’s type-token ratio when they wrote his lines, but it is possible that there was at some level of consciousness, an increased set of vocabulary choices for Jim.

6.2.2 Key Words

This next section analyses the keywords for Jim. Firstly, the results for the Jim corpus, as written by all six writers, are shown in Table 22. Then the keyword results for each writer’s separate Jim corpus are discussed in turn.

Table 22: Key Word List for Jim (6 writers combined)

AntConc KeyWord calculation: “Jim” combined corpus by all six writers			
Rank	**Frequency**	**Keyness**	**Keyword**
1	65	+ 274.31	christine
2	87	+ 187.49	joe
3	53	+147.05	indeed
4	103	+141.72	ah
5	70	+ 116.89	jazzer
6	93	+ 115.18	lynda
7	792	+ 101.4	of
8	21	+ 90.19	orchard
9	207	+ 81.39	very
10	28	+ 77.91	cider

As Culpeper found, some keywords were more closely influenced by context than by characterisation. The top result, “Christine”, is the name of Jim’s companion, who does not appear very frequently in the show, and when she does appear, is often in scenes with Jim, so it is more likely that Jim, rather than any other character, would use her name. This level of storylining is often made by the production team, rather than an individual scriptwriter, so is not indicative of authorial decisions. Further, since the show is a radio programme, there is an increased tendency for characters to mention each other by name (especially those characters who appear less frequently) in order to remind the audience who is speaking, so this result is explained by the genre of radio drama, rather than authorial choice. The second highest result (Joe) is also a character name. Joe Grundy was a prominent character in the show for many years, so it is perhaps surprising that his name is featured as one of Jim’s keywords, when he appeared in numerous scenes with many different characters. However, this is likely because Joe would often feature in scenes with members of his family, and so would be referred to as Dad and Granddad by many of the characters with whom he shared a scene, thereby making ‘Joe’ a less frequent term of address for other characters to use, which makes it more likely to feature as a keyword for Jim. “Orchard” and “cider” are also explained by a storyline in which Jim was helping to organise a community orchard, which was linked to the Grundys’ “cider club”, so the higher-than-average use of these two words is indicative of “a particular context”, as Culpeper discussed, rather than characterisation.

"Indeed" and "Ah" are interesting results, because they are not linked to any particular context, but are used frequently by Jim as terms of agreement and / or acknowledgement. This seems indicative of his characterisation: that he regularly has conversations which involve a polite exchange of views, resulting in acknowledgements such as "ah" and "indeed," both of which seem quite reserved as a response. The pragmatics study (Chapter 7) found that there was inter-author variation in the way that writers used pragmatic noise items such as *oh*. The presence of "ah" (also a pragmatic noise item) on the keyword list for Jim's lexis (combined corpus) could be skewed by the presence of certain writers who use pragmatic noise items such as "ah" very frequently, but this would not explain why the token shows up as being more frequent for Jim than for other characters.

"Indeed" and "ah" could be indicative of a character who engages in rational discussions and exchanges of view. "Ah" is often used as a triumphant flourish, when Jim outflanks his conversational opponent, as in this example:

OLIVER So Caroline and a number of others tell me. Rest assured – it won't be me having sleepless nights about bread, Jim.

JIM (SMUGLY) **Ah.** But do you have the wisdom of the ancients on your side?

(Writer 1, 2012)

Culpeper and Kytö state that most frequently, "ah" can be divided into "three broad functional groups, namely, the use of AH to express: (a) emotional distress, (b) surprise, or (c) correction. Broadly speaking, these correspond to the following functions: emotive expressive, cognitive expressive and conative" (2010:225). One explanation for "ah" being a keyword might be that Jim is prone to correcting others, and often uses "ah" to do so.

"Of" also features as a keyword. The presence of "of" in this keyword list could be partly explained by Jim's tendency to verbosity, as illustrated by these three separate lines of dialogue:

> "I'm in something **of** a quandary"
>
> "The clearing **of** winter quarters. A timeless ritual."
>
> "No clue to distinguish the garb **of** a large estate-owner"
>
> (Writer 1, 2010)

Again, the high use of "of" could be a sign of Jim's pompous style, that in each case the line could have been written in fewer words, without using "of".

Keyword list by Individual Writer

The keyword lists from individual writer's corpora are now discussed. For Writer 1, "ah" was a highly ranked keyword. Only Writer 1 had "ah" has a top ten keyword, although it did rank at number 11 for Writer 2. It occurs 43 times in the data, with a keyness score of +99.99 which suggests quite strongly that it is a lexical choice of Writer 1 for Jim, to a much greater extent than any of the other writers. This is the same writer who has a notably higher use of *oh* in the Pragmatics study (Chapter 7), suggesting an intra-author consistency at the pragmatic level of analysis.

As with the combined corpora keyword analysis, some of the keywords which appear can be explained by context. For example, Greenacres is the name of the house Jim is buying, so is often mentioned by him when discussing updates on his house move, and is more revealing of topic than character, or indeed authorship, in line with Wright's (2014) argument that content words on their own can be more revealing of topic than authorship, and in this case, characterisation.

Table 23: Writer 1 Keyword Analysis

Rank	Freq	Keyness	Keyword
1	43	+ 99.99	ah
2	14	+ 59.95	christine
3	19	+ 57.37	hmmm
4	23	+ 51.67	jazzer
5	425	+ 48.37	the
6	60	+ 37.03	very
7	5	+ 36.68	greenacres
8	35	+ 36.65	quite
9	8	+ 33.6	nathan
10	4	+ 33.5	crossword

Table 24: Writer 2 Keyword Analysis

Rank	Freq	Keyness	Keyword
1	4	+ 33.75	Anaerobic
2	139	+ 29.32	the
3	4	+ 29.3	digester
4	2	+27.29	coerced
5	2	+27.29	egregious
6	2	+27.29	forging
7	7	+ 25.99	local
8	5	+ 23.82	christine
9	2	+ 23.47	soccer
10	2	+ 21.74	recitation

In Writer 2's keyword results, the highest score is 'anaerobic". Again, this does not reveal anything about character or authorship, but is simply capturing technical term used in a storyline where Jim raises concerns about large-scale agriculture. The keyword list for Writer 2 seems to show Jim's tendency to use formal Latinate words in everyday conversation, for example, "egregious", "recitation" and "coerced". It is interesting that soccer is one of Jim's keywords, because the informality of the slang word "soccer" seems at odds with Jim's portrayal as a formal, sometimes quite stuffy, character. However, both "coerced" and "soccer" only appear twice each, and for both tokens, this is when Jim is discussing the crossword with Christine:

```
CHRISTINE How many letters?

JIM       Seven.
```

CHRISTINE Anagram of **soccer**?

JIM That's only six. Oh hang on… **Coerced.**

(Writer 2, 2011)

Similarly, "forging" appears as a key word when it appears only twice. It appears in the same sentence:

> "it is heartening to discover a young man creating his own..." no, no "forging! ... forging his own opportunities..." If I don't say it the wretched editor will."
>
> (Writer 2, 2013)

"Forging" only occurs twice because Jim repeats himself, and the line is written to convey the impression of somebody reading aloud their own work as they draft a written piece. Again, its occurrence does not seem to suggest characterisation or authorship because it is an immediate repetition, rather than a chosen word which might be associated with Jim. In fact, it is Jim's second choice of word, after he initially uses "creating". The use of "no, no" and the repetition of "forging" give the impression he has only just thought of this word, and perhaps would not ordinarily use it, so its appearance as a keyword for Jim is a little misleading. The keyword results for the remaining four writers showed similar patterns.

Table 25: Writer 3 Keyword Analysis

Rank	Freq	Keyness	Keyword
1	24	+ 127.2	christine
2	23	+ 57.36	mike
3	22	+ 48.41	jazzer
4	21	+ 47.7	joe
5	28	+ 46.35	lynda
6	5	+ 39.62	literary
7	13	+ 38.45	indeed
8	4	+ 37.07	jocelyn
9	3	+ 33.52	horticultural
10	102	+ 32.58	yes

Table 26: Writer 4 Keyword Analysis

Rank	Freq	Keyness	Keyword
1	13	+ 41.91	robert
2	3	+ 37.02	wagtail
3	10	+ 35.58	indeed
4	8	+ 34.51	daniel
5	3	+ 30.29	political
6	12	+ 27.43	may
7	2	+ 24.68	farrago
8	2	+ 24.68	ornithological
9	2	+ 24.68	ornithology
10	2	+ 24.68	yellowlegs

Table 27: Writer 5 Keyword Analysis

Rank	Freq	Keyness	Keyword
1	11	+ 67.67	orchard
2	8	+ 60.1	romans
3	17	+ 34.63	joe
4	7	+ 33.17	copy
5	4	+ 32.94	sacred
6	9	+ 32.27	cider
7	183	+ 31.22	of
8	11	+ 29.97	read
9	4	+ 29.28	varieties
10	87	+ 28.12	they

Table 28: Writer 6 Keyword Analysis

Rank	Freq	Keyness	Keyword
1	25	+ 98.05	joe
2	5	+ 63.19	che
3	7	+ 51.88	onions
4	9	+ 44.13	christine
5	10	+ 38.39	indeed
6	6	+ 38.1	orchard
7	3	+ 37.91	fletch
8	4	+ 37.11	riley
9	8	+ 36.99	cider
10	4	+ 27.67	percent

For Writers 3, 4, 5 and 6, keyword lists were also dominated by character names, in particular the names of characters who frequently appeared in scenes with Jim. Other keywords were indicative of topic rather than characterisation or authorship. For example, Writer 4 has a storyline where Jim has a birdwatching rivalry with another character, Robert, which explains the presence of the keywords *Robert, wagtail, ornithological, ornithology* and *yellowlegs.* Many of the terms which appear are simply because of a given storyline which the writer is covering. In some cases, this leads to a name of a minor passing character (e.g. 'Fletch' in Writer 6's corpus), which has appeared as a keyword, simply because this character only appeared in one episode, so was featured by that one writer and no others.

Perhaps with a larger corpus for each writer, there might be more keywords which are indicative of characterisation, but in these corpora, those words, such as "coerced", which seem potentially indicative of characterisation, are often explained by the circumstance of a particular scene or storyline (in this instance, that Jim was reading aloud a crossword clue). A few lexical items, such as "ah" and "indeed" which are less context-bound appear often enough to suggest an association with the way Jim is portrayed, but this use did not discriminate between authors.

6.2.3 Latin-derived Words

The next test analysed the first 500 and last 500 words in each writer's Jim corpus, marking out Latinate words. The results are shown in Table 29, and Figure 6, below.

Table 29: Number of Latinate words used by Jim

	2010	2015
Writer 1	65	19
Writer 2	27	20
Writer 3	21	24
Writer 4	29	36
Writer 5	17	23
Writer 6	17	23

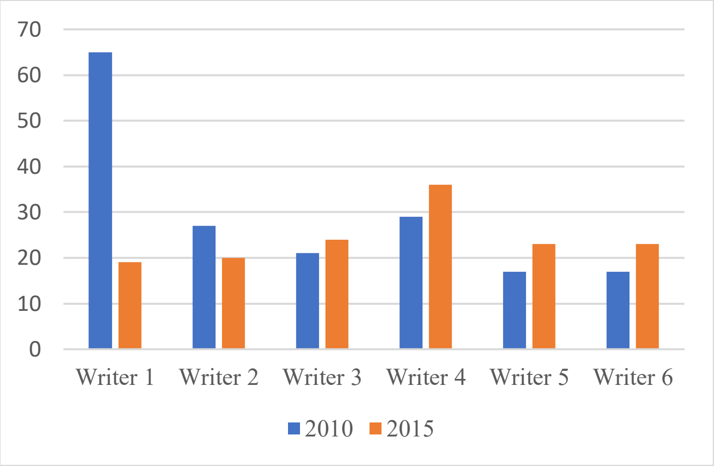


Figure 6: Jim's Latinate words 2010 compared to 2015

One assumption is that the earlier lexis would have a higher use of Latinate words as a newer, more two-dimensional character, and that this would reduce as the character became more rounded, less reliant on the linguistic signalling of being the educated professor. However, for five of the six writers this was not the case: there was only Writer 1 who showed a notable decrease in the use of Latinate words.

Writer 1 uses Latinate words over twice as frequently as the next closest writer. However, this is partly explained by the plot: in 2010 Jim is throwing a Latin-themed party, so is discussing foods of ancient Rome such as "Garum sauce" and figures such as Catullus, which increases the frequency of Latinate words for Writer 1. In some cases, within Writer 1's 2010 corpus, the Latin form is clearly a prestigious form, for example, "to help you through the **labyrinth** of possibilities," where "labyrinth" could more simply have been "maze". It would be possible to phrase this more simply, such as "to help you choose." In other cases, the Latinate form does not seem particularly prestigious, as in, "dig out these old **lecture notes**" where the meaning of the words conveys education and formality, but it is hard to think of a simpler, Germanic, less formal phrasing for "notes". Dividing the Latinate words into their grammatical classes, these results suggest the possibility that Latinate verbs, adverbs and adjectives seem to be more revealing of a wish to convey formality compared to nouns, where the writer has a more restricted choice (if any) of words to use, such as "notes" or "mobile" which do not convey any sense of being a prestige choice of word.

Writer 2 uses words for Jim which are markedly formal, as in these examples:

> "They pick their teeth and **eructate**"
>
> (Writer 2, 2010)
>
> "The lifestyle of your average **bovine**"
>
> (Writer 2, 2010)

In both examples, simpler, less formal words could easily have been selected. Whilst *bovine* might be used in agricultural settings in a professional, rather than pretentious manner (for example, discussing Bovine Tuberculosis), this particular occurrence is not a conversation about farming, so was inferred as being used to communicate Jim's somewhat pretentious style._In some cases, the content of the conversation heavily influenced the number of Latinate words. For example, in Writer 3 (2010)'s corpus, Jim is buying a house, so the Latinate words include terms associated with house buying, such as 'completion' and 'agent', which is to be expected because the legal language often has etymological roots in Latin.

Whilst it was expected that Jim's high use of Latinate words might soften over time, this did not turn out to be the case. However, the presence of some slang words did appear in the later corpus, which suggests that Jim's character became more fully fleshed, moving away from a stereotype. Culpeper writes that, "Characters that are able to switch from one type of lexis to another are liable to be perceived as rounder characters" (2014:187). This happens to an extent with Jim, that in the later data (Writer 3, 2015) he talks about having a "fry-up", and describes someone as being "quite matey" (Writer 5, 2015), so arguably it is the increase in code-switching between Latinate words and more informal language, which suggests a more rounded character, rather than a decrease in Latinate words.

It is highly subjective to decide whether a character's switch between formal and informal language is a failure of identity disguise (inconsistent), or if it is a successful creation of a complex, rounded character. Lexical choices such as "matey" are not ones that would be associated with Jim, but according to Culpeper's views on code switching, this creates a rounder character. However, these results still do not discriminate between the different scriptwriters. This particular method, of comparing Latinate to Germanic words, captures some of the inter-character variation, but does not discriminate between authors.

The same analysis was then carried out on the first 500 words in the data by David Archer, a college-educated farm owner, and then Eddie Grundy, an uneducated character. Based on their educational levels, Jim would be expected to use the highest number of Latinate words, followed by David, then Eddie the fewest. The results are shown in Table 30 / Figure 7 below.

Table 30: Latinate word totals for Jim, David and Eddie (out of 500)

2010	Jim	David	Eddie
Writer 1	66	19	14
Writer 2	28	10	10
Writer 3	21	11	2
Writer 4	29	17	5
Writer 5	17	9	10
Writer 6	16	7	6

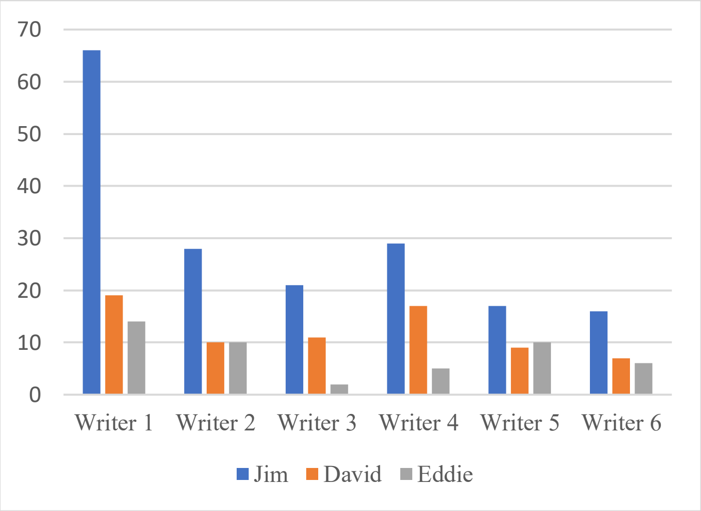


Figure 7: Latinate words used by Jim, David and Eddie

As can be seen from the chart, Jim uses an observably higher number of Latinate words. For Writer 1 and Writer 2, Jim uses Latinate words around three times as frequently as David; for Writers

3, 4, 5 and 6, this is around twice as often. Interestingly, there is not the same difference between David and Eddie that might be expected. Nearly all of the Latinate words used by David are fairly ordinary words, such as “professional”, “obviously” and “tactful” (Writer 1), and “quantities”, “traditional” and “desperate” (Writer 2). Similarly, when Eddie uses Latinate words, they do not convey formality: for example, “collect” (Writer 3), and Writer 4’s, “it’s a different way of paying my **respects**, ain’t it?” The use of “ain’t” in the tag question suggests a marked informality in that turn. It is also possible that the loosely bound idiomatic phrase “paying my respects” means that the individual words within the phrase are less deliberately chosen to create characterisation: the Latinate word appears as part of a common idiom.

These results show observable differences in the way that the writers create Jim, compared to other characters. The results also show that while cumulatively, a higher number of Latinate words suggests a higher level of formality and education, this can be influenced by context. Culpeper argues:

> We have to move beyond the simplistic 'X linguistic feature = Y personality feature' equation which has bedevilled more traditional language attitudes research. What a particular form means in one context may differ from what it means in another. (2014:166)

In this corpus, the number of Latinate words was sometimes influenced by context, such as a conversation about house-buying (“completion”) or a council meeting (“quorum”). Also, a qualitative assessment of the Latinate words used shows that many of them, for example those predominantly used by Eddie, are not themselves especially prestigious (“exactly”, “provide”).

From these observations it can be seen that there is a marked difference between the number of Latinate words used by Jim compared to Eddie, similar to the contrast between Lady Capulet and the Nurse. However, while there is a contrast between the two extremes, it interesting to see that David, who might be expected to sit midway between these characters, is only in this position for Writers 1, 3 and 4. The Latinate words he uses are also much more ‘ordinary’ words such as ‘decent’, ‘sign’, ‘afford’, rather than Jim’s more notably formal choices such as “frivolous”, “arduous” and “endeavour”. It seems that measuring the proportion of Latinate vocabulary is a useful, but fairly blunt, tool for capturing levels of formality in characterisation. However, it does not distinguish between words of Latin origin which are fairly ordinary words such as “notes” or “collect”, and those which signal a more educated character, such as “ostracisation” (Writer 5), or even “fascinating” (Writer 5). Measuring the number of Latinate words in a given length of text can provide information about characterisation, but did not show observable differences between the different writers. In the one case where it did (Writer 1, 2010), the unusually high levels of Latinate words could be explained by the context, because Jim was planning a party themed on ancient Rome, and so was discussing customs, costumes and people from that period, which significantly increased the number of Latinate vocabulary items.

6.2.4 Conclusion

Vocabulary richness, key word analysis and an analysis of Latinate words were all revealing about Jim’s characterisation, and the results suggest that writers are able to adapt linguistic traits to imitate each other. Chapter 5 showed the relatively high type-token ratio used for Jim, and the analysis in 6.2.1 showed that this was consistent across all six writers and between 2010 and 2015. This suggests the writers are able to match quite successfully at the level of lexical choice. The tests are less able to discriminate between authors, or interrogate which features of authorial style remain. Identifying Latinate words seems too blunt a tool for measuring formality, because it did not distinguish between common words from Latin such as “mobile” and more unusual ones such as “eructate”. However, using this method did create a framework for selecting lexical items for a qualitative comparison, which was able to illustrate how Jim’s vocabulary was more pretentious than Eddie’s. The method was more successful at identifying the outlying characters, Eddie and Jim, rather than the ‘middling’ character, David.

An area which Culpeper discusses that has not been explored here is syntax. This is potentially an interesting area for authorship analysis. Culpeper writes:

> Whilst research on real-life talk has not established a clear relationship between syntactic complexity and cognitive organisation, fictional texts, whether in dialogue or monologue, have exploited what appears to be a schematic relationship between syntax and cognitive organisation, such that the more simple the syntax the more simple-minded the character, and vice versa. (2014:203)

A further area for research could be an exploration of syntactic complexity, investigating the extent to which writers are able to make their syntax more or less complex, depending on which character is being written. Assuming that professional writers will be at the more adept end of a scale of syntactic complexity, it would be interesting to see how successfully and consistently writers are able to simplify grammatical structures to convey the "more simple-minded" character Culpeper describes.

6.3 Study 2: The Dialect of Jack 'Jazzer' McCreary

6.3.1 Introduction

This second character study explores how the writers use dialect, also exploring the second of my sub-questions exploring imitation of linguistic features. In *The Archers,* there are a number of characters who are not from Borsetshire, the fictional region where the drama is set. Scriptwriters can be asked to write dialogue for any of these characters, so it is interesting to see how they write in dialects different from their own. Many of the characters who have a 'nonstandard' accent in the drama are pinned as originating from the fictional village of Ambridge which is in the equally fictional county of Borsetshire, described as being somewhere near the region of Worcestershire. Since this is not a real place, the Borsetshire accent does not have a real-life existence: whilst it is possible to analyse the consistency with which the writers mirror each other, it does not allow an analysis of how successfully they portray the Borsetshire accent, since there is no original accent for them to imitate. Carrying out a study of how non-Scottish writers create the linguistic persona of a Scottish character is a way of exploring how authors try to write "otherness".

6.3.2 Dialect Results

First, the dialect items in each writer's corpus (up to the 1938 words) were counted, and coded as grammatical, lexical or eye-dialect (pronunciation). These simple quantitative results show that the writers approach writing dialect in significantly different ways. Writer 4 has by far the highest number of dialect features in the writing, at almost double the usage of the next closest writer (Writer 3), and over 10 times as frequently as the lowest writer (Writer 6).

Table 31: Dialect items per 1000 words

	Writer 1	**Writer 2**	**Writer 3**	**Writer 4**	**Writer 5**	**Writer 6**	**River City**
Dialect items per 1000 words	**17.03**	**13.42**	**35.60**	**61.92**	**11.87**	**5.68**	**5.68**

6.3.3 Dialect Results by Feature

The results are now compared in the three linguistic categories of grammar, lexis and pronunciation (Table 32). Interestingly, Writer 6, whose dialect usage was notably lower, had exactly the same results as for the extract of the *River City* script, which is produced in Glasgow. The reason that Writer 4 is so much higher than Writers 1 and 3 is the relatively heavy use of "eye dialect" which serves as a pronunciation guide, and because of the frequent use of enclitic -nae/ -na endings. All six writers use a similarly small number of Scots / Scottish grammatical features. Aside from Writer 4, the remaining five writers also have a similarly low number of pronunciation features, ranging between 0 and 3. Looking only at the results for lexis, Writer 3 and Writer 4 are notably higher than the other writers, whilst Writer 6 is notably lower. These results are now discussed in more detail.

Table 32: Number of Dialect Features in standardised script length

	Writer 1	**Writer 2**	**Writer 3**	**Writer 4**	**Writer 5**	**Writer 6**	**River City**
Grammar	3	4	3	3	3	1	1
Lexis	30	20	63	72	18	10	10
Pronunciation	0	2	3	41	2	0	0
Total	**33**	**26**	**69**	**120**	**23**	**11**	**11**
Dialect items per 1000 words	**17.03**	**13.42**	**35.60**	**61.92**	**11.87**	**5.68**	**5.68**

Pronunciation

Writer 4's decision to use "eye dialect" as a pronunciation guide is distinctive from the other five writers. Examples include:

> At this time **o'** the morning, who's counting?
>
> We're **no'** going to get loaded otherwise.
>
> Out you go **wi' ye.**
>
> Pint **o'** lager, please.
>
> (Writer 4)

Writer 4 consistently used pronunciation in the remainder of the corpus, which spanned a seven-year period. "With" was also used. There were 11 instances of the abbreviated "wi'" in the seven-year corpus (6693 words) compared to 34 instances of "with" written in the standard format, so although Writer 4 has a relatively high usage of "wi'" compared to the other writers, it is not used consistently, and is not the most frequent way that "with" appears in the text. Whilst Writer 4's relatively high use of eye-dialect might set them apart from the other writers, this is a very specific set of circumstances, and might not necessarily be applicable to the same writers in other literary texts, or even in other drama scripts: in these data, all writers know that they are writing dialogue for a Glaswegian-born actor, who can add in the dialect during recording, as in this example (by Writer 2), where the studio changes are noted in bold:

> No, I **wouldnae** ~~wouldn't~~ care for it. On your **ain** ~~own~~ all day, no-one to talk to. That way madness lies. (Writer 2, 2017)

This process may have influenced the decision of the other five writers to avoid eye-dialect. In a novel, where there is no performance process and the words are read directly by the intended audience, the same five writers who choose to use very little eye dialect here might use it much more heavily. If this were a script for an as yet uncast actor (and therefore potentially not Glaswegian), the writers might make a different decision, so it is not necessarily a marker which would be consistent across media.

In some cases, the apostrophe is inaccurate, a so-called "apologetic apostrophe" (Costa 2017:54) frequently, as in these examples:

“You're **no'** just a crate jockey.”

“I’m **no’** sayin’”

“Is it **no’**?”.

This apostrophe is not necessary because the word is not eye-dialect for a glottal stop. Instead, the character is using the Scots words “no”, not an abbreviation of “not”. As Costa writes:

> Forms of written Scots are loosely united by a set of more or less accepted rules, often based on the 1947 document, such as the rejection of the “apologetic apostrophe”—the use of an apostrophe where English has a consonant, said to construct Scots as a form of defective English. (2017:54)

This could be a mishearing by the writer, punctuating the word as if there is a glottal stop, where the Scots words does not require one. This is an indication that the writer is not Scottish, and is hearing the word as the closest English English equivalent. However this only likely to be a useful way of distinguishing authorship in cases of literary dialect. It is likely that in other forms of informal written conversation, such as on social media, that the same writer would simply leave off the apostrophe.

Lexis

Writer 3 and Writer 4 used a higher number of lexical features of Scots / Scottish Standard English dialect. One frequently-used token is “aye”. For Writer 1, “aye” accounts for 43.3% of the Scottish lexis, which is almost identical to Writer 3 (42.4%). Writer 6, who uses relatively few dialect items predominantly uses “aye” as a dialect item, where it accounts for 69.2% of the writer’s Scots / Scottish Standard English lexical choices. Compared to the *River City* script, where “aye” only appears 3 times in the same length of data, it seems that Writer 1, Writer 3, Writer 4, and Writer 6 all use “aye” for Jazzer to a much greater extent that is used in *River City*, and also higher than Writer 2 and Writer 5. Interestingly, Writer 1, who had a very high rate of *oh* usage in the Pragmatics chapter, has an equally high rate of “Aye” in this analysis. However, Writer 6, who was notably lower in his *oh* usage, had a high rate of using “Aye”. Whilst it could be seen that there is a pragmatic consistency for Writer 1 between their creation of different characters, this same pragmatic consistency was not found with Writer 6.

The enclitic ending -nae is used by the writers (hasnae and didnae, for example). Unsurprisingly, Writer 4's usage is the highest. It is notable that Writer 4 switches between an enclitic *-nae* ending and a *-na* ending. As stated in the Education Scotland guide to Scots, *-nae* is more common in North Eastern Scotland, and *-na* further south, so this intra-author variation seems telling. It is possible that a non-Scottish writer has blurred the different pronunciations, and this inconsistency could mark a vagueness in their recognition and recreation of the accent, rather than a specific local dialect, as discussed in Wells (1982).

As stated in the Methodology section, the Scots "no" and the enclitic *-nae*/*-na* endings were counted as lexical choice, but, as grammatical function words, they are arguably more informative about sentence structure, than as content words. Looking at the lexical results, and discounting all the "non-content" words (hasnae, didnae etc), as well as the pragmatic noise items / response words (Aye, eh, och) left a surprisingly small group of remaining lexical items. The remaining words from the six writers were predominantly: wee, lassies, and 'wains'. As an aside, "wains" is a misspelling of "weans". The consistent misspelling, in all but one instance (by Writer 6) might suggest a mistake that is being sub-edited in during the production process, although the fact that multiple writers across multiple scripts (which would not all be edited by the same person) weakens this suggestion. The content words which are examples of Scottish dialect do suggest a certain romanticised view of Glaswegian – arguably these words are quite nostalgically and famously Scots words, similar to the use of "braw" in Humza Yousaf's political tweet. It could be argued that the writers are indexing a certain image of Scotland and whether consciously or subconsciously, are invoking a certain type of Scottishness. There is a wealth of dialect words which are never used.

Sometimes the dialect words seem to be used slightly inaccurately. For example, the word "cludgie", meaning toilet, is used twice (once by Writer 1 and once by Writer 3). In the first case, the writer has Jazzer saying that he would "rather stick my head down a cludgie", and in the second case, Writer 3 has Jazzer suggesting to his friend that they make a sharp exit from a pub while their dates are "in the cludgie." While both writers have correctly used "cludgie" to mean "toilet", the context seems slightly off. Looking at the occurrences of "cludgie" in the Scottish Corpus of Text and Speech (SCOTS), "cludgie" collocates more frequently with "the", and only appears as "a cludgie" twice: firstly when there is a specific reference to getting a new cludgie in the close, and once in a poem. It seems to have quite a specific meaning, as being a domestic toilet, quite often a shared toilet in tenements. In the 19 occurrences in SCOTS, it often appears in a nostalgic 'memory' discourse, rather than current usage, where "lavvy", "toilet" and "bog" are used. In SCOTS, "cludgie" is not used to refer to public toilets. Although this does not mean that it could never refer to a public toilet, it seems to be more closely associated with a domestic setting. It is also an interesting dialect choice for a Glaswegian speaking to someone English. "Cludgie" is not commonly used in England, yet both times, Jazzer deliberately picks a Scots term, and the person he is speaking to shows no surprise and needs no explanation. It could be argued that 'cludgie' is used to display Jazzer's Scottishness, and does not follow patterns of linguistic convergence that might be expected, of Jazzer adapting his language to his English friends.

A similar scenario happens with "lassie". Writer 3 has Jazzer talking a number of times about the "cute lassies" he has met. However using the SCOTS corpus, it seems that "lassie" and especially "wee lassie" is used to describe young girls, rather than grown women. The only times that "lassie" describes an adult woman is when the speaker is generationally older and is describing the woman in a paternal tone. The word is being used correctly, but a Scottish person might query the lexical choice, and it seems that, as with "cludgie", the referent differs from the examples in SCOTS.

Although this analysis found that the content words which formed the lexical aspect of Jazzer's dialect were very limited in number and indexed a certain type of Scottish stereotype, it is also possible that the method of analysis led to a self-fulfilling prophecy: lexical items were only included if they seemed Scottish, so this excludes other slang words (an important part of Glaswegian). However, there are numerous Scots words, listed in taxonomies such as Hagan (2002) which are just not used by Jazzer, in favour of the very-well known dialect words such as "wee". In four hours of the programme *River* City, the word ‘lassie’ is not used once, suggesting that it is perhaps overused by *The Archers* writers. This could be that the writers have informally and collectively used a high instance of “wee” and “lassie” as part of Jazzer’s individual linguistic style. Equally, it could be that English writers have focused their construction of a Scottish dialect on a very small number of well-known Scottish words.

Grammar

The results for grammar showed a surprisingly low usage by the writers. This could partly be that by including the enclitic – nae endings under ‘lexis’, as well as the Scots ‘no’ meaning not, the coding was set up to find more lexical items. However, there were still many features which did not occur at all in the writers’ data: there were no double modals, although Grieve et al. found that these were rare, so this is not surprising; there were no Scots versions of past-tense verbs, e.g. “brocht”; and a lack of the Scots reflexive pronoun (e.g. “I’m away to my bed”), a complete absence of the Scots plural rule (e.g. “my foot’s gey siar”). Writer 2 had some examples of ellipsis, which contributed to the dialect. For example:

> "that you and Harry's moving in together."
>
> "I'm off for to get a drink."
>
> "You've only to ask Ed."

Subjectively, these all contribute to the dialect, and it is interesting that Writer 2, whose lexical dialect was lower, was slightly higher than the other writers in grammatical dialect occurrences. Arguably, increasing the use of grammatical portrayals of dialect, for example, “missing him something terrible”, is a subtler way to create dialect, rather than a reliance of famously Scottish words such as “wee” and “lassie.”

It is speculative to find a lack of authenticity based on what writers *could* have written, but there are, for example, many Scottish and Glaswegian words and phrases which could have been used, but are not. One notable absence is the lack of “How?” to mean “Why?”. This is a common

Glaswegian feature, which does appear in the *River City* script, but is absent from *The Archers* corpora. Other absences are “staying” to mean ‘living at’, “flitting” for moving house and so on. Another notable absence is the lack of fixed phrases and idiom, which is occurs frequently in other types of Scottish literature, and is present in the SCOTS corpus. One further missing feature that we might expect, is the use of taboo words (Macafee), but because of the nature of the data, writers are restricted in their use of strong language.

6.3.4 Results by Individual Writer

Having analysed the relative patterns of dialect usage for the six writers whose work spans seven years, what follows is a step-by-step analysis of how each writer creates Scottish dialect, using the entire length of data for each writer, instead of the first 1938 words only. This is not intended to be exhaustive, but is a selection of observations to compare the way the writers create Jazzer’s dialect.

Writer 1

Of the six writers, Writer 1 uses the greatest variety of lexical dialect items: fashes, bampot (spelled once as ‘bampot’, but four times as ‘barmpot’ despite the Scots Dictionary spelling of ‘bampot’), mithering, cludgie, and the North Eastern / Borders word, “scran”. The enclitic -nae ending appears only 15 times in over 13000 words, compared to 163 instances of “n’t” suggesting that the writer predominantly uses the standard English form. This could be that Jazzer, as a character who has lived in England for 10 years by the start of the data, and is always heard speaking to non-Scottish people, tends to use the Standard English form.

Writer 2

Writer 2, along with Writer 6, has a light approach to using dialect. Interestingly when other writers are portraying Jazzer’s attempts to meet women, Jazzer refers to them as ‘lassies’, emphasising his Scottishness. When this situation occurs in Writer 2’s data, ‘English English’ slang is used to describe women, such as "the blonde", and "stuck-up cows". This foregrounds a more sexist, predatory approach to women, rather than Jazzer's slightly whimsical Scottishness. All the lines that follow are taken from the same part of the same scene:

> Why is it that all **the best girls** are spoken for?
>
> Who's **the blonde** with the flower stuck behind her ear?
>
> You saying I can't pull **a posh bird**?
>
> I have met **a few stuck-up cows** in my time, but that one...
>
> I remember when **posh young women** used to appreciate a bit of rough.
>
> (Writer 1, 2013, emphasis added)

The lexical choices above suggest a more sexist, predatory version of Jazzer, as opposed to a slightly more whimsical Scottishness when the character talks about "lassies". In other instances though, Jazzer seems to use levels of politeness and hedging that we might not expect to find in someone who – in his own words – is seen as "a bit of rough". When asking his close friend if he can stay in her spare room Jazzer says:

> "You wouldn't notice I was here."
>
> (Writer 2, 2010)

Again, it is problematic to speculate on what he *could* have said, but using examples from literature and dialectology, we might expect something more direct, and more of a "chancer" approach, changing the subjunctive to the future tense: "You'll no notice I'm here." Writer 2 elsewhere uses this present tense for a future action, "I'm away to my bed", so we might expect to find it here.

Writer 3

Writer 3 uses "Och" and "Ach" and is the only one to do so consistently. These pragmatic noise items are famously Scottish, so could perhaps be indicative of an English person writing dialect, but they are also accompanied by frequent occurrences of the less iconic "eh?", which suggests a high usage of discourse markers and pragmatic noise generally. The writer tends to avoid using eye dialect, but and for the word "my" has 50 uses of "my" spelt the standard way, no instances of "ma" (which is used by other writers, especially Writer 4 to denote 'my'), but one occurrence of "mae": "I'm away to mae bed." Unlike "nae" which is a Scots dialect word, featuring in the National Dictionary of Scots, "mae" does not exist as a Scots word. Since this occurrence only happens once, it is possible that it was sub-edited in while the script was being prepared for studio. It only occurs in a sentence which uses a Scots grammatical structure: "I'm away to my bed". Perhaps it was this famously Scottish English phrase which led the writer to introduce eye-dialect for this particular occurrence of "my", but not for any of the other frequent occurrences of the word.

Lexically, there are number of mentions of iconic Scottish symbols in Writer 3's corpus: bagpipes, Burns Night, haggis, 'first footing' and Glasgow kiss. Whilst the first four occur in the context of discussing Scottish celebrations, and Jazzer could be deliberately emphasising his own pride in Scottish traditions, and using lexical terms for which there is no English alternative, the use of "Glasgow kiss" is interesting. Again, Jazzer could deliberately be performing his own Scottishness, and indexing the stereotype of Glaswegians as tough and straight-talking. Equally, this could be an English person using a famous example of something which is said *about* Glaswegians to attempt to *enact* a Glaswegian. Anecdotal evidence found during earlier studies (see Braber and Butterfint, 2008 and Braber, 2009) suggests that in spite of the negative stereotypes around Glaswegians, or perhaps even because of it, many Glaswegians are fiercely proud of Glasgow and use Glaswegian, with its covert prestige to signal solidarity among working-class speakers and the desire to maintain distinctiveness from other social groups (see for example Stuart-Smith, Timmins and Tweedie (2007)). At times, there have been complaints that Jazzer portrays a national stereotype of a hard-drinking, porridge-eating Scot (Ferguson and Singh Kohli, 2010). Debating this in *The Guardian,* the

“Glasgow kiss” is one of the examples given by expat Scot Euan Ferguson, along with ‘deep fried Mars Bars’, as being something that non-Scottish (often English) people use to invoke Scottishness, rather than something which Scottish people themselves actually use.

Writer 4

Writer 4 has by far the heaviest use of dialectal features, and is distinguishable from the other five writers through the high use of eye-dialect, as discussed above. The troublesome ‘braw’ is also used: “No, the pigs have been braw.” Again, it is speculative to suggest that other options might be more convincing, but it is noticeable that there are other options, less frequently-used outside Scotland, such as ‘gem’ or ‘greet’, which would fulfil the same message, but were not chosen.

The notion of “performing” Scots arises quite literally in a scene in Writer 4’s data. There is a scene where Jazzer bumps into a village acquaintance, Lilian, and an Englishman, Paul, who is having an affair with Lilian. In an attempt to disguise himself, Paul spontaneously decides to present himself as a Scottish man, which backfires when he discovers that Jazzer is Scottish.

> JAZZER Now, where exactly are ye from?
>
> PAUL Where… exactly are *you* from?
>
> JAZZER D’ya need to ask? Soon as I open ma mooooth? Glasgow!
>
> PAUL Yes… aye… o’ course. I thought so.
>
> JAZZER But you! I’m getting a bit of Fife, bit of Dumfries, bit of Edinburgh.
>
> PAUL Aye, that’s right. Well done! All of those. I moved around a lot. And I’ve been in England for years. Years and years.
>
> (Writer 4, 2013)

When Paul replies, “Yes… aye…. O’course”, the self-correction from “Yes” to “aye” suggests that this is a conscious feature of Scottish dialect, both for Paul and for the writer. Interestingly there is eye-dialect for “ma mooth” but not for Glasgow, which we might expect to see as “Glesga”. So it seems that there is a conscious performance of Scottishness by Paul, which takes him ‘up’ to a certain level of Scottishness, but as we might expect, this is outdone even further by the purportedly genuine Scot, Jazzer. However, since Jazzer is performing Scottishness, we might expect him to go even further in this instance and use “Glesga” instead of Glasgow. This could be viewed as a linguistic leaking of the non-Scottish ‘Glasgow’, or a limitation of the resources available to the writer in the construction of a Glaswegian man overtly indexing a strong Glaswegian identity.

There is another moment of Jazzer performing Scottishness, when he describes Burns’ Night celebrations:

> Rabbie –
>
> They both are. You pipe the haggis in, you say a poem to it...
>
> Aye! You stab it with your dagger...
> You’ll be saying you’ve never eaten haggis next.
>
> Sassenach.
>
> (Writer 4, 2011)

Jazzer corrects Robbie to Rabbie, is eulogising about the ceremony of addressing the haggis, but then refers to a “dagger” rather than the Scots word, “dirk” for dagger. Again, it seems that if Jazzer is consciously performing Scottishness, it is surprising that there are some Scottish words he chooses to avoid, in favour of their English counterparts.

It is notable that while Writer 4 uses “ma” as eye dialect for “my”, the same does not occur for the Geordie character, Ruth. Both writers use “mae” and “ma” for Jazzer, but not for the character of Ruth Archer. Ruth’s Geordie accent and strongly accented pronunciation of “oh no”, and “my” have been widely parodied (e.g. Sanderson, (2006)), yet in the scripts, all six writers use the standard English spellings and avoid eye dialect. One explanation is that, as Wells (1982) argued, the closer one is to ‘home’ the more accurately the dialect is heard. An English or Welsh writer might well hear a Scottish dialect as “other”, compared to a northern English dialect, which is still heard as sufficiently ‘close to home’ as to be written with the standard spelling. Another possible explanation is that Jazzer is a more broadly drawn comedy character in the drama, and as such, his portrayal is somewhat more ‘broadbrush’ than Ruth’s. This raises questions about which pronunciations authors perceive as “other” to the extent they are written in eye dialect and has elements of similarity to Ruzich and Blake’s findings regarding *The Help*, where black characters’ dialect was “othered” using eye-dialect, and the strong Southern American accents of white characters was largely minimised.

Another aspect of character creation – as with sociolinguistic analysis – is to consider the character’s background and education. There are a couple of moments where Jazzer seems unaware of cultural information which we would expect a Glaswegian man of his age to know. In this first example, discussing independence, he states:

> “It will be if we get independence. I’m surprised Rhys hasnae got leeks on the menu. And the Irish could have… what do they eat in Ireland? Shamrocks?”

Jazzer’s surname is “McCreary”, an Irish surname; he refers to his mother as “ma” not “maw”, which is an Irish influence, Glasgow is a city with historically high immigration from Ireland, and yet he seems unable to name any Irish food (e.g. Irish stew, colcannon) and instead settles for ‘Shamrocks’. Whilst it is clearly intended for comedy effect within the drama, as a piece of linguistic identity assumption it would be surprising that Jazzer seems so unaware of very basic information about Irish culture, and shows how the identity assumption can be weakened when these linguistic resources (Grant and MacLeod, 2018) are ignored. However, it is likely that the writer has knowingly sacrificed linguistic authenticity to prioritise the comedy.

Writer 5

Writer 5 does not use as much dialect as Writer 4, but one word which occurs relatively frequently is "wee", especially as a second adjective in a noun phrase:

```
"poor wee girls"
"poor wee lassies"
"crafty wee lasses"
```

Whilst other writers use 'wee' as a comedy diminutive ('wee bit hammered' to describe someone extremely drunk; 'wee chat' to describe a difficult conversation / dressing down from a manager), 'wee' here is used very much as a fond phrase to describe Jazzer's pigs. It also occurs as: "wee wains", which is problematic because "wains" (normally spelled 'weans') is a contraction of "wee ones", which we would expect Jazzer to know. For the non-Scottish writer, who presumably is not aware of the etymology, "wee wains" is not necessarily a tautology and follows a similar linguistic pattern to "poor wee girls", which the writer also uses. For the speaker who would know the etymology, "wee wee ones" is awkward and a Glaswegian would likely avoid the tautology.

Writer 6

Writer 6, who also had by far the fewest uses of pragmatic noise (Chapter 7), is similarly lower in their use of dialect features, but interestingly, has identical levels of dialect to the *River City* script. One line which is notable is when Jazzer says to his female friend, "Don't' be a Jessie!" Whilst this is appropriate for the context and tone, it seems an odd thing to say to a female friend. The definition of Jessie is: "*n*. Sc. usage: a contemptuous expression for an effeminate man" so its usage here betrays an unfamiliarity with the context in which it is normally used and seems to be a misapplication of the insult.

6.3.5 Conclusion

There are numerous points of consistency in the ways the writers approach Jazzer's language: in particular, the use of 'aye', some of the writers' usage of the tag question 'eh?', and the use of 'wee'. However, there are inconsistencies in the overall level of usage of dialect items between the writers, in particular between those who use eye dialect and those who do not. Further, there seems to be an inconsistency between the enclitic *-nae* / *-na* endings, which would be expected to be consistently either *-nae* or *-na* but not a mix, because they convey a speaker's accent.

There are also notable absences in the creation of an authentic Glaswegian dialect. Jazzer is a character who sometimes enacts his own Scottishness, but in the occasions where he is doing this, we might expect him to be even more Scottish than he is – some Englishness still creeps in. More grammatical features might be expected in his dialect too. Another notable absence is the lack of words which are commonly used in England but have a different meaning in Scotland, such as "stays at" meaning "lives at" and "How?" meaning "why?", and these are all ways in which his dialect could be more subtly indexed.

It is important to note, that the writers' purpose is, of course, to write an entertaining script, and so their purpose is very different from criminal or investigative cases of identity disguise. Arguably, the salient features of Jazzer as the jack-the-lad Scottish womaniser are indexed for comedic effect and dramatic colour, so it is unfair to criticise this as a failure of identity disguise, when the writers' purposes are closer to those outlined by Stockwell (above) of invoking certain elements of cultural schematic knowledge. However, there is no reason why using some of the subtler grammatical level dialect items would detract from this, and could strengthen Jazzer's linguistic characterisation.

This study shows that analysing dialect to investigate authorial synthesis can reveal differences between certain pairs of writers in their use of lexis, grammar and eye-dialect, when they are writing in a dialect, partly by the frequency of dialect items, but also broken down by the type of dialect items they use. Although studying dialect produced results which demonstrated inter-author variation, many of the results, such *canna*, or *cannae* could simply be detected using standard stylometric tests, such as a word-n-gram-based analysis, or a keyword analysis, without needing to search specifically for the use of dialect items. However, if, as Grant and MacLeod's (2016) experimental data showed, impersonators are able to focus on features such as the non-standard spelling of eye-dialect, it is quite possible that somebody attempting authorship synthesis would focus on these and be able to imitate them successfully. In my data, the writers are not attempting to match each others' spelling in a way that somebody trying to adopt another person's online identity might. In such scenarios, where imposters are able to focus on and imitate dialect features such as variant spellings, it might be that the writers would imitate spelling variations, but that higher-level features are areas in which we can detect weaknesses in the imitation.

6.4 Study 3: The (im)politeness strategies of Lynda Snell

The third qualitative analysis of characterisation explores the (im)politeness strategies of Lynda Snell. This also addresses the second of my research questions, which considers how closely the writers match each other in the creation of a character, focusing on Lynda's (im)politeness strategies. As explained in the Methodology in Chapter 4, this study compares how the different writers portray Lynda directing the pantomime.

6.4.1 Lynda's (im)politeness: Observations and Discussion

Lynda uses a number of imperatives in Writer 1's rehearsal scenes. In Culpeper's taxonomy, these are bald, on-record FTAs, where the speaker uses language in the most direct, clear, unambiguous and concise way possible, as in this example:

> "Pick up your things backstage, then wait with Hayley until you're collected."
>
> (Writer 1, 2010)

Lynda is described by other characters as being bossy, and bald statements such as these, support that view. However, in context, Lynda is speaking in her role of director of the village pantomime, so when she issues commands, this is not necessarily a form of impoliteness because she is fulfilling an agreed role, rather than assuming a position of power without authorisation. She is also speaking to children in this exchange, which – according to Brown and Gilman's framework for measuring FTAs – reduces the weightiness because her relative power is much higher than her Hearers'. In other instances, when directing adults, the imperative is softened slightly by the presence of a feature of formal politeness, "please", at the end: "Starting positions, please!" However, in Locher's terms, this use of "please" seems an unmarked, expected, use of politeness. Imperatives such as these, set her up as being an assertive character. She uses similarly "bald" approaches in politeness strategies, such as: "So, once you've had your notes, go and enjoy the rest of your evening" (Writer 1, 2010). Although the statement is an imperative, there is no real imposition. As Culpeper (1996) notes, such statements are in the Hearer's interest, so are unlikely to be interpreted as impolite. Culpeper observes that, "An order could be conceived as polite in a context where it is thought to be of benefit to the target (for example, "Go on, eat up" as an order for a dinner guest to tuck in to some delicacy)" (Culpeper, 1996:351).

Writer 1 also uses imperatives when Lynda is directing her cast:

> Creep furtively onto the landing, till you accidentally tread on Tommy's tail... Then rush to hide behind the grandfather clock. Where both of you lurk ominously, as Alice comes on in her nightdress.
>
> (Writer 1, 2010)

In this series of instructions, there are no mitigating linguistic strategies, but simply a list of instructions. Possibly, some of these instructions are an audio drama device to provide a description for the audience to explain the action taking place, so their primary function is arguably to give the audience a visual picture rather than to convey Lynda's personality. However, the instructions could still have performed the same visualising function even if the writer had chosen to soften them with use of modality, formal politeness markers, or questions such as "could you…?".

Writer 1 also uses imperatives to create comedy: for example, "Prop yourself up a touch more, dear. That top's perhaps a little lower-cut than you realise" (Writer 1, 2010). This is addressed to Sabrina Thwaite, a well-established as a non-speaking "silent character" in the programme, who always dresses and behaves in a sexualised way, raising eyebrows among other residents. Again, here, the imperatives are performing a dramatic function which is not just about Lynda's characterisation, but is also a comedic comment on Sabrina Thwaite.

In contrast to Writer 1's bald, unhedged imperatives, Writer 2 shows Lynda directing the cast, using these instructions:

> "And Lilian... Lilian, can I beg you, **please**, to spend a little time between now and curtain up this evening going over your script? And as for you, Jean... what can I say? Beautifully spoken, as ever. But could you **please**, **please**, **please** try and remember your moves."
>
> (Writer 2, 2015, emphasis added)

Here, the instructions are framed as a question, "can I beg you?", with the additional politeness of "please", and in the second sentence, "please, please, please." Framing this instruction as a question ("can I beg you…?") could be seen as very deferential, and less assertive than Writer 1's portrayal of Lynda's approach to directing. This could partly be explained by the different contexts: in Writer 1's instructions, "Creep furtively onto the landing…" and so on, it is a very early stage in rehearsals and Lynda is simply setting out what needs to be done. The example from Writer 2 occurs later in rehearsals and the cast have failed to follow instructions to Lynda's expected standards, so the different discursive approach – the repeated use of "please, please, please", and the questions, might not necessarily suggest a greater deference in the characterisation of Lynda; it could be reflective of her increased frustration and desperation. These lines also follow a speech where Lynda has thoroughly criticised her cast for performing terribly in the dress rehearsal, so the politeness features of "can I beg you?" and "please" seem to suggest frustration not deference. Locher (2006) discusses how overly polite language can become rude, and arguably the repetition of the polite words 'please' emphasises Lynda's strength of feeling, which could embarrass Jean, and in so doing, becomes impolite. Analysis is based purely on the text, and not the performance, but from the text alone, it is certainly possible to infer Lynda's frustration.

In a later rehearsal scene for a different year's panto, Writer 2 does show Lynda using imperatives with no hedging, as in this example, when Kirsty is dismayed that Justin Elliot has been cast as her onstage father:

> KIRSTY You know he's given Rob Titchener a job? When virtually everyone else in the village has recognised what a total scumbag the man is, and completely ostracised him, Justin Elliott turns round and says, Oh well, let's let bygones be bygones.
>
> LYNDA **Use it, Kirsty.**
>
> KIRSTY What?
>
> LYNDA Colinette, your character, feels trapped by her manipulative father, who won't allow her to marry the man she loves. So **let your feelings about Justin inform your performance.**
>
> (Writer 2, 2016, emphasis added)

As Culpeper argued, it is not possible to link a feature to a level of (im)politeness without considering context: although Writer 2 is using unhedged imperatives, Lynda's speech seems to be more encouraging and mentoring towards Kirsty, rather than making demands. Tonally, Writer 4 has very similar scenes, also between Lynda and Kirsty. In two similar speeches, Lynda speaks to Kirsty with a very similar pragmatic force:

LYNDA Remember what I said, Kirsty. At this point, you take both his hands and look up into his eyes.

(Writer 4, 2013)

Another similar example by Writer 4 is:

LYNDA: So let’s see some of that. Kirsty, you need to look at Rob the way… well, the way you look at Tom. With undying affection.

(Writer 4, 2013)

In both speeches, there are similarities to Writer 2. The imperative “Use it, Kirsty” by Writer 2, is tonally similar to “Remember what I said, Kirsty” by Writer 4 and “let’s see some of that.” The tone seems to be supportive and encouraging in each speech, rather than officious. One explanation might be that the conversation is between two characters who are established as being fairly close colleagues, and socially equal. Audience (within the fictional world, not the listening audience) also seems to influence tone. When Writer 2 has Lynda chastising Jean, she is speaking to her in front of the whole cast, which makes the weightiness of the impoliteness seem more severe, because it would presumably be more embarrassing to Jean than if the conversation had been a private one. These examples show ways in which Writer 1 and Writer 2 both have Lynda using imperatives with no or few mitigating strategies, but as the different contexts show, she is not necessarily behaving impolitely.

Writer 6 similarly shows Lynda using imperatives to encourage her cast, as in this example:

LYNDA Good. Close your eyes. Kirsty, you too.

KIRSTY Why? What are we doing?

LYNDA Just, just humour me. I want you to take yourself back to a romantic time in your life.

(Writer 6, 2013)

The example between Kirsty and Lynda seems to be different tonally from Writer 6’s other scenes where Lynda is directing a group: for example, when she calls out “come on – chop-chop!” to her cast (Writer 6, 2016). These extracts show ways in which the scriptwriters are very similar in the way they achieve a cohesive characterisation of Lynda in her more sensitive moments, coaching her cast.

During rehearsals, Writer 1 and Writer 4 both have Lynda blocking out a scene by telling the actors what to do and where to stand. Lynda’s instructions, such as “Creep furtively onto the landing”, could be seen as assertive, as a statement from Lynda to Sabrina, but at the higher level of literary organisation discussed in Simpson (1989), they can also be seen as a message from writer to audience, to explain what is happening to the non-seeing audience. Whilst the speeches are largely similar, Writer 1 often has Lynda offering reassurance and encouragement at the start of speeches, as in these examples, all taken from the same episode in 2010:

LYNDA **That’s it.** Nicely curled up, Sabrina.. Prop yourself up a touch more, dear. That top’s perhaps a little lower-cut than you realise.. Eddie, Nathan.

LYNDA **That’s right.** We’ll start with you miming it.

LYNDA **That’s it.** Creep furtively onto the landing, till you accidentally tread on Tommy’s tail...

LYNDA (SOTTO) **Very good, very good.** (CALLS) **That’s it Harry.** On you come. Remember you're half-asleep, so you think she's a vision.

LYNDA **Marvellous!** Then if Patrick were here…

LYNDA (STANDING UP) **Very much better.** We'll call a halt there.

(Writer 1, 2010, emphasis added)

There is a strong tendency for instructions to be preceded by some sort of reassurance or compliment.

When Writer 4 has a similar scene of blocking out the play, Lynda also offers reassurances, as in these examples:

LYNDA **Well done, Harry**! And there should be a big "aw..." from the audience.

HARRY (AS SELF) Hope so.

LYNDA So then you put your bundle on your back and trudge miserably upstage.

HARRY (MOVING FURTHER OFF) Like this?

LYNDA **Lovely.** (CLAPS HANDS) Right, Act one scene ten next please.

(Writer 4, 2010)

There is a similar tendency to soften the instructions with a reassurance or compliment at the start of each speech, which is consistent for both writers. Writer 6 also has a similar tendency to show Lynda starting each line with some form of confirmation or praise:

LYNDA **Good.** Take your places then. Tom, your first entrance is upstage right.

TOM (GOING) OK. (KIRSTY AND TOM CLIMB THE STEPS)

LYNDA And Kirsty –

KIRSTY (GOING) Yeah, I know.

LYNDA **Marvellous!** Let's make a start.

(Writer 6, 2013, emphasis added)

These three extracts again show the similarities between the writers in the way they depict Lynda's characterisation. For the other three writers, these "blocking out" scenes did not occur in their scripts so a comparison was not possible.

Whilst the tendency to give compliments and reassurances is common to the three writers, there are differences in the creative decisions concerning Lynda's directing lines. Writer 1 has Lynda issue an instruction, which is presumed to be followed, and the statements are separated. In contrast, Writer 4 uses the continuous present for instructions and physically directs her cast:

LYNDA You all move clockwise round the table. (ACTORS MOVE) About half way round... That's it. Then back... (ACTORS MOVE) Then clockwise again... All the way round this time. (ACTORS MOVE) And when you get back to where you started, rats break away, stage left...

(Writer 4, 2010)

The phrases ("about half way round", "Then back") in this extract produce quite a different tone to Writer 1's rehearsal scene. In Writer 1's equivalent scene, Lynda is portrayed as stationary: instructing the cast through verbal explanations rather than with implied physical gestures, and she paints a picture about the music and emotions. In Writer 4's more active rehearsal scene, there is less extraneous detail, and it is implied that Lynda is physically demonstrating how the cast should move.

Seven years later, Writer 4 uses similar language patterns in rehearsals:

LYNDA You come over here!

NEIL (APP) Where?

LYNDA Here. **Next to him**, Eddie.

EDDIE (APP, FED UP) All right.

LYNDA And Kirsty, **you next to** Harrison

KIRSTY (APP) Can't we all just go home?

LYNDA Certainly not. (GOING) **Now, fairies behind them. Susan in the middle.** Come on, come on, our audience is waiting!

(Writer 4, 2017)

It is interesting to see that, seven years apart, there are linguistic similarities, such as the elided verbs in "Then back" and "then clockwise again", and "Now, fairies behind them", which have echoes in "Next to him, Eddie" and "you next to Harrison." There is a similar sense of Lynda as an active director, physically showing her cast where to move, partly suggested by the deictic "here", in contrast to the directing scenes of Writer 1 and Writer 2, where move specific instructions are given in the dialogue itself, rather than being implied in the (unseen) action. However, as in the previous Writer 4 rehearsal scene, which begins "Well done, Harry!", Writer 4 has written other scenes which are closer in stance to those written by Writer 1 and Writer 2. It is possible to observe that Writer 4 has a tendency to write rehearsal scenes with deictic instructions and elided verbs, but also sometimes writes rehearsal scenes which are tonally closer to those of Writer 1 and 2. Whilst Writer 4 shows some consistency in the protagonist's stance in two scenes, seven years apart, it is not possible to say that the writer would *only* use this technique to script rehearsal scenes with Lynda blocking out the play's movements. For this reason, the 'active' scenes could tentatively be associated with Writer 4, but the more instructional stance is not exclusive to Writer 1 or Writer 2, and, indeed, there is no evidence that Writer 1 or Writer 2 would never use the stance adopted by Writer 4's Lynda.

In a different scene by Writer 4, Lynda is directing the action, and also uses unhedged imperatives, and ellipsis, creating a sense of immediacy between director and cast. This reinforces the tendency of Writer 4 to stage rehearsals with a consistently similar tone:

LYNDA Stage right... pass him the bucket, Nathan... Loom over them, Sabrina...

EDDIE This bucket feels -

LYNDA Don't talk, you're in mortal danger, Eddie! Throw the confetti!

(Writer 4, 2010)

In this scene, compared to the other two rehearsal scenes, Lynda is far blunter, interrupting Eddie when he says "this bucket feels –". However, this is partly explained by the plot of the scene: somebody has switched the confetti in the bucket for wallpaper paste, to play a prank, so it is dramatically necessary for Eddie to foreground the prank by noticing that something is amiss with the contents of the bucket, but for this not to be fully discovered before he throws the bucket over Sabrina Thwaite. Lynda's interruption, then, is not just about her urgency and bluntness as a director, but is also important to the plot: her insistence increases the scene's urgency in the build up to the prank. Also, if Eddie is allowed to stop and investigate why the bucket is suspiciously heavy, the prank will not happen.

One area where the writers show some difference is in the interpersonal phrasing of instructions. Compared to other writers, there seems to be a tendency for Writer 3 to couch directives in personal terms where Lynda refers to herself and her thoughts, rather than a reliance on imperatives and "Let's", as used frequently by Writer 4. These lines, all from Writer 3, show the increased tendency to couch directives in interpersonal pronouns.

LYNDA No, no, Jim, **I’m** sorry but **I** can’t have **you** dropping your voice like that.

LYNDA On every ‘honourable man’. **You**’re going to have to really push those out to the audience.

LYNDA Jim, **I** think **you** greatly underestimate your own vocal capacity. (HARDER) The fact is **I** could hardly hear you.

LYNDA Preparing the brawn, **I** believe.

LYNDA Be that as it may, **I’m** sure on the night **you’ll** have a rapt audience.

LYNDA Yes – (CALLING) Now this is where **I’m** hoping the lights can gradually fade, Neil?

LYNDA **I thought** it was a most commendable effort.

(Writer 3, various years, emphasis added)

Writer 3 has a tendency to use verbs relating to cognition, such as ‘seem’, ‘believe’, ‘thought’ and ‘hope’. This creates a slightly different character style for Lynda, compared to other writers’ version of Lynda. Here she seems a little more thoughtful in her relational work, and makes an effort to use interpersonal language, rather than directives to a group, softening the FTA directives. Of course, this is not exclusive to Writer 3. In this extract, Writer 4's Lynda uses a similar strategy:

LYNDA	Well … At the moment, **I really don't feel you're** achieving the right level of intimacy. And now it's occurred **to me** that **you two** probably don't know each other very well in real life. **You aren't** aware of what makes each other tick.

(Writer 4, 2013)

As with the directorial stance in rehearsal extracts quoted above, it is possible to see that certain writers have tendencies to use certain techniques when showing Lynda doing relational work, but it is not possible to show that certain techniques are exclusive to one writer and authorially distinctive.

A different aspect of relational work is the way Lynda responds to conflict or criticism. Writer 2 shows Lynda feeling piqued that two of her cast members are attempting to re-write her pantomime script. Lynda's response here is to use a constative reply to counter Fallon's criticism, rather than to attack directly:

FALLON	I just don't understand my motivation.
LYNDA	(OFF A BIT IN HALL) Fallon, this is a pantomime.

(Writer 2, 2010)

When Fallon continues to criticise the script in the same scene, Lynda remains polite:

FALLON	It's just the way it's written…
LYNDA	May I remind you, this is a tried and tested script. It went down very well in Sunningdale.

(Writer 2, 2010)

In this second extract, Lynda refutes Fallon’s face-threatening comments with constative comments, similar to the pragmatic force of “this is a pantomime”. Lynda’s responses to Fallon’s (im)politeness lend an authority to the argument. By avoiding the personal pronouns, Lynda counters the FTA by remaining polite, but not giving ground: she describes the merits of her own pantomime script in a detached, neutral manner, as if presenting fact rather than opinion.

Writer 4’s Lynda uses a similar strategy in these two extracts. Firstly, Writer 4’s Lynda argues, “This is no time to talk about films!” (Writer 4, 2016), which echoes a snappy response by Writer 2’s Lynda, “We are not here to have fun!” In both cases, Writer 2 and Writer 4 have Lynda avoiding a bald on-record piece of impoliteness, and instead, write her as initially couching her disagreement in a slightly more indirect constative statement, until the argument builds to a more direct confrontation.

This style of responding to confrontation also occurs in the following extract, when Kate has failed to make suitable costumes for *Mother Goose* as promised:

LYNDA (BEAT) There must be more to come, surely.

KENTON (READS LABEL) No… Just says ‘Mother Goose, Pool of Beauty’.

LYNDA But it’s nothing like Kate’s sketch. You can’t come out of it in just that! There’s barely anything of it!

(Writer 4, 2016)

Similarly to Writer 2, Writer 4 uses two constative statements to protest against Kate’s unsuitable costume design (“But it’s nothing like Kate’s sketch” and “There’s barely anything of it!” However, embedded in the middle is a blunter response, in the directive “You can’t come out of it in just that!" There is a similarity to Writer 2's Lynda, with the presence of an additional directive to argue more directly with the other character present.

An issue with analysing (im)politeness or relational work for authorship attribution purposes is that writers can and do draw on multiple techniques. This extract, by Writer 6, shows a shift in Lynda's (im)politeness strategies in two consecutive lines:

LYNDA You must have some good memories with Tom. Maybe you could channel those into your performance.

KIRSTY Lynda –

LYNDA (ALOUD) Right – that's enough! From now on I want silence!

(Writer 6, 2016)

Lynda softens her directive to Kirsty with the hedge, "Maybe", which is in contrast to her unhedged statement of "Right – that's enough!" when she addresses the cast at large. This blunt instruction is followed up with another statement, this time an assertive, but with a directive force: "From now on I want silence!". As would be expected in a drama, the characters use different strategies within a scene to create tonal variety and interest. Since the primary purpose of the data is to entertain its audience, it would be surprising if Lynda did not use a variety of (im)politeness strategies to interact with other characters.

6.4.2 Conclusion

Whilst some similar strategies and some differences have been identified, it is difficult to use these strategies to discriminate with any accuracy between the writers. One reason is that each context is different, so comparisons are problematic; in Ohmann's (1964) terms, it is impossible to separate the content from the style. Even in the comparable situations of pantomime rehearsals, each situation is affected by context, storyline, and internal audience, which makes it problematic to compare levels of politeness between different scenes. For example, if one rehearsal is running successfully, a later one will be full of problems. In other cases, such as the bucket-throwing prank, the needs of the plot constrain the politeness strategies which Lynda can use – she has to urge Eddie to act quickly for the mechanics of the prank to work. In other instances, such as Sabrina's low-cut top, what might seem to be a bald FTA, is arguably softened, because a long-term listener would recognise this as a joke about a renowned silent character.

Perhaps different, more tightly controlled data which does not have the deliberate variations that a dramatic text does – for example customer services logs – would show clearer differences in (im)politeness techniques which are attributable to differences in authorship. In *The Archers* scripts, the differences in (im)politeness can be partly, or even mainly, attributed to differences in context, rather than differences in authorship. Further, the selected examples showed ways in which writers were able to imitate each other's (im)politeness strategies very closely, and that even where differing tendencies could be observed, these were not exclusive to a single writer.

6.5 Three Character Studies: Conclusion

From these three character studies, I found that the study of Jazzer's dialect was the most promising of the three studies for analysing linguistic imitation and authorship synthesis. As was shown, the writers had different approaches to writing dialect, which are both individually consistent and showed variation from other writers in the group. It was possible to demonstrate ways in which the portrayal of dialogue was ideologically loaded and revealing about what the writer heard as "other", for example, comparing the heavy use of Jazzer's dialect to the virtually non-existent use of Ruth's Geordie dialect. This, in turn, could be used to infer information about the writer's own linguistic resources and profile.

It was also possible to identify at the levels of lexis, grammar and pronunciation (eye-dialect) examples of linguistic leakage, or a failure to suppress the English authorial voice, as suggested by the over-use of politeness features. One way in which the dialect study was able to demonstrate weaknesses in the writers' linguistic identity disguise was through the inaccurate meanings: for example "lassie" was correctly included as a Scots dialect word by Writer 3, but was used to describe young women, rather than the more commonly applied meaning of young girls.

In the studies of Jim's lexis and Lynda's politeness, there were discernible similarities, but one shared problem was the heavy influence of plot and context on any linguistic observations. Also, the various analytical frameworks used (vocabulary richness, keyword analysis, Latinate words and (im)politeness strategies) returned results which gave information about events and relationships within the fictional world. The keyword results for the Jim Target corpus showed results which are not context-bound (*ah*, *of* and *the*), and could potentially be more revealing of Jim's habitual patterns of speech, but in the relatively small corpus, it was not possible to demonstrate this. The keyword results for the smaller corpora for individual writers were words relating to current storylines. As such they did not discriminate between authors.

In terms of my superordinate research question, it seems that the writers are able to imitate each other successfully in their lexical choices, and in some shared features of politeness strategies. In the context of drama, where there is a focus on effect, characterisation can be viewed as a higher-level objective. One explanation is that those linguistic techniques which are most telling about characterisation receive a greater level of focus, as illustrated by the close imitations and shared linguistic techniques.

7. The Functions of *Oh*

7.1 Introduction

My final study explores whether there are discernible differences between six scriptwriters in the way they use language, by carrying out an inter-author and inter-character comparison of the function of the word *oh* in the data. This study addresses the third and final of my research sub-questions, asking whether 'higher-level' pragmatic features can provide a base for authorship analysis in cases of linguistic identity disguise. As discussed in 2.5.1, Grant and MacLeod advocate using all four domains of language to explore the creation of linguistic identities:

> We see identity as a phenomenon best classified at the level of social behaviour, but it must be kept in mind that the identities projected by individuals are produced with the resources available to them at all three of the other linguistic levels. (2020:39)

My study uses pragmatics to consider whether analysing the function of a word, rather than the word itself, can discriminate between authors. Firstly, the functions of *oh* are discussed, with examples from the data. Then, in 7.4, the token *oh* was counted, to compare frequency of usage. This was broken down by character, to analyse whether *oh* was used more by some characters than others. Finally, a pragmatic analysis was carried out, using the coded functions of *oh*, to evaluate whether considering the pragmatic function of a word was able to discriminate between more pairs of authors.

The main, and largest, corpus for this study was the Helen and Rob corpus, which consisted of all the duologues from a major storyline about coercive control in *The Archers,* beginning in 2013 when long-standing character Helen Archer met the newly-arrived Rob Titchener. The data include their relationship, from the beginning of their affair, through to their marriage, and ends when Helen eventually stabs Rob. The second corpus is of Lilian, another long-standing character, who began an affair with her partner's recently discovered half-brother, Paul. The third corpus was also of two long-standing characters: Elizabeth, the widowed owner of a stately home, and her married manager, Roy.

7.2 Codifying *Oh*

This next section discusses the examples of each function of *oh*, illustrated with examples from the data. The relevant literature informing the coding has been discussed in 3.6, and the test-re-test reliability procedure for the coding is discussed in 4.6.2. The following examples describe and discuss functions for which *oh* was used in my data, to illustrate in more detail the different functions considered in the final analysis of this chapter. The first of these categories is *oh* as conventionalised phrase.

7.2.1 *Oh* as Conventionalised Phrase or Vocative

There are many instances of *oh* being used in a two-word, or occasionally three-word phrase. Examples in the data include:

> LILIAN: **Oh, Paul.** I can't tell you how much I've been looking forward to this.
>
> (Writer 5, 2012)
>
> PAUL: **Oh, Lilian.** Look maybe I could say I was ill, or –
>
> (Writer 5, 2012)
>
> ROB: **Oh Helen,** it's such bliss to have you here.
>
> (Writer 2, 2014)
>
> HELEN: **Oh, my goodness,** what am I going to cook?
>
> (Writer 2, 2014)

Similar examples provided by Aijmer (1987) include *oh dear*, *oh for God's sake*, *Oh God*, *oh my word*, *oh hell*, and *oh heavens*. As she points out, the *oh* is optional except for with 'dear'. This use of *oh* corresponds with Culpeper and Kytö's examples of the historical "O", suggesting that it could be viewed as a homophone of *oh*, which has since become a polysemous word with standardised spelling. Sometimes writers used a comma between the *oh* and the name or conventionalised phrase, but not always, and this was not a criterion for inclusion, so in the examples above, all four examples were coded as conventionalised phrase / vocative address. In each one, the *oh* carries an emotional intensifier, except for "oh my goodness", where it forms part of a conventionalised phrase. In this extended example, below, compare the vocative appearances of "Oh Helen" and "Oh Rob" to the third occurrence of *oh*, "Oh! Henry!". This last occurrence was coded as "3. Surprise" rather than Vocative, because the *oh* is tonally separated from "Henry", almost as two separate, subsequent thoughts. The punctuation shows a separation between the pragmatic intent of "Oh!" before the new tonal unit, Henry. The prosody of "Oh Rob" would be significantly different from "Oh! Henry!" In the former, the *oh* operates to intensify the speaker's emotions; in the latter, it signifies surprise.

> ROB **Oh** Helen, it's such bliss to have you here.
>
> HELEN It's such bliss to be here. (ANOTHER LINGERING KISS) Oh Rob... (DOOR OPENS. HENRY ENTERS)
>
> HENRY Mummy.
>
> HELEN **Oh!** Henry. What's the matter darling? Can't you sleep?

(Writer 2, 2014)

Although lines spoken to Henry were excluded when compiling the Helen and Rob corpus, this particular instance was retained because the "Oh!" is spoken to Rob, as well as to Henry. The distinctions between categories are of course not entirely watertight. For example,

> HELEN Mm. **Oh, Rob.** I just want to be honest. With mum and dad. With everyone. I want us all to be clear about things.

(Writer 3, 2014)

Arguably, this instance of "Oh, Rob" could have been coded as "3. Surprise" because it marks the moment when Helen has a new thought, but the coding was prioritised in the order it is listed, so that *oh* plus the vocative use took precedence.

7.2.2 *Oh* as Agreement

In these data, agreement was defined as two broad areas: agreement of opinion, and consent. Culpeper and Kytö note that *oh* reinforces an affirmative answer to a yes/no question; of these two, *oh* is used far more frequently in positive answers, i.e. collocating with 'yes' rather than 'no' (2010:242). However, this is perhaps truer historically than in present-day English. Often agreement manifests itself as "oh, yes", but this is not the only way that agreement appears in the corpus, and, secondly, the appearance of "oh yes" does not automatically mean that the occurrence was coded as agreement. First, an example of straightforward agreement:

```
LILIAN    You were right. It is popular.
PAUL      You like it, though?
LILIAN    Oh, yes. It's very smart.
```

(Writer 4, 2012)

In this case, Lilian has not received any new information, and is simply expressing her agreement. However, the tokens, "oh yes" do not automatically mean the occurrence of *oh* was coded as agreement: in some cases "oh yes" was used as a form of acknowledgement. For an occurrence to be coded as 'agreement' rather than 'acknowledgement' it was necessary for the speaker to have knowledge of the prior statement, in contrast to 'acknowledgement', where the speaker is receiving the new information (cf. Heritage's (1984) "information receipt") along with their concurrence to the previous statement. This occurs here:

```
LILIAN    You can't manage anything else?
```

PAUL I've had plenty, thanks. You could feed the five thousand with this.

LILIAN Well, it won't go to waste. I'll drop some off at mum's later.

PAUL **Oh yes**- you said she's been ill.

(Writer 1, 2010)

In this example, Paul is not agreeing that Lilian should visit her mother, but rather is acknowledging that he has recalled being told that Lilian's mother has been ill. This second occurrence was coded as "4. Acknowledgement" because he is acknowledging the receipt of information.

"Oh, yes", (or its variant form, "oh, yeah") was unsurprisingly the most frequent way in which the function of 'agreement' appeared, as in this example:

ROB Well, I – I'd better get on. Busy, you know.

HELEN **Oh yes**. Me too.

(Writer 3, 2013)

In the wider context of this interaction, Helen has bumped into Rob in a bank. There is mutual attraction but Rob is still married, and after an awkward exchange of small talk, Rob makes his excuses to leave. Arguably Helen's "oh yes" could be coded as "4. Acknowledgement" in response to the new information that Rob is busy, but the major tone of the exchange is that both parties mutually feel the need to end their encounter, and as such, 'oh yes' marks Helen's agreement with Rob's justification for leaving, and the following "Me too" supports this view. At a more literal level of textual analysis, Helen is agreeing with Rob that she appreciates he is busy and needs to leave, and also that she too is busy so cannot stay to chat. Subtextually, there is a mutual sense of needing to end their conversation because they are both denying their attraction towards each other, and depending on performance, Helen's "oh yes" could signal awkwardness at the situation, or an overt casualness which is at odds with her feelings towards Rob, so although it is agreement, it has elements of other functions. At a discourse analytical level, Rob is beginning a ‘closing sequence’ to their conversation, and Helen is agreeing with this and continuing the ‘closing section’ of their talk.

Other, less frequent, occurrences included “oh, I will”, and “oh, I do”, and “oh, me too”. For example:

ROB But as soon as possible I want everyone to know we’re together.

HELEN Oh, me too. I want to shout it from the rooftops!

(Writer 3, 2013)

Somewhat perversely there were instances of “oh no” being used to indicate agreement, as in this example:

HELEN Do you mind if I…

ROB **Oh no.** Plenty of room. Sit down.

HELEN (SITS DOWN)

(Writer 1, 2013)

In its response, the use of “oh no” answers the negated form of the question (“do you **mind** if I?”). This could be coded as “spokenness / downplayer” (discussed below), but since the pragmatic thrust of the sentence Helen giving consent for Rob to sit next to her, it was coded as agreement.

As a side note about the presence of “oh no”, I considered using this as a separate category in its own right, or combined with “oh yes”, where each *oh* was being used as an intensifier’ to add emphasis to the agreement / disagreement. However, in the corpus, the majority of occurrences of “oh no” did not mirror the intensifying effect of the majority of the occurrences of “oh yes”. Sometimes the *oh* seemed to heighten the emotion of “oh no”, mirroring the intensifying function of *oh* in “oh yes!", as in this example, where Helen is panicking about having to spend time on her own with her coercive husband:

```
ROB       Either way, your mum can have Henry, so
          it's just the two of us.

HELEN     Oh no! (QUICK) I mean, I don't want him
          staying out overnight in case he does wet
          the bed again.
```

(Writer 3, 2015)

However, "oh no" mostly occurred with less of an intensifying function, and sometimes (as Heritage and Aijmer each observed) to show that an answer has not received a great deal of prior thought, as in this example:

ROB So when's it to be? Tonight? Tomorrow?

HELEN **Oh**, no, not this early in the week. Um … How about Thursday?

(Writer 3: 2015)

It would have been problematic to group these two together in a single category, or into complementary categories because "oh no" was often used with quite a different function from "oh yes".

7.2.3 *Oh* as Surprise

Oh is used to express surprise, but in this coding, it is a specific interpretation of surprise, corresponding with Aijmer's (1987) description:

> *Oh* and *ah* can be associated with an interruption or intervention in the conversation at the point at which a person reacts to an unexpected situation. This is the case if the speaker suddenly has a certain insight but also if he guesses or infers something, remembers or recognizes something, notices or observes something, or successfully solves a problem. (Aijmer, 1987:63)

This description carries a sense of the surprise being defined in terms of its structural position within the conversation, rather than the content of information – i.e. the content of the surprise does not necessarily have to be surprising to the listener, but the sense of interruption to the train of thought needs to be surprising. Often, there is an overlap between function in the sentence and the speaker's emotion. This example was coded as surprise.

```
(BUSY  PEOPLE  TALKING  EATING.  CUTLERY  TRAYS  GOING
DOWN ETC. ROB AND HELEN ARE EATING SANDWICHES)

HELEN     Oh,  look.  Someone's  left  a  copy  of  the
          Borchester Echo behind.
```

(Writer 5, 2014)

The sense of surprise in this codification requires it to be self-initiated, or non-linguistic, as in this example, and the level of surprise can be extremely mild, as in this example.

7.2.4 *Oh* as Acknowledgement

As described by Macaulay (2005) and Heritage (1984), Acknowledgement can be a neutral, or more emotional, show of receiving information, as in the example below:

```
HELEN     You're away again?

ROB       No...  er...  (SIMPLE)  Jess  is  coming  up
          this weekend.

HELEN     (TRYING TO HIDE DISAPPOINTMENT) Oh.
```

(Writer 1, 2013)

Sometimes the *oh* is a complete sentence, or line, as in this example. In other cases, it is part of a longer sentence, and indicates the character's and/or scriptwriter's stance on the information received (which corresponds to Evaluation, discussed by Aijmer, passim):

```
          HELEN     Yes - but you soon put it right.
```

ROB The moment I noticed. Only, apparently that's just 'covering up'.

HELEN **Oh**, that's outrageous.

(Writer 1, 2015)

7.2.5 *Oh* as Downplayer / Conversational element

The fifth category in the coding is the "conversational oh". Culpeper and Kytö state that part of the reason why pragmatic noise appears, "may be the wish to (re)create an illusion of spokenness" (2010:200). Whereas most literature on the functions of *oh* (Heritage (1984) Ameka (1992), and Aijmer (1987) among others) analyses naturally-occurring data, Culpeper and Kytö is highly relevant for my data because it explores the use of *oh* in play-texts. Often *oh* occurs as the answer to a question. For example:

ROB Er .. How long are you going to be?

HELEN **Oh**, I don't know. Another hour at least.

(Writer 2, 2015)

Sometimes it is less enthusiastic than agreement. For instance:

ROB Unless you and Tom don't want me helping out ...

HELEN **Oh no** – it's great you're taking such an interest, Rob.

(Writer 1, 2015)

Another way in which an occurrence of *oh* could be coded as "5. Spokenness" is hesitation. For example:

PAUL Come on, Lilian, you're pretty resourceful. You'd think of something.

LILIAN **Oh** … I don't know…

(Writer 3, 2012)

Hesitation is a slightly awkward fit within the category of spokenness, but spokenness seemed to be the best fit category since hesitation is a feature of spoken language, and occurred so infrequently that a separate category was unnecessary.

7.2.6 *Oh* Topicaliser / continuer

The final category included in this analysis is of 'oh' as a continuer. This occurred so infrequently that it would – for statistical purposes – have been preferable to subsume it within another category, potentially as part of 'acknowledgement'. However, its function is different from a simple receipt of information, and rather than evaluating the information received, it prompts the first speaker to continue (or in the case of a topicaliser, it prompts the previous speaker to focus on a particular aspect of the conversation, as indicated by the fellow conversant.) One example of hesitation is in this exchange:

```
HELEN     I'm free for a couple hours and I just
          thought…

ROB       Oh yes? What did you think?
```

(Writer 6, 2013)

Schiffrin discusses *oh* as a topicaliser or continuer, using the term, "backchannel" (where *oh* does not cause the first speaker to end their turn, but continues almost over the second participant's *oh*). An alternative view to *oh* being a backchannel is that the hearer's placement of *oh* shows the first speaker which area to focus on, and is thus a topicaliser.

```
LILIAN    He's fine though. Busy hatching his plans
          for when he gets out.

PAUL      Oh really?
```

LILIAN He wants to set up a new company.

(Writer 6, 2010)

Paul’s use of “oh really” is partly an acknowledgement of Lilian’s information, but is Janus-like in the sense that it acknowledges Paul’s previous statement and simultaneously prompts his next one. It is also a useful feature in a radio drama, because as Wyatt and Grove observe, if a character does not speak for a while, the audience may lose the sense of their presence (2013). If one character has a long speech, topicalisers such as *oh* provide an audible way of keeping the listening character(s) ‘present’ in the scene. As such, its use may be indicative of text-type, rather than authorial style. However, since all the writers are writing for the same-text type, any variation in use is an authorial variation.

7.3 Results and Discussion

7.3.1 Descriptive Results: *oh usage*

At the structural level of language, the total occurrences of *oh* for all six characters in the combined three corpora were counted and normalised per 1000 words (Table 33 / Figure 8). These results show that two writers (Writer 1 and Writer 2) use *oh* almost twice as frequently as Writer 6 at 8.67 and 8.86 uses per 1000 words, compared to 4.52 words per 1000 for Writer 6.

Table 33: Combined occurrences of *oh* per 1000 words (normalised)

Writer 1	8.67
Writer 2	8.86
Writer 3	7.81
Writer 4	6.05
Writer 5	6.41
Writer 6	4.52

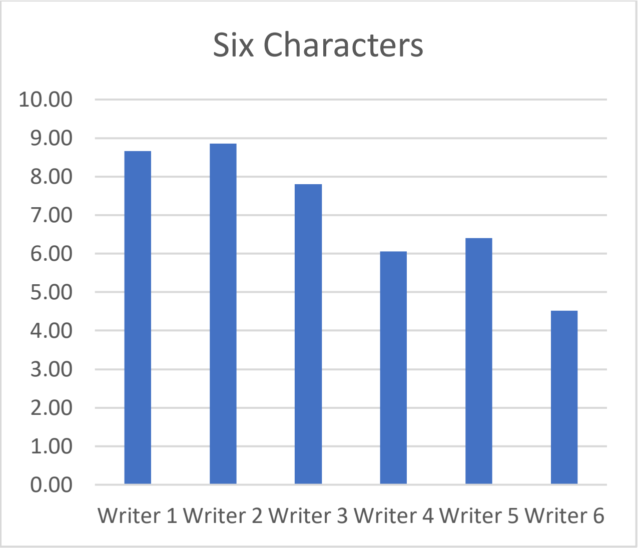


Figure 8: *oh* usage by writer (all six characters)

7.3.2 Helen and Rob Results

Secondly, the frequency of *oh* was counted using only the Helen and Rob corpus; firstly as a combined corpus, and secondly for each character. The results for Helen and Rob combined show a fairly similar pattern (Table 34/ Figure 9). This is to be expected because the Helen and Rob corpus was by far the largest of the three corpora, so had most influence on the combined results above (Table 32 / Figure 9). In the Helen and Rob corpus, Writer 1 has a relatively high frequencies of 8.66 occurrences per 1000 words, which is more than double Writer 3's frequency, with Writer 2 the next closest. Writer 6 remains notably lower than the other writers. Writers 3, 4 and 5 are all in a broadly similar range, with only 0.43 difference between highest to lowest of these three.

Table 34: Helen and Rob corpus *oh* usage

	Word count	*oh*	Normalised per 1000
Writer 1	9925	86	8.66
Writer 2	5788	43	7.43
Writer 3	8559	54	6.31
Writer 4	7270	49	6.74
Writer 5	5663	37	6.53
Writer 6	7437	30	4.03

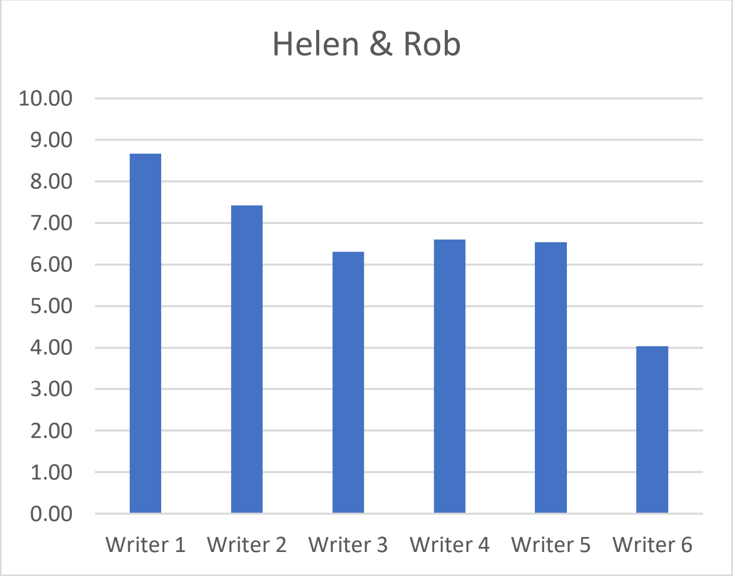


Figure 9: *oh* usage by writer from the Helen and Rob corpus.

However, when breaking the results down by character, and separating Helen's lines from Rob's, the results show some even greater distinctions between pairs of authors (Table 35 / Figure 10).

Table 35: Use of *oh* in Helen and Rob corpus, separated by character

	Helen			**Rob**		
	Word count	***Raw Freq. of Oh***	**Normalised per 1000**	**Word count**	***Raw Freq. of Oh***	**Normalised per 1000**
Writer 1	3887	55	**14.15**	6038	31	**5.13**
Writer 2	2532	25	**9.87**	3256	18	**5.53**
Writer 3	3891	39	**10.02**	4668	15	**3.21**
Writer 4	2718	38	**13.98**	4552	11	**2.42**
Writer 5	2427	21	**8.65**	3236	16	**4.94**
Writer 6	3388	14	**4.13**	4049	16	**3.95**

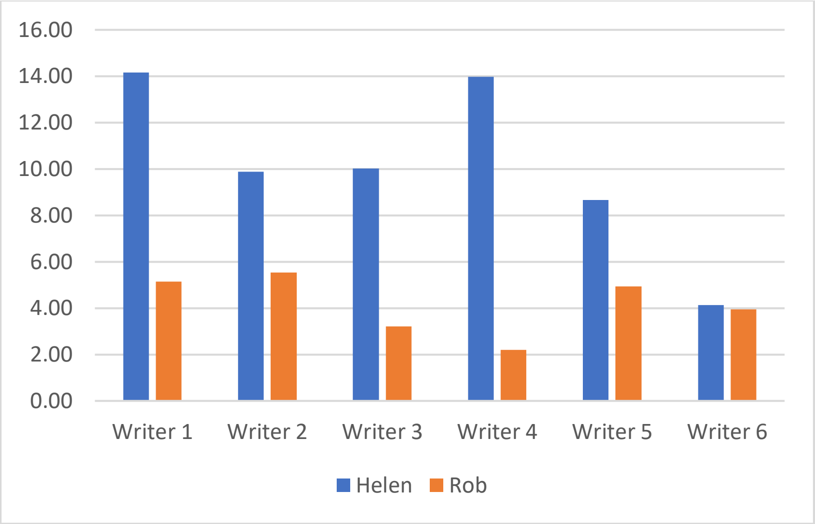


Figure 10: Helen and Rob, usage of oh, normalised per 1000 words

With the exception of low *oh* user Writer 6, the five remaining authors all use *oh* far more frequently in Helen's lines than in Rob's, most significantly Writer 4 with 13.98 for Helen compared to 2.42 for Rob. For Helen's lines, Writer 1 uses *oh* more than 3 times as frequently as Writer 6, showing an observable difference in the way Writer 1 writes Helen compared to Writer 6. Separated by character, Writer 4's use of *oh* becomes notably higher than Writer 2's (13.98 per 1000 compared to 9.87 per 1000), whereas the combined results for Helen and Rob had shown Writer 2 to be the second highest user. Results shows a greater disparity in the way Writer 4 writes Helen compared to Rob. Writer 2 also showed a higher usage of *oh* for Helen than Rob, but the difference between the two characters is less marked.

These results suggest that five out of the six writers adapt their usage of *oh* as part of the way they create characters, because there is a marked difference between the frequencies of Helen and Rob. Yet, although five of the six writers use *oh* far more frequently for Helen than for Rob, there is still variation relative to each of the other scriptwriters in how often they use the token (Figure 10), with Writer 1 and Writer 4's Helen having a notably higher usage, while Writers 2, 3 and 5 have a fairly similar occurrence rate. This suggests that individually, writers have adapted their style to write Helen compared to Rob, but are not particularly consistent as group.

7.3.3 Lilian and Paul, and Elizabeth and Roy Results

The results for the two smaller corpora, "Lilian and Paul" and "Elizabeth and Roy" are less distinctive (Figure 11 and Figure 12 respectively). There is a tendency for the writers, in particular Writer 3 and Writer 6, to have Lilian using *oh* more frequently than Paul. The results for Elizabeth and Roy are more mixed, although Writer 5 has Elizabeth using *oh* relatively frequently, whereas Roy does not use it at all.

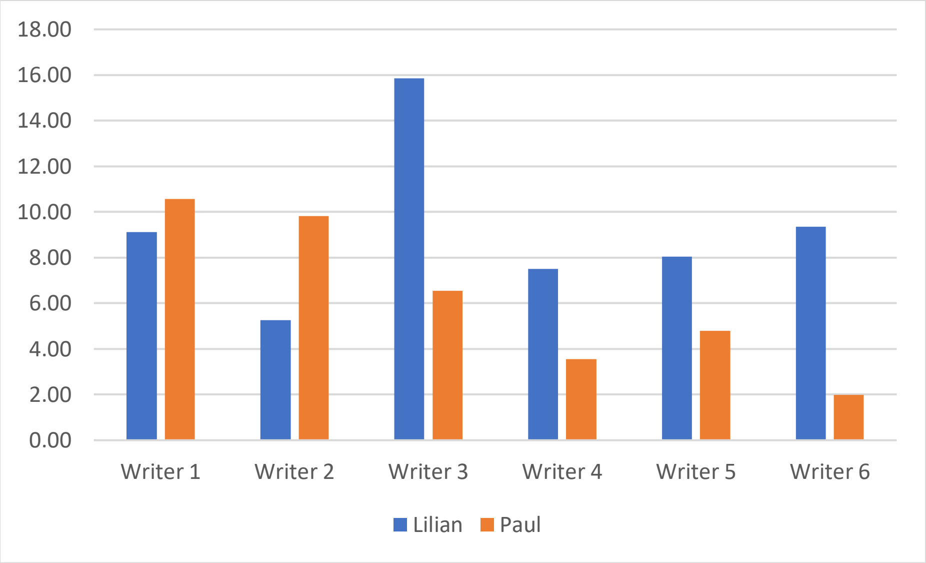


Figure 11: Lilian and Paul, usage of *oh*, normalised per 1000 words

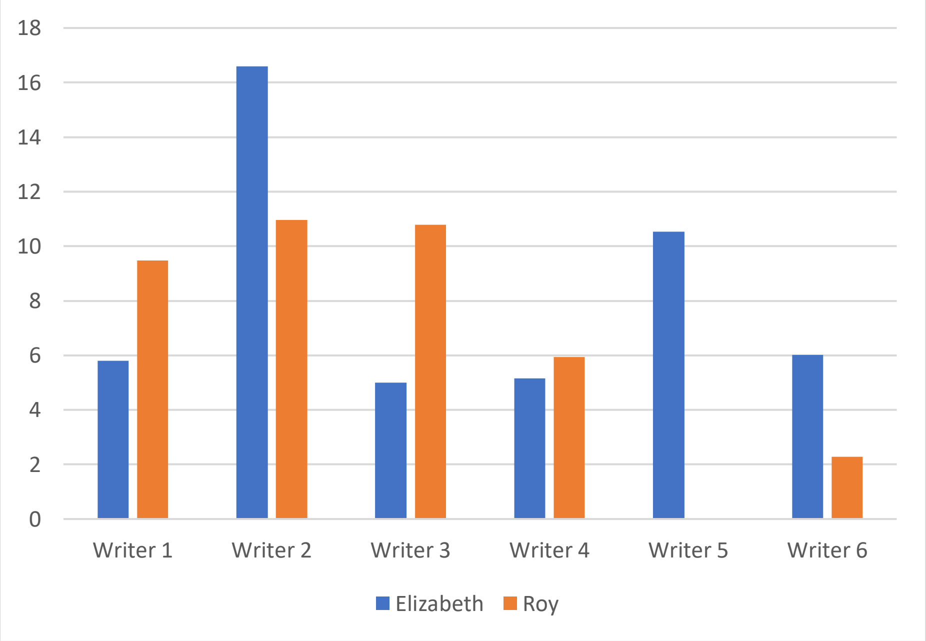


Figure 12: Elizabeth and Roy, usage of *oh*, normalised per 1000 words

These are both smaller corpora than Helen and Rob, so the quantitative results are less stable. For example, Writer 5's corpus for Elizabeth and Roy was only 613 words, compared to 5663 words for the Helen and Rob corpus. Writer 5 only had 3 occurrences for *oh* for Elizabeth and zero for Roy, so these frequencies are too low for a statistical analysis. In the combined corpus, a structural level analysis of *oh* discriminates between certain pairs of authors: Writer 1 and Writer 6, and then Writer 2 and Writer 6. Without the presence of Writer 6 in the corpus, it would be difficult to discriminate between any pairs of writers with any confidence, based on the frequency of *oh*. In the Helen and Rob corpus, a frequency count of *oh* shows differences between certain pairs of authors: Writers 1 and 5, Writers 1 and 6, and for Helen's lines only, between Writers 4 and 5, and between Writers 4 and 6.

The results for the Helen and Rob corpus suggest that five writers (with the exception of Writer 6) have used the pragmatic marker with a different frequency, depending on which character they are writing, using *oh* more frequently for Helen than Rob. For five writers (except Writer 6) there is some consistency of characterisation: Writer 6 seems not to use *oh* heavily as a linguistic feature, and also does not use it to discriminate between the characters being written. The difference in frequency of *oh* also suggests it would be possible to distinguish between certain pairs of authors, even though they are all writing in the voices of other characters.

7.4 Pragmatic Results and Discussion

Moving from a structural level analysis, where *oh* is examined as a single, structural level word, to a pragmatic examination, I explore my final sub-question, asking: is it possible to discriminate between pairs of authors based on the pragmatic uses of *oh*, and if so, does this increase the number of authors who can be discriminated from each other? Each use of *oh* was coded, as outlined above (7.2) and counted. This was then converted to a normalised figure, per 1000 words. A normalised count was made for each writer, for each character, and for each type of usage of *oh*, for example, 'Writer 1, Helen, Agreement'.

When breaking down the frequency of *oh* by writer, then by character and then by function, the figures are too small for a formal statistical analysis, so a quantitative, descriptive approach was used instead. (See Appendix 5d for full results).

The three categories with the highest number of results were:

- Vocative / conventionalised phrase
- Acknowledgement
- Spokenness.

Vocative / conventionalised phrase

Firstly, the writers' use of *oh* in Helen's lines is considered, and then their use of *oh* in Rob's characterisation. When Writers 2, 3 and 4, and to a lesser extent, Writer 5, write Helen's lines, the use of the vocative / conventionalised phrase is relatively high, as shown in Figure 13.

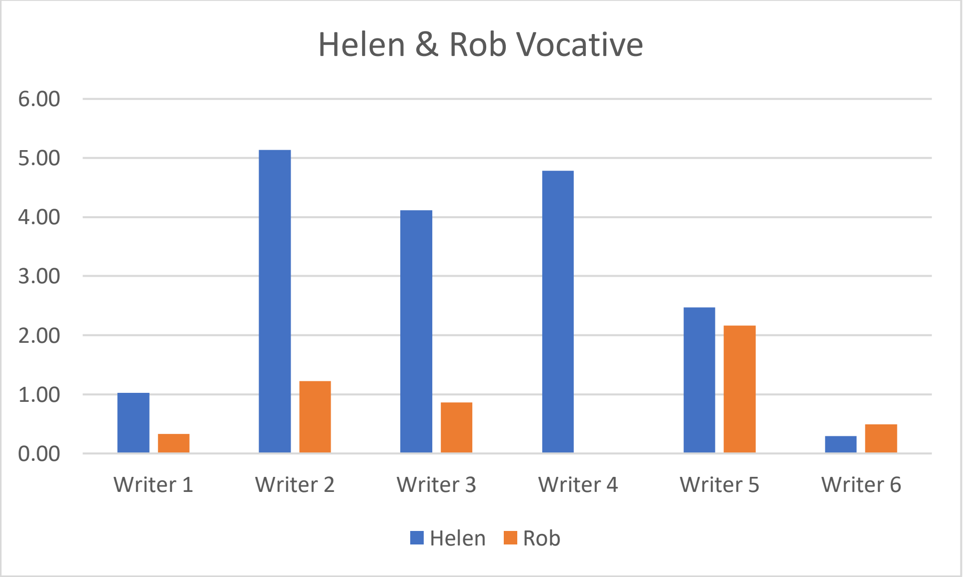


Figure 13: Vocative / Conventionalised Phrase for Helen and Rob

Writer 1 was one of the most frequent users of *oh*, but is a notably lower than average user for this pragmatic function, with a normalised frequency of 1.03 occurrences per 1000 words for Helen's lines, and even lower (0.33) for Rob's. This function only appears four times in Writer 1's 'Helen' corpus. When it does appear, there is a negative stance for three out of four occurrences, e.g. "oh dear".

In Helen's lines, the three higher uses of the Conventionalised phrase / vocative *oh* are Writers 2, 3 and 4. Interestingly, when comparing these three higher usage writers, there are differences in the way they use the vocative / conventionalised phrase. Writer 2's Helen corpus contains a mix of emotions, but there is a tendency towards a neutral or negative tone. Examples include:

> (EXASPERATED MUTTER) Oh, for heaven's sake...
>
> Oh, darling, I didn't mean to nag.
>
> Oh Rob, that's wonderful.
>
> Oh, my goodness, what am I going to cook?
>
> Oh dear.
>
> Oh no!
>
> (Writer 2)

Out of the 13 instances of *oh* being used in this way, only two are positive. Both examples are "Oh Rob", once as a complete turn (at the beginning of an interrupted sex scene), and "Oh Rob, that's wonderful", in response to some welcome family news. In 11 out of 13 cases there is a negative or apologetic tone to the usage. Writer 2's use of *oh* as vocative contrasts sharply with Writer 3's usage of *oh* in the vocative / conventionalised phrase, especially when considering the collocate "Oh Rob". Out of 16 examples of the vocative function, 12 uses of *oh* collocate with "Rob", and the tone is overwhelmingly positive. Examples include:

> (HAPPY SIGH) When you put it like that… oh Rob.
>
> Oh, Rob, you're so lovely to me.
>
> (Writer 3)

Writer 4 also has a high usage of the vocative / conventionalised phrase function for Helen, but unlike Writers 2 and 3, there is not a heavy use of *oh* plus addressee, and there is a more even split between positive and negative stance.

```
Oh, Rob!  Not even a little one?
```

(positive – Helen is asking for clues to Rob's surprise evening out)

```
Oh please, Rob. It is an emergency, after all.
```

(negative – tentatively asking a reluctant Rob to help with childcare)

```
Oh, thank you so much. You are an absolute star.
```

(positive)

(Writer 4)

The writers show differences in the way they use the vocative *oh* for Rob, compared to Helen. There is a strong differentiation between Writer 3's frequency of *oh* for Helen and the use of *oh* for Rob in this function: there are 16 for Helen compared to 4 for Rob. Interestingly, all of Rob's usage of the vocative function use exactly the same phrasing: "Oh Helen." The tone is less consistently positive than Writer 3's Helen lines. In fact, in all four cases there is a sense of reluctance or gentle reprimand:

```
HELEN     It's been the same for me. It's been awful

ROB       Oh Helen -


ROB       Oh, Helen, it's difficult. Mum was very
          close to Jess.


ROB       Oh, Helen, can't we just leave things as
          they are? Like I said, it'll only cause
          hassle.
```

HELEN I said. I didn’t want to spoil our evening out.

ROB Oh Helen…

(Writer 3)

There is a difference in characterisation by Writer 3, in terms of frequency difference of *oh* between the two characters, and the positive and negative stance, but there remains an intra-author consistency in terms of the high collocation frequency of *oh* plus name of addressee. Writer 4’s Rob lines do not use the vocative / conventionalised address form, so again, there is a contrast in the way the authors write Helen compared to Rob. Writer 5 also has a high occurrence of the collocation “oh Rob”, even though the overall usage of *oh* as a conventionalised phrase / vocative is almost half that of Writer 3. Despite the similarity in collocation of *oh* plus “Rob”, the pragmatic stance is far more negative and includes tones of gentle chiding, disagreement and worry, such as this extract where Rob and Helen are discussing Rob’s divorce from Jess:

ROB Jess has only just agreed. I haven’t had anything official from her solicitor yet.

HELEN Oh Rob! You don’t think she’ll go back on it.

(Writer 5, 2014)

The typical use of “Oh Rob” in Writer 4’s corpus is notably less effusive than in Writer 3’s. When Writer 5 writes Helen’s lines, there are five uses of “conventionalised phrase”: five are *oh* plus addressee, and two are the conventionalised phrase, “oh dear.” Four out of five of “oh—plus-addressee” are “oh darling”. The last instance is “Oh Henry”, spoken to a silent Henry, so every single instance of Rob addressing Helen preceded by an *oh*, is “oh darling”, showing some character consistency in the way Rob is written, which is different from the way the writer creates Helen.

On a quantitative level, the frequency of *oh* as vocative can distinguish between certain pairs of authors – the high users, Writers 3, 4 and 5, compared to the lower users – Writers 1, 2 and 6 (although Writer 6 is a low user of *oh* per se). Whilst the frequency of the function by character does not discriminate sharply between Writers 2, 3 and 4, a qualitative analysis can show differences between the ways the writers use them, including differences in pragmatic stance, and different collocation frequencies. If the criteria are further tightened, to include only *oh* plus an addressee (e.g. "Oh, Rob, and "Oh, darling", but not "oh my goodness"), then some of the pairings become even more marked (Table 36). Writer 3 and Writer 4 are even higher in their use of "oh plus addressee" as can be seen in the table below. For example, the difference between Writer 1 and Writer 3 becomes even more noticeable.

Table 36: Collocation of *oh* plus addressee for Helen

Writer	No. of vocative	Collocation of "oh plus addressee"	% of occurrences
Writer 1	1	4	25
Writer 2	5	13	38
Writer 3	12	16	75
Writer 4	4	13	31
Writer 5	4	6	67
Writer 6	0	1	0

As can be seen, Writer 3 and Writer 5 use *oh* preceding an addressee as a higher proportion of their vocative usage of *oh* than the other writers.

2. Agreement

The second category discussed is 'Agreement'. Overall, the numbers for agreement are much lower. The highest figure for any one writer for 'agreement' was four occurrences per corpus. One notable use is Roy's way of agreeing with statements. Four of the writers use "oh yeah" for Roy, but never for another character in the combined corpus. For example, Writer 1 uses "oh yes / yeah" a total of four times. Three out of the four instances are "oh yeah", rather than "oh yes." "Oh yeah" is not a form used by Writer 1 for any other character in the corpus. For Writer 6, three out of five occurrences of *oh* collocate with "oh yeah". Writer 3 also has one occurrence of "oh yeah" and none of "oh yes". (Neither Writer 2 or Writer 5 use "oh yeah" or "oh yes" for Roy). At a structural level, the writers have a collective habit of using "oh yeah" for Roy, which stands in contrast with a preference for Elizabeth – the widowed owner of a stately home – to use "oh yes". The shared pattern of using "oh yeah" suggests a way in which the writers are successfully synthesising their language at a structural level, compared to greater inter-author variation at a pragmatic level.

3. **Surprise**

With the more tightly defined criteria for surprise, the occurrence rates were too low for comparison.

4. Acknowledgement

Writers 1 and 4 have a notably heavy uses of acknowledgement for Helen's lines, shown in Figure 14, in particular single word turns consisting only of *oh*.

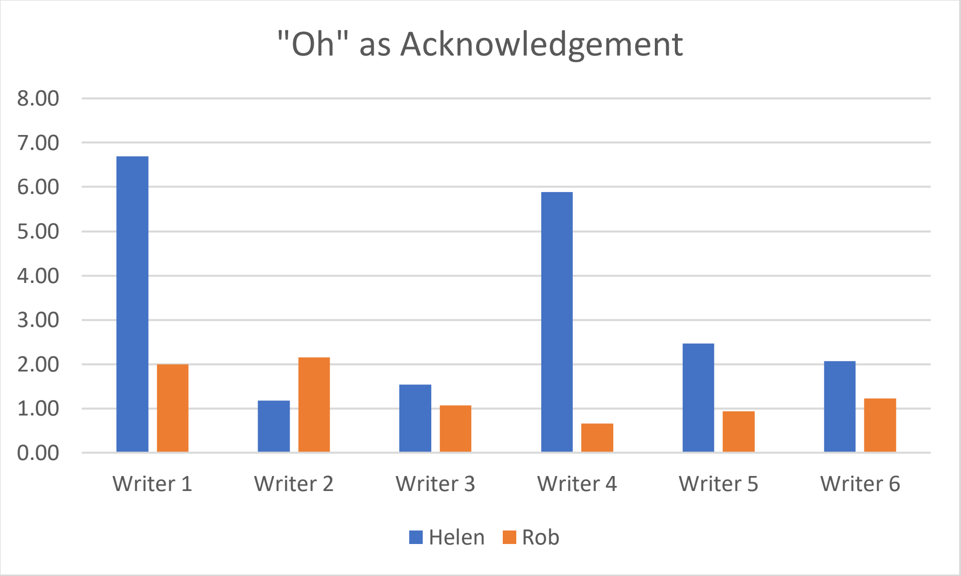


Figure 14: Helen and Rob, *oh* as Acknowledgement

Whilst a straightforward frequency count of *oh* would not distinguish between certain pairs, e.g. Writer 1 and Writer 2; Writer 3 and Writer 4, breaking the occurrences of *oh* into pragmatic function does allow us to discriminate between those pairs of authors. Writer 1 has a significantly higher use of *oh* as Acknowledgement compared to Writer 2, and Writer 4 has a significantly higher use than Writer 3. It is therefore possible that a two-step process, first considering overall frequencies, and secondly considering frequency of the pragmatic category "acknowledgement" could perhaps be used to discriminate between a greater number of pairs of authors.

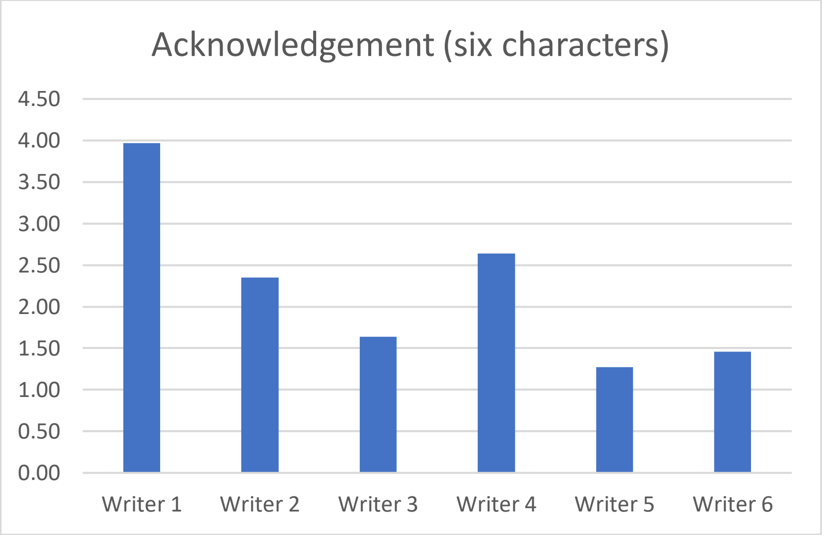


Figure 15: *oh* as Acknowledgement in combined corpus

As seen in Figure 15 above, Writer 1 uses *oh* as acknowledgement more frequently than the other writers, whether they are writing Helen's lines, or throughout their corpora, suggesting this is an area of identity construction which remains constant to the writer, regardless of which character's voice they are scripting. The use of "oh as acknowledgement" as a single word turn is also high. In Writer 1's Helen lines, 14 out of 26 uses of "oh as acknowledgement" have *oh* as a single-word turn. In Writer 1's Rob lines, this occurs in 4 out of 12 acknowledgement examples. In Lilian's lines this occurs 2 out of 4 times, in Paul's lines, 1 out of 6 times, in Elizabeth, it is 2 out of 3 times, and in Roy's lines this happens 2 out of 8 times.

Writer 4 is the second highest user of *oh* as Acknowledgement. In Writer 4's Helen lines, there are 16 occurrences of 'acknowledgment': 11 of these feature *oh* as a single-word turn. Rob only uses it three times, none of which is a single-word turn. At one level, this could show that Writer 4's use of pragmatic markers demonstrates a way the writer adapts linguistically to create Helen's dialogue compared to Rob's. From a Critical Discourse analysis, this could be partially explained by power relations: Helen was found to be the victim of coercive control in a relationship, and the high use of *oh* as a single turn could reflect this: receiving information and opinions but feeling unable to comment or evaluate them, or unable to follow up the information with questions or opinions of her own. However, when compared to Lilian's lines there is also a high use of *oh* as a single-word turn form of acknowledgement (3 out of 6 occurrences). Lilian is a confident, flamboyant character, and the same power differentials do not apply.

Another way to view this could be that *oh* as a single-word turn of acknowledgment is a feature associated with Writer 1 more than with any other writer, regardless of which character is speaking. In this sense, it shows their consistency as writers. This figure would be even higher if it included *oh* as a single sentence within a longer turn. For example, in Writer 1's Helen lines, 21 out of 24 cases of "acknowledgement" have *oh* as a single word sentence, often a single word turn. This is also high in the Lilian-Paul corpus and (to a lesser extent, the shorter Elizabeth-Roy corpus), regardless of which of the couple is speaking.

The tone of "oh as Acknowledgement" is often quite negative, as in this example:

```
ROB  No... er... (SIMPLE) Jess is coming up this
weekend.

HELEN     (TRYING TO HIDE DISAPPOINTMENT) Oh.
```

(Writer 1, 2013)

There is often a sense of disappointment, or receiving unwelcome news or opinions. This is true of Rob's lines as well as Helen's, as in the example below:

```
HELEN     But - until we know where we are with Dad.
```

ROB Oh.

(Writer 1, 2014)

In one instance, Helen agrees with Rob, so at the character-to character level (Culpeper’s “first order” pragmatics) this should seem quite a positive use of *oh*.

[ROB (UNLOCKS SEAT-BELT) You might want to straighten your hair before we go in.

HELEN Oh. Yes. (CHECKS HAIR IN SUN-SHIELD MIRROR)] (Writer 1, 2015)

However, at a discourse level, and from the audience perspective, there is a negative sense to this. Rob’s control over Helen extends to telling her how to dress and style her hair, so even if Helen is agreeing in the moment, the audience response (Culpeper’s “second-order” pragmatics) is likely to be a more negative interpretation.

Comparing the tone of *oh* as acknowledgment as used by Writer 3 compared to use by Writer 1 and Writer 4, there seems to be a difference. Even where Writer 3 is using *oh* as a brief statement of acknowledgment, there is often a more positive tone, as in these examples:

ROB We’ve had a row. A big row.

HELEN Oh. I see.

(Writer 3, 2013)

At this stage, Rob is still married to Jess, so whilst the news of a row might ordinarily elicit a sympathetic response, from a self-interested perspective, this could be interpreted as welcome news by Helen.

HELEN Yes. (BEAT) So – the weekend. Are you – are you going to Hampshire?

ROB I’m not sure. I haven’t decided yet.

```
HELEN     Oh! I see.
```

(Writer 3, 2013)

When Writer 3 writes Rob's lines, there is a similarly positive stance to *oh* as acknowledgment, for example, "oh, brilliant" and "Oh. Very nice." This echoes the findings for Writer 3's use of *oh* as vocative / conventionalised phrase, where the pragmatic stance was often a positive one.

Writer 6's use of *oh* as a single word turn most frequently occurred in response to something physical, rather than information as in these two examples:

```
HELEN     Okay. (KNOCK AT THE DOOR)

ROB       Oh –
```

(Writer 6, 2013)

```
ROB       Just  wait!  (TURNS  ON  STEREO.  'IS  THIS
          LOVE'  BY  CORINNE  BAILEY  RAE  PLAYS,  LOW,
          FOLLOWED IF NEEDED BY 'ALL OF ME' BY JOHN
          LEGEND)

HELEN     Oh –
```

(Writer 6, 2015)

The response is to a physical change, rather than receipt of information. Although the number of occurrences is low, when *oh* as a single-word turn does occur, it tends not to be a response to information.

5. Spokenness / casualness

Writer 5 and Writer 6 were lower in their usage but Writers 1, 2, 3 and 4 were all broadly similar (Figure 16).

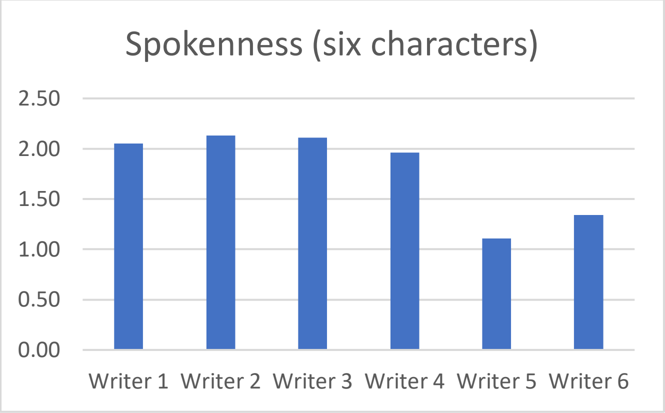


Figure 16: *oh* as Spokenness

Pragmatically, the use by the different writers does not seem to show any significant features which would allow us to discriminate. As a category, it seems to have a secondary purpose, in that it removes a large number of *oh* occurrences into a contained category, which then allows a closer pragmatic analysis of "vocative / conventionalised phrase" and "acknowledgement".

6. Continuers / topicalisers

These occur infrequently and used similarly for all writers and characters where they do occur.

7.5 Conclusion

My aim was to explore whether higher-level domains of language were better able to discriminate between authors. A straightforward frequency count of the occurrences of *oh* showed some inter-author variation, with Writer 1 and Writer 4 using *oh* almost twice as frequently as Writer 6. Analysing patterns of usage by character, the results showed that *oh* was a feature which writers used in different frequencies for Helen than for Rob, with Writers 1-5 having a higher usage for Helen than for Rob. For Writer 3, the normalised results showed that *oh* was used by Helen on average 10.02 times per 1000 words, but only 3.21 per 1000 by Rob. Similarly, for Writer 4, *oh* was used by Helen on average 13.98 times per 1000 words, but only 2.42 per 1000 by Rob. At this structural level of language, the writers were successfully able to synthesise the way they adapted the frequency of *oh* for Helen, compared to Rob.

In the pragmatic analysis, sub-dividing the occurrences of *oh* by function showed inter-author variation: Writer 1 has a significantly higher use of *oh* as Acknowledgement compared to Writer 2, and Writer 4 has a significantly higher use than Writer 3. It is therefore possible that a two-step process, first considering overall frequencies, and secondly considering frequency of the pragmatic categories could perhaps be used to discriminate between more pairs of authors.

The pragmatic interpretations were also interesting in terms of authorial suppression, such as Writer 1's strong tendency to use *oh* as acknowledgement, regardless of whether the lines are spoken by Helen or by Rob. Using Grant and MacLeod's terms of persistence features, and linguistic leakage, this seems to show that certain writers have certain persistent linguistic traits in their functions for which they use *oh,* because the pragmatic function is consistent by author, regardless of which character is speaking.

My superordinate research question asked to what extent dramatists are able to create linguistically distinctive characters, and maintain the consistency of those characters' style, whilst simultaneously suppressing their own authorial style. From these results, it seems that writers are simultaneously able to create linguistically distinctive characters at a structural level, as indicated by the variation in the use of *oh* between Helen and Rob, but at the pragmatic level, there is evidence of a persistent authorial style which has not been suppressed. This suggests that the writers are better able to imitate each other at a structural level of language, because fewer pairs of authors could be distinguished from each other, and less able to suppress their own authorial style at a pragmatic level, because patterns of intra-authorial consistency were discernible.

8. Conclusion

8.1 Introduction

This chapter reflects on the findings from my three main studies, and considers how they might be applied to forensic enquiries. I begin by re-stating my research aims and summarise the three main studies carried out. The contribution to research is evaluated, followed by discussions of conscious identity performance, and the issues surrounding the use of fictional characters to explore linguistic identity disguise. Some future directions are indicated, before I state my final conclusions.

My superordinate research aim was to investigate the extent to which dramatists are able to create linguistically distinctive characters and maintain the consistency of those characters' styles, whilst simultaneously suppressing their own authorial style. I defined *character style* as linguistic traits which could be identified with a particular character, but acknowledged the issues with using this term, because of the embedded discourse structure of drama (Short, 1989:149), which means that character style logically cannot be completely separate from authorial style. I defined *authorial style* as the linguistic traits and patterns which were found to be consistently used by an author regardless of which character's voice they were writing.

Each of the three main analyses addressed one of my research sub-questions in turn. These sub-questions were:

1) To what extent do quantitative, structural-level analyses identify character style rather than authorial style?

2) Are writers able to identify consistent intra-character features?

3) Can higher-level pragmatic features provide a base for authorship analysis in cases of linguistic identity disguise?

The next section summarises my three studies, before discussing the main findings in relation to my research aims.

8.2 Summary of Analyses

The first study (Chapter 5) was a quantitative exploration of the data, to look for patterns of variation and consistency by writer and by character. I carried out a word-n-gram-based study to identify bi-grams and tri-grams which were used by only one or two writers, and might therefore be potentially indicative of individual authorial style. Selected n-grams were investigated using concordance plots, to analyse the distribution of these n-grams throughout the character corpora. The purpose of the word-n-gram-based study was to produce explicable results which would track the usage of particular n-grams by character, to analyse whether an n-gram was distributed randomly across characters (and therefore more indicative of authorial style) or if the n-grams were only used by certain characters (and therefore more indicative of the writer modifying their language to create a specific character's style). This was followed by three commonly-used stylometric tests: these were Average Word Length, Average Turn Length and Type-token ratio, and were carried out on each character in the 20-character corpus to look for patterns of inter-author and inter-character variation and consistency. The purpose was to analyse whether writers modified their style for different characters in ways that affected the results of these commonly-used stylometric tests.

The second study (Chapter 6) addressed the second of my sub-questions. It focused on the distinctive linguistic traits of three larger-than-life characters, by analysing Jim's lexis, Jazzer's dialect and the (im)politeness strategies of Lynda Snell. The aim of these three qualitative studies was to investigate how closely the writers were able synthesise character style for these three characters.

My third, and final study (Chapter 7) carried out a pragmatic analysis to explore whether pairs of authors could be distinguished from each other, not just by the frequency of a token, *oh*, but by the function of that token, to examine whether the analysis of higher-level domains of language made it possible to distinguish between more pairs of authors than a structural-level analysis which considered only the frequency of occurrence. The main findings from all three studies are discussed in turn in 8.3, and then in combination in 8.4.

8.3 Conclusion from Three Studies

The word-based-n-gram test was intended to show the distribution across characters of authorially individuating n-grams, to examine whether those n-grams were being used for specific characters, or if the distribution was spread evenly across characters. Although I began with a moderately large dataset, the All-Character corpus was divided by writer and then by character, and standardised to match the lowest word counts of the corpora used, meaning each individual character corpus only contained 2804 words. As a result, the raw frequencies produced were too low to evaluate n-gram distribution by character. There were suggestions of authors using particular n-grams for certain characters, such as Writer 6's use of "yeah yeah" for Pip, and Writer 4's use of "yeah yeah" for Tom. There were also suggestions of writers using individuating n-grams across a number of characters, such as Writer 1's "one or two", which suggested a failure to suppress authorial style when writing individual characters. These figures, however, were too low to be anything more than tentative suggestions.

The remaining three stylometric tests found that the authors were able to alter their writing style in ways that produced observable, quantitative variation between different characters written by the same author. Notably, all six writers had higher average word lengths and higher average turn lengths for three characters (Lynda, Jim, and to a slightly lesser extent, Brian). The writers also varied their linguistic style for other characters, such as Helen and Pip, who tended to have lower figures for average word length and average turn length. Of the two results, the six writers were more closely matched in average word length than average turn length. An issue with this method was the lack of explicability: whilst Pip's shorter word and turn length could be explained by her fraught relationship with her parents, there was no literary interpretation why Helen was written with lower average sentence length and turn length by multiple writers. On the whole, the writers followed a shared pattern of increases and decreases for the twenty different characters, suggesting some convergence in the way they create the characters' dialogue. In turn, this convergence suggests a level of intention about the way the writers adapt their voices for each character.

For certain characters, such as Brenda, the average word length for all six writers was less closely matched. This suggests that some characters are linguistically less unifying than others, when authors attempt character-level synthesis. It is possible that the characters whose stylometric results did not follow any shared pattern among the writers were synthesised in other ways which did not affect the results of stylometric tests. Reflecting on dualist concepts of style (discussed in Section 2.4.2), it could be argued that for certain characters, the character synthesis was carried out by modifying the *content* of the character, rather than the linguistic style, or, at least, these three measurable elements of linguistic style. Examples of this from the data are Brenda's careers ambitions, Jennifer's snobbishness and Eddie's get-rich-quick schemes, all of which are consistent character traits as written by all six writers, but are traits which may manifest themselves in ways that do not have an effect on stylometric results.

Chapter 6, "Three Character Studies", analysed imitation of distinctive features of characterisation. In stylistics and in literary criticism, characterisation is often discussed with reference to plot, and to characters' emotional and cognitive trajectories. In Chapter 5, my focus was on linguistic features, both structural level and discursive features, rather than a focus on the characters' actions, decisions or emotional states. One reason for this is because many decisions regarding plot are made by the production team, rather than by the individual writer, so could not be interpreted as stylistic choice. A second reason was to focus on the character-to-character level of discourse, to find results which would be applicable to forensic investigations of linguistic identity disguise, rather than the playwright-to-audience level of discourse explored in literary stylistics.

Analysing linguistic features of characterisation found that writers were adept at altering their lexical choices to create the more distinctive characters, such as Jim and Lynda. For the characterisation of Jim, all six writers used a consistently high number of Latinate words, evoking a formal style. A weakness in this approach is that the Latinate analysis seemed better at discriminating between characters who were at either extreme, and was not sophisticated enough to discriminate between characters who were not outliers. The process required a two-step approach of selecting Latinate words and then subjectively evaluating individual words for levels of formality, which then raises questions of methodological rigour and replicability. Whilst the process was able to show how the writers successfully adapted their lexical choices to create distinctive characters, it was not a process which identified authorship in these data.

The keyword analysis was less illuminating about characterisation or authorship, perhaps because of the size of the data: words were identified as keywords on the basis of only two or three occurrences. Examining those keywords in context, it was often found that the keyword was not suggestive of any particular character traits: for example, the word “fletch” appeared as a keyword, but was explained by the presence of a character called Fletch who featured in a single episode with Jim and not with any other characters. Further, the non-visual nature of the medium means that proper names often appear as keywords because characters tend to address each other by name more frequently than in other forms of drama, to inform the audience which characters are present. This revealed information about genre, rather than authorial style. Some of the keywords, such as *ah* and *indeed* were suggestive of the type of conversation Jim typically has: exchanges of views, and calm, measured responses, rather than more emotional responses such as “oh my goodness” which other characters use. As with the Latinate analysis, the results did not show differences in individual authorial style, with the exception of a high usage of *ah* by Writer 1. More frequently, keywords could be explained by context or by decisions that were made by the production team rather than individual writers.

Analysing (im)politeness strategies showed some similarities in the ways that writers had Lynda complimenting and coaching cast members during rehearsals. There were differences in the levels of reassurance, and the way the reassurance occurred within Lynda’s directives. For two writers, the reassurance or compliment tended to be placed at the start of a line, for example, “That’s it. …” and “Marvellous…”. However, these patterns were only tentative. The main issue with trying to discern authorship from (im)politeness strategies was that no two scenes could ever be the same. For example, in one scene, Lynda is directing rehearsals, but the dramatic interest of the scene is predominantly about a disagreement between two characters, Kirsty and Rob, so the tension between Kirsty and Rob is foregrounded, not Lynda’s directing. In a separate rehearsal scene, the audience’s perspective is with Lynda, rather than the cast members, because we hear her private "sotto" asides in between her instructions to the cast. Any differences between the way that Lynda is portrayed by the two different scriptwriters could be explained by a number of reasons: whether the storyline is supposed to show rehearsals going well, or rehearsals going badly; whether the core of the drama in the scene is about the show's success, or whether the panto rehearsals are merely a backdrop or dramatic device to progress a different storyline.

The character study which seemed to show the most promising distinctions between the six writers was the study of Jazzer's dialect. Analysing dialect at the levels of grammar, lexis and pronunciation (manifested as eye-dialect in the scripts) showed that there were observable differences between the ways that the six different writers used dialect. One of the writers was notably higher in the number of lexical items of dialect included. Writer 2, who was overall lower in the use of dialect included some grammatical level features of dialect, most notably, ellipsis, which seemed to create a subtler evocation of a Scottish dialect. The study of dialect showed differences between certain pairings of writers, and it would be interesting to apply the same methodology to other characters; for example, the perennially disadvantaged Grundy family, who often use non-standard English grammar, such as "you was going to do that." One issue with applying this study to forensic settings is that dialect is a feature which would likely be presented very differently in text-types other than drama scripts: in drama, writers are often giving a flavour of a character's dialect, rather than attempting a fully authentic representation of dialect, as might be the requirement in a forensic authorship synthesis task.

There were occasions when one might have expected Jazzer's lines to be written with a higher number of dialect items, particularly in the moments when Jazzer, as a character, is indexing his own Scottishness. In these moments it was possible to detect the linguistic leakage of a writer showing features of "English English" where Scottish English might be expected instead. Having analysed dialect at the levels of grammar, lexis and pronunciation, it would be interesting to examine the evocation of dialect at a more discursive and pragmatic level, and to link it to characterisation. This, however, raises questions about drawing on stereotypes and archetypes to perform identity disguise.

The study of *oh* (Chapter 7) found observable differences between the writers. At a structural level of linguistic analysis, it could be seen that some writers used *oh* more than others. However, authorial discrimination became more noticeable at the pragmatic level, when writers used *oh* for differing functions, and with differing evaluative stances. In this analysis, linguistic leakage could be detected: writers were able to vary the frequency of *oh* for different characters, but retained intra-author consistency in the function for which it was used. It stands to reason that pragmatic noise is a promising feature to choose to analyse pragmatic and semantic differences because the non-standard nature of the words, and their spellings, make them more open to individual interpretation and usage. Many forms of computer-mediated discourse retain these features of spoken English in the text, often to evoke a sense of informal, spontaneous conversation, so it would be interesting to apply this methodology to data such as tweets and other social media posts to analyse variation in pragmatic noise at a structural-level (spelling, frequency, placing), but also at a pragmatic level, studying variation in its function within a text.

It was interesting to observe that Writer 1, who had a consistently high use of *oh* in the pragmatics study, also had a high use of *aye* in the study of Jazzer's dialect, and *ah* in the study of Jim's lexis, suggesting a consistent pattern of beginning lines of dialogue with a form of acknowledgement, regardless of which character is speaking. A word frequency analysis would not necessarily identify Writer 1's tendency to begin lines with a form of acknowledgement, because this function is realised by different tokens. At the structural level, the writer uses *oh* for Rob (among other characters), *ah* for Jim and *aye* for Jazzer, showing an ability to modify lexical choice in the creation of individual character styles. However, at the pragmatic level, there is an intra-author consistency of beginning lines of dialogue with a form of acknowledgement. This intra-author consistency at the pragmatic level co-occurs with intra-character consistencies at the structural level.

8.4 Combined Observations

This section now considers the results in combination. From my results, it is possible to see that authors can and do modify their character style to create different linguistic identities, and that these modifications are such that a stylometric, structural level analysis can detect these differences, reinforcing findings in Adversarial Stylometry research (2.5.3) about the vulnerability of standard stylometric tests in cases of linguistic identity disguise.

Two analyses used pragmatics frameworks: Lynda's (im)politeness, and the study of *oh.* These studies showed different levels of promise in their ability to discern authorship. The linguistic features of Lynda's (im)politeness were heavily influenced by context, by audience, and by the needs of plot and character development, and did not discriminate between authors with any certainty. In contrast, the pragmatics study of the functions of *oh* was more revealing. A possible reason why the study of *oh* was better able to determine authorship than the (im)politeness study is because it uses a specific, structural-level feature to study a pragmatic function. *Oh* is a discrete feature, which was easy to identify, extract and compare, whereas the study of Lynda's (im)politeness examined a broad collection of linguistic strategies which were extremely sensitive to context.

The combined results show that there are certain characters whose distinctive characterisation is manifested linguistically in features such as formal language, whereas other characters are less linguistically distinctive. These other characters may well be considered idiosyncratic through their opinions or behaviour, rather than their linguistic style: for instance, being serially unfaithful, the village gossip, or a ruthless business owner. It seems that those characters who were linguistically unifying and produced similarly marked results in basic stylometric techniques were those who are more heightened in their characterisation, compared to the more middle-ground characters. In terms of character synthesis, it seems that having a 'target' who is linguistically distinctive makes imitation more successful, in these data.

8.5 Contribution to Research

This thesis has contributed to research in Forensic Linguistics by taking the idea of playwriting as identity disguise and using it as a proxy for forensic issues of anonymity and linguistic imitation. It has shown that when authors deliberately alter their writing style to create different characters, it has an observable effect on stylometric features commonly used in authorship analysis, such as word length and sentence/turn length, bringing those features into question as reliable style markers in cases where writers have deliberately modified their style to attempt identity disguise. It has shown that writers are able to adapt their style to the extent that these basic stylometric results are altered, but that some authorial traces remain: for example, writers with a lower average TTR for one character tended to have a relatively low TTR across all their characters. The most promising contribution is probably the analysis of pragmatic functions, because it used a replicable methodology which showed that features of authorial style could be identified when the writers used the same lexical token to perform different functions. This could be extended to other discourse markers, and other linguistic features.

Grant and MacLeod (2016) found that impersonators could be trained to improve their identity performance at different linguistic levels. While the writers are not linguistically trained, they are selected for their skill and experience, and these studies show that often, on an instinctive level, they are able to achieve closely matched identity performances of characters, in particular through lexical choice and politeness strategies, without necessarily having any specific knowledge of linguistic analysis to do so. Improving our understanding of which linguistic domains authors are more successfully able to imitate could be helpful in tasks when analysts are trying to identify deceptive linguistic identity performances.

The study of Jazzer's dialect has shown that when writers use dialect, they are able to maintain some consistency over a long (seven year) period of time, in particular through content word choice. However, there are cases where a writer's use of dialect is consistently variable: for instance, the writers who used the Scottish enclitic endings of “disnae” and “wasnae” and so on, did not use these endings exclusively, and often reverted to the English English forms of “doesn't” and “wasn't”. Also, some writers fluctuated between differently accented endings, for example, “canna” and “cannae”. The writers were arguably indexing Scottishness for dramatic purposes, not necessarily trying to fully realise a Glaswegian dialect, which is a limitation of using fictional data to explore forensic identity disguise. Even so, measuring the presence of dialect items through an analysis of grammar, lexis and pronunciation was a useful way of examining how consistently and how convincingly authors were able to perform a sociolinguistic identity that did not match their own linguistic resources. The analysis was able to identify instances of linguistic leakage, and also weaknesses of identity assumption if a character did not draw on the sociolinguistic resources they might be expected to possess.

8.6 Methodological Questions

My results suggest that pragmatics is a promising area of research for identity disguise. It might be that analysing a larger proportion of the available data, rather than only analysing scenes with a shared context, would have yielded more significant results. Arguably it was a methodological weakness to select only scenes where Lynda was directing plays. The aim was to reduce contextual variation to allow a clearer focus on variations in (im)politeness strategies, but even when the situation was overtly very similar, dramatically and tonally, there were so many differences that a comparison was problematic.

Studying drama scripts has allowed an analysis of a much larger corpus of identity disguise than would have been possible using data such as online chatrooms, where the available data consisting of cases of known identity disguise are more limited. However, the embedded discourse structure of drama influences the text’s production, and the authorship analysis results, so the application of findings from my thesis to forensic settings requires caution.

8.7 Conscious Identity Performance

The overarching aim of my thesis is to explore the extent to which writers suppress their authorial style when they are writing in the voices of individual characters. A related question is the extent to which writers are conscious of the linguistic means they use to imitate others' linguistic styles and to suppress features of their own authorial style. One possible reason why pragmatic noise is a promising way to distinguish between authors arises from Brinton's description of pragmatic markers as "semantically empty", which could suggest they are used less consciously by writers, whereas those linguistic features which are more obviously important to an utterance's referential meaning, and to characterisation, receive a greater level of focus. As Burrows wrote in his study of Jane Austen's grammatical words, "It is a truth not generally acknowledged that, in most discussions of works of English fiction, we proceed as if a third, two-fifths, a half of our material were not really *there"* (1987:1). Being high frequency but receiving less focus, could mean that writers are less aware of these words when writing in character: the immunity to conscious manipulation that McMenamin advocates.

MacLeod (2020) also discusses those aspects of language which speakers often pay little attention to, and ignore. She details the linguistic training given to undercover officers to prepare them for authorship synthesis and argues that officers can be trained to recognise previously unnoticed aspects of their language. Writers in collaborative projects such as multi-authored dramas are not routinely given linguistic training, and nobody would suggest that scriptwriters require a statistical level knowledge of sociolinguistic features. However, part of the process of creative writing relies on observation, and as such, those features which are above conscious manipulation, may be deliberately altered to create varying characterisations. McMenamin refers to language choices as both conscious and unconscious (2002:26) but highlights the problem that:

> Not enough is known about the composition to establish precisely what in writing is conscious or unconscious. The reasons for this are the difficulties associated with such studies, i.e., that every writer's level of conscious choice of forms in writing is different, and that writers demonstrate varying levels of consciousness in language production, e.g., unconscious, subconscious, semiconscious and conscious. (2002:169)

It would be an oversimplification to suggest that conscious versus unconscious use of style are binary opposites, or even on a continuum. Bucholtz and Hall describe how the situation is more nuanced:

> Any given construction of identity may be in part deliberate and intentional, in part habitual and hence often less than fully conscious, in part an outcome of interactional negotiation and contestation, in part an outcome of others' perceptions and representations, and in part an

> effect of larger ideological processes and material structures that may become relevant to interaction. It is therefore constantly shifting both as interaction unfolds and across discourse contexts. (2005:606)

Reflecting on some of the influences that Bucholtz and Hall list, and applying them to drama, it could be argued that the scriptwriters are conscious of the effect they are producing, without any focus on the linguistic theory underpinning these creative decisions. Sometimes this can result in very similar approaches, as in some of Lynda's relational work in her directing scenes. It is also possible that some characters foreground certain linguistic domains: for example, if Lynda's position as the village's self-appointed organiser means the writers are more aware of the way she makes requests and demands, they are able to carry out an effective character-level synthesis and mirror each other closely. With another character, where (im)politeness strategies are less foregrounded, the synthesis may be less successful in this particular domain. This in turn creates methodological questions, discussed in the Literature Review, about whether a forensic linguist should use a pre-selected set of style markers, or should choose analytical frameworks in response to the content of the text.

8.8 Creating characters: schemata and stereotypes

The creation of characters could be considered in terms of Grant and MacLeod's resources and constraints model, where the characters' linguistic histories provide a resource of how they should speak, but also a constraint, in that it might break the suspended disbelief if a character suddenly speaks in a way that is perceived to be 'out of character'. One of the layers of influence that Bucholtz and Hall describe could also be applied to drama: the layer of "larger ideological processes and material structures". In a drama, this could be seen as the influence of genre on characterisation: the fictional and dramatic nature of the text also has a bearing on the linguistic style of the characters.

Schema theory is also important in fictional identity disguise. McIntyre and Bousfield discuss the concept of a model person, and its importance in readers' analyses of fiction:

> The concept of a model person, or 'cardboard cut out' (Brown and Levinson 1987: 58-9), is a concept which is helpful in the analysis of fiction. All readers draw on schematic expectations of character types (arising from societal experience, including reading literature) which we

> rely on when building a mental image of fictional characters (see Culpeper 2001: section 2.3)." (2017)

Schematic knowledge is an important resource and constraint for writers, because it can help to create different characters, and can allow us to use shortcuts to convey a character. However, it can be a constraint if the character is reduced to the level of a stereotype. The analysis of Jim's vocabulary showed that his characterisation seemed more interesting, and less reliant on a stereotype when he used linguistic code-switching, and combined some slang words with his more formal, 'professorial' speech. In a long-running drama, characters can be established over time, with unfolding personalities (as discussed in 2.7.2). Within the bounds of the core personality of a character, writers are able to stretch the identity of that character. For example, Writer 1 and Writer 4 both push the identity of Jazzer as a hard-drinking, womaniser to a greater extent than the other four writers. There is an expectation that characters will behave in consistent ways. If characters deviate too much from their known persona, especially in a format with no visual aids for the audience, it can cause confusion about who is speaking, and can make characters seem less credible.

In forensic contexts, issues of credibility do arise in identity performances by undercover officers, which suggests a shared requirement of conveying credibility for the officer as well as the dramatist. Chiang and Grant (2019), among others, discuss the presence of Perverted Justice decoys on chatrooms where paedophiles communicate. The fora are low-trust environments, where offenders have reasons to doubt that their interactants are who they claim to be. Further, the Undercover Officers have goals to apprehend suspected abusers: just as the fictional data are driven by the plot points of ongoing storylines, so too do undercover officers aim to achieve some progress as their interaction unfolds, which suggests elements of similarity in the process of authorship synthesis.

8.9 Future Directions

Having reviewed and discussed the main findings of my thesis, I now consider possible areas for future research. One possible area is to extend the analysis on Jazzer's Glaswegian dialect to consider all non-standard portrayals of grammar and lexis, in order to analyse how consistently writers can maintain a 'non-standard' dialect. This could have useful applications in forensic cases of identity disguise, where people online are conversing informally, using slang and in-group markers, and an undercover officer might be required to impersonate someone who is indexing a particular sociolinguistic identity.

A further area of research is to extend the analysis of the functions of *oh* by analysing more items of pragmatic noise. Another possibility is to replicate the two-step process used in the study of *oh* by selecting specific structural-level features, for example pronouns, or imperatives, and analysing them, firstly for frequency, then secondly through a pragmatics framework. This could be a method for analysing politeness, or other interactional phenomena. This process would provide a replicable methodology, and could produce explicable results, because it should be possible to track and compare variation in pragmatic function when the analysis is pinned to a quantifiable structural-level feature. However, this approach would be more time-consuming on a large dataset as the pragmatic coding cannot be automated.

Another possible avenue of research is to build on the analysis which looks at what *could* have been said. Whilst the predominant preference in linguistics has been for naturally-occurring data, it is still possible to glean information about the success of an identity performance by also considering what a writer could have done or chose not to do. For example, there were cases in the dialect study, where a writer used just enough Scottish dialect to index the character of a Glaswegian, but might have been expected to use more. Considering what *could* or *might* have been written is more speculative than methodologies commonly used in forensic linguistic enquiry, but it is very much part of the discussion in creative writing tasks and for drama practitioners. As such, it has possibilities to be utilised when discussing linguistic identity performance, because even in forensic settings, the task of authorship synthesis is inherently a creative one.

This study has shown that identity performance takes place at a number of different linguistic levels, from structural-level changes, such as increases and decreases in average word length, up to higher-level language use, such as the level of implicature in (im)politeness strategies. It has shown that writers are more able to alter structural-level features, such as the choice of content words, which in turn influences the results of commonly-used stylometric markers, in particular average word length and average turn length. In contrast, writers are less successful at anonymising higher-level features, such as the pragmatic functions of words, and the semantic force of words. Even so, there were some areas of close imitation, such as Lynda's (im)politeness strategies, suggesting that even without linguistic training, writers were able to match certain pragmatic elements.

8.10 Concluding Remarks

What is clear from the analysis, is that there is no 'one size fits all' method of attributing authorship in cases of linguistic identity disguise. The distinguishing features vary between characters and between pairs of authors: what may distinguish between one pair of authors will not work for all pairs of authors, and may only work for some characters. For example, the writers all wrote Jim and Lynda's dialogue consistently, but had other characters whose stylometric results were far more disparate. Considering all possible levels of language improves the likelihood of identifying authorship, because the study has shown that at higher levels of language analysis, the scriptwriters showed more intra-author consistency.

There are methodological difficulties in attempting to code higher levels of language, as found in the analysis of Lynda Snell's (im)politeness strategies, because each scene was heavily context dependent, and often influenced by the requirements of the plot, although a possible methodological approach is suggested in 8.9. Grant and MacLeod discuss the issues with analysing language beyond the visible, structural features on the page:

> Such a pragmatic focus may be inherently more difficult to analyse consistently, since coding 'involves an interpretive, subjective component' (Herring, 2004:18) but we would argue that it captures an essential element of identity performance. (2018)

In the pragmatic coding of *oh* it was necessary to change the categories of coding numerous times in order to achieve a desirable level of accuracy to ensure replicability. This raises methodological issues and practical issues, as it becomes time-consuming, which limits the amount of data that may be analysed. It is also entirely likely that a different dataset would require at the very least tweaks to the coding, if not an entirely new way of categorising the pragmatic codes. This could be a particular issue in forensic settings, such as authorship synthesis tasks, where an undercover officer may only have a very limited period of time to research a persona before having to perform that identity.

Overall, my findings show that certain elements of successful character-level stylistic imitation can be identified using structural-level analyses, such as studying lexical choice. It may be that this is linked to the level of attention consciously paid to features such as lexis, although providing evidence to support this claim is problematic, as McMenamin (2020) notes. My findings also show that analysing higher-level domains of language can successfully identify persistent elements of individual authorial style.

9. References

Afroz, S. et al. (2012) Detecting hoaxes, frauds and deception in writing style' IEEE Symposium on Security and Privacy, San Francisco, 20-23 May, pp.461-475.

Agha, A. (2003) The social life of cultural value. Language & Communication, 23(3-4), pp.231-273.

Agha, A. (2006) Language and social relations. Cambridge: Cambridge University Press.

Aijmer, K. (1987) 'Oh and Ah in English conversation', in Meijs, W. (ed.) Corpus Linguistics and Beyond. Amsterdam: John Benjamins, pp.61–86.

Aijmer, K. (2002) *English discourse particles: evidence from a corpus*. Amsterdam: John Benjamins Publishing. Vol. 10.

Aitken, A.J. (1984) 'Scottish accents and dialects', in Trudgill, P (ed.) *Language in the British Isles*. Cambridge: Cambridge University Press, pp.94-114.

Ameka, F. (1992) 'Interjections: the universal yet neglected part of speech', *Journal of Pragmatics*, 18(2-3), pp.101-118.

Ameka, F. (2006) 'Interjections', in: Östman J.-O., and Verschueren J. (eds.) *Handbook of pragmatics*. Amsterdam: John Benjamins, pp.1-22.

Archer, D., et al. (2012) *Pragmatics: an advanced resource book*. London: Routledge.

Argamon, S, et al. (2007) 'Stylistic text classification using functional lexical features', *Journal of the American Society for Information Science and Technology,* 58(6), pp.802–822.

Argamon, S., et al. (2009) Gender, genre and writing style in formal written texts. *Text & Talk*, 23(3), pp.321-346.

Argamon, S., and Koppel, M. (2012) A systemic functional approach to automated authorship analysis. *Journal of Law and Policy*, 21, pp.299-315.

Austin, J.L. (1962) *How to do things with words*. Oxford: Clarendon.

Baayen, H. et al. (2002) 'An experiment in authorship attribution', *Textuelles* (1), pp.69-75.

Banks-Smith, N. (2016) 'Old is the new new in Ambridge', *The Guardian*, 16 February. Available at: https://www.theguardian.com/tv-and-radio/2016/feb/16/nancy-banks-smith-on-the-archers-old-is-the-new-new-in-ambridge (Accessed January 2021).

Barr, S. (2018) 'Dinner, supper or tea?', *The Independent,* 26 May. Available at: https://www.independent.co.uk/life-style/food-and-drink/dinner-supper-tea-which-one-uk-brits-debate-evening-meal-yougov-a8363331.html. (Accessed January 2021).

Barron, A. (2013) 'Instant messaging', in Herring, D. et al. (eds.) *Pragmatics of Computer-mediated Communication*. Berlin: de Gruyter, pp.135-162.

Bell, A. (1984) 'Language style as audience design', *Language in Society*, 13(2), pp.145-204.

Bennison, N. (1998), 'Accessing character through conversation: Tom Stoppard's *Professional Foul*', in Culpeper, J. et al. (eds.) *Exploring the Language of Drama.* Abingdon: Routledge, pp.77-92.

Berman, R. (2008) 'The psycholinguistics of developing text construction', *Journal of Child Language*, 35(4), pp.735–71.

Bernstein, B. (1962) 'Linguistic codes, hesitation phenomena and intelligence', *Language and Speech*, 5(4), pp.221–240.

Biber, D. (1988) *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., et al. (2000) *Longman grammar of spoken and written English*. London: Longman.

Bloch, B. (1948) 'A set of postulates for phonemic analysis', *Language*, 24(1), pp.3-46.

Bousfield, D. (2008) *Impoliteness in interaction.* Amsterdam: John Benjamins. Vol. 167.

Bousfield, D. (2014) ‘Stylistics, speech acts and im/politeness theory’, in: Burke, M. (ed.) *The Routledge handbook of stylistics*. London: Routledge, pp.118-136.

Braber, N. & Butterfint, Z. (2008) ‘Local identity and sound change in Glasgow: a pilot study’, *Leeds Working Papers in Linguistics and Phonetics*, 13, pp.22–43.

Braber, N. (2009) ‘“I'm not a fanatic Scot, but I love Glasgow”: concepts of local and national identity in Glasgow’, *Identity*, 9(4), pp.307-322.

Braber, N. (2018) ‘Performing identity on screen: Language, identity, and humour in Scottish television comedy’, in: Bassiouney, R. (ed.) *Dialect and identity performance*. London: Routledge, 265-285.

Bradac, J. J. (1982) ‘A rose by another name: attitudinal consequences of lexical variation’, in Ryan, E., and Giles, H (eds.), *Attitudes toward language variation: social and applied contexts*. London: Arnold, pp.99-115.

Bradley, A.C. (1905) Shakespearean tragedy: lectures on *Hamlet, Othello*, *King Lear, Macbeth*. London: Macmillan.

Brennan, M.R. and Greenstadt, R. (2009) ‘Practical attacks against authorship recognition techniques’ *Twenty-First IAAI Conference*. California, USA, 14-16 July. Available at: Practical Attacks Against Authorship Recognition Techniques (drexel.edu) (Accessed: October 2017).

Brennan, M., et al. (2012) ‘Adversarial stylometry: circumventing authorship recognition to preserve privacy and anonymity’, *ACM Transactions on Information and System Security (TISSEC)*, 15(3), pp.1-22.

Brinton, L. (1996) *Pragmatic markers in English: grammaticalization and discourse functions*. Berlin and New York: Mouto de Gruyter.

Brown, R. and Gilman, A. (1989) ‘Politeness theory and Shakespeare's four major tragedies, *Language in Society*, 18(2), pp.159-212.

Brown, P., and Levinson, S.C. (1987) *Politeness: some universals in language usage*. Cambridge: Cambridge University Press. Vol. 4.

Bucholtz, M. and K. Hall. (2004) ‘Language and identity’, in Duranti, A. (ed.) *A Companion to linguistic anthropology*. Oxford: Blackwell Publishing, pp.369-394.

Bucholtz, M. and Hall, K. (2005) ‘Identity and interaction: a sociocultural linguistic approach’, *Discourse Studies*, (7), pp.585–614.

Burrows, J.F. (1987) *Computation into criticism: a study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon Press.

Burton, D. (1980) *Dialogue and discourse: A sociolinguistic approach to modern drama dialogue and naturally occurring conversation*. London; Boston: Routledge & Kegan Paul.

Caliskan, A., and Greenstadt, R. (2012) ‘Translate once, translate twice, translate thrice and attribute: identifying authors and machine translation tools in translated text’, *IEEE Sixth International Conference on Semantic Computing,* 19-21 September, pp.121-125.

Carter, E. (2021) ‘Distort, extort, deceive and exploit: exploring the inner workings of a romance fraud’, *The British Journal of Criminology*, 61(2), pp.283-302.

Carter, R., and Simpson, P. (eds.) (1989) *Language, discourse and literature: a reader in discourse stylistics.* London: Unwin Hyman.

Chiang, E. & Grant, T. (2017) ‘Online grooming: moves and strategies’ *Language and Law / Linguagem e Direito,* 4(1), pp.103-141.

Chiang, E. and Grant, T. (2019) ‘Deceptive identity performance: offender moves and multiple personas in online child abuse conversations’, *Applied Linguistics,* 40(4), pp.675-698.

Clarke, I. and Grieve, J. (2017) 'Dimensions of abusive language on Twitter' *Proceedings of the first workshop on abusive language online* (pp.1-10) 14(9). Available at: https://doi.org/10.1371/journal.pone.0222062 (Accessed January 2018).

Coates, J. (2015) *Women, men and language: a sociolinguistic account of gender differences in language*. London: Routledge.

Coulthard, M. (1994) 'On the use of corpora in the analysis of forensic texts', *The International Journal of Speech, Language and the Law,* 1(1), pp.27-43.

Coulthard, M. (2004) 'Author identification, idiolect, and linguistic uniqueness', *Applied Linguistics*, 25(4), pp.431-447.

Coulthard, M. (2020) 'Experts and opinions: in my opinion', in Coulthard, M. and Johnson, A. (eds.) *The Routledge handbook of forensic linguistics*. London: Routledge, pp.523-538.

Coupland, N. (2001) 'Dialect stylization in radio talk', *Language in Society*, 30(3), pp.345-375.

Coupland, N. (2004) 'Age in social and sociolinguistic theory', in Nussbaum, J.F. and Coupland, J. (eds.) *Handbook of communication and aging research*. London: Routledge, pp.69-90.

Crystal, D. (2006) *Words, words, words.* Oxford: Oxford University Press.

Crystal, D and Davy, D. (1969) *Investigating English style* (1st ed.). Abingdon: Routledge.

Culpeper, J. (1996) 'Towards an anatomy of impoliteness', *Journal of Pragmatics*, 25(3), pp.349-367.

Culpeper, J. (2005) 'Impoliteness and entertainment in the television quiz show: 'The Weakest Link', *Journal of Politeness Research*, 1(1), pp.35-72.

Culpeper, J. (2011) *Impoliteness: using language to cause offence.* Cambridge: Cambridge University Press.

Culpeper, J. (2014) *Language and characterisation: people in plays and other texts*. London: Routledge.

Culpeper, J. et al. (2003) ‘Impoliteness revisited: with special reference to dynamic and prosodic aspects’, *Journal of Pragmatics*, 35(10-11), pp.1545-1579.

Culpeper, J., and Kytö, M. (2010) *Early modern English dialogues: spoken interaction as writing*. Cambridge: Cambridge University Press.

Culpeper, J., and Haugh, M. (2014) *Pragmatics and the English language*. London: Macmillan International Higher Education.

Cutler, M. (2019) ‘Mary Cutler Puts Down Her Pen’, *BBC Radio 4*, [n.d.]. Available at: https://www.bbc.co.uk/programmes/articles/54FrhTn3dyFk66Z7h4Gr0hV/mary-cutler-puts-down-her-pen. (Accessed January 2020).

Davies, K. (2011) ‘Mousey - and *The Archers* archive’, *BBC Radio 4,* 2/9. Available at: https://www.bbc.co.uk/blogs/thearchers/entries/e2c47789-8fc1-3a00-aa47-9c3bbea55228. (Accessed: January 2020).

Davies, K. (2013) ‘Writing *The Archers*: from idea to airwaves’, *BBC Radio 4,* 21/10. Available at: https://www.bbc.co.uk/blogs/writersroom/entries/989184e1-10a1-3c2d-916e-cfbf67c7a334. (Accessed: May 2019).

Day, S., et al. (2016) ‘Adversarial Authorship, AuthorWebs, and Entropy-Based Evolutionary Clustering’. *25th International Conference on Computer Communication and Networks (ICCCN).* Hawaii, 1-4 August. pp.1-6. Available at: doi: 10.1109/ICCCN.2016.7568489.

Douglas, F. (2019) ‘English in Scotland’, in Kachru, B.B. et al. (eds.) *The handbook of world Englishes*. Oxford: John Wiley, pp.17-33.

Dynel, M. (2013) ‘On impoliteness and drama discourse: an interview with Jonathan Culpeper’, *International Review of Pragmatics*, 5(1), pp.163-188.

Eckert, P. (1989) *Jocks and burnouts: social categories and identity in the high school.* USA: Teachers College Press.

Eckert, p. (2021) ‘Afterword’, in Hall-Lew, L., et al. (eds.) *Social meaning and linguistic variation: theorizing the third wave*. Cambridge: Cambridge University Press, pp.382-387.

Eder, J., et al. (2010) *Characters in fictional worlds*. Berlin and New York: De Gruyter.

Edmonson, W. (1981) *Spoken discourse: a model for analysis*. London: Longman

Enkvist, N.E. (1964) ‘On defining style: an essay in applied linguistics’, in Spencer, J. (ed.) *Linguistics and Style*. Oxford: Oxford University Press, pp.1–56.

Enkvist, N.E. (1978) ‘Stylistics and text linguistics’, in Dressler, W. (ed.) *Current trends in textlinguistics.* Berlin, Boston: De Gruyter, pp.174-190.

Eunson, B. (2019) ‘The importance of using Scots language in the classroom’, *Times Educational Supplement*, 7 March. Available at: https://www.tes.com/news/importance-using-scots-language-classroom. (Accessed: March 2021).

Fawcett, B. and Hearn, J. (2004) ‘Researching others: epistemology, experience, standpoints and participation’, *International Journal of Social Research Methodology*, 7(3), pp.201-218.

Ferguson, E., and Singh Kohli, H. (2010) ‘Is *The Archers* guilty of national stereotyping?’, 28 November. Available at: https://www.theguardian.com/theobserver/2010/nov/28/archers-jazzer-scottish-stereotype. (Accessed: March 2021).

Fischer, K. (2000) *From cognitive semantics to lexical pragmatics: the functional polysemy of discourse particles*. Berlin and New York: De Gruyter.

Forster, E.M. (1927) *Aspects of the novel*. London: Edward Arnold.

Foster, D. (2000) *Author unknown: on the trail of anonymou*s. New York: Henry Holt.

Fox Tree, J.E., and Schrock, J.C. (1999) ‘Discourse markers in spontaneous speech: oh what a difference an *oh* makes. *Journal of Memory and Language*, 40(2), pp.280-295.

Fox Tree, J.E. (2015) ‘Discourse markers in writing’, *Discourse Studies*, 17(1), pp.64-82.

Fraser, B. (1990) ‘An approach to discourse markers’, *Journal of Pragmatics*, 14(3), pp.383-398.

Furkó, P., and Abuczki, Á (2014) ‘English discourse markers in mediatised political interviews’, *Brno Studies in English*, 40(1), pp.45-64.

Geraghty, C. (1981) ‘Continuous serial: a definition’, in Dyer, R. (ed.) *Coronation Street*. London: BFI. pp.9-26.

Gibbons, J. (2003) *Forensic linguistics: an introduction to language in the justice system*. Oxford: Wiley-Blackwell.

Goffman, E. (1959) *The presentation of self in everyday life*. USA: Anchor.

Goffman, E. (1981) *Forms of talk*. Philadelphia: University of Pennsylvania Press.

Grant, T. (2010) ‘Text messaging forensics Txt 4n6: idiolect free authorship analysis?’, in Coulthard, M. and Johnson, A. (eds.) *The Routledge handbook of forensic linguistics*. Abingdon: Routledge, pp.508 – 522.

Grant, T. (2012) ‘Txt 4n6: method, consistency, and distinctiveness in the analysis of SMS text messages’, *Journal of Law and Policy*, 21, pp.467-494.

Grant, T., and Baker, K. (2001) ‘Identifying reliable, valid markers of authorship: a response to Chaski’, *Forensic Linguistics*, 8, pp.66-79.

Grant, T. and MacLeod, N. (2016) ‘Assuming identities online: experimental linguistics applied to the policing of online paedophile activity’ *Applied Linguistics*, 37(1), pp.50–70.

Grant, T., et al. (2017) *Quantitative research methods for linguists: a questions and answers approach for students*. Abingdon: Routledge.

Grant, T., and MacLeod, N. (2018) ‘Resources and constraints in linguistic identity performance: a theory of authorship’, *Language and Law/Linguagem e Direito*, 1, pp.80-96.

Grant, T., and MacLeod, N. (2020) *Language and online identities: the undercover policing of internet sexual crime*. Cambridge: Cambridge University Press.

Greenhill, S. (2015) *The Archers* ‘must not be EastEnders in a field’, *Daily Mail*, 27 January. Available at: (https://www.dailymail.co.uk/news/article-2927570/The-Archers-not-EastEnders-field-Director-general-signals-roots-fans-complained-racy-plots.html.) (Accessed: January 2020).

Grice, H.P. (1975) ‘Logic and conversation’, in Cole, P. & Morgan, J. L. (eds.) *Syntax and semantics 3: speech acts*. New York: Academic, pp.41–58.

Grieve, J. (2007) ‘Quantitative authorship attribution: an evaluation of techniques’, *Literary and Linguistic Computing,* 22(3), pp.251-270.

Grieve, J., et al. (2019) ‘Attributing the Bixby letter using n-gram tracing’, *Digital Scholarship in the Humanities*, 34(3), pp.493-512.

Grove, C., and Wyatt, S. (2013) *So you want to write radio drama?* London: Nick Hern.

Hagan, A. (2002) *Urban Scots dialect writing.* Bern: Peter Lang.

Hall-Lew, L., et al. (eds.) (2021) *Social meaning and linguistic variation: theorizing the third wave*. Cambirdge: Cambridge University Press.

Hansen, M.B.M. (1998) *The function of discourse particles: a study with special reference to spoken standard French.* Amsterdam: John Benjamins Publishing. Vol. 53.

Haworth, K. (2013) ‘Audience design in the police interview: the interactional and judicial consequences of audience orientation’ *Language in Society*, 42(1), pp.45-69.

Heritage, J. (1984) ‘A change-of-state token and aspects of its sequential placement’, in: Maxwell, A.J., and Heritage, J. (eds.) *Structures of social action. Studies in conversation analysis.* Cambridge: Cambridge University Press. pp.199–345.

Herman, V. (1995) *Dramatic discourse: dialogue as interaction in plays*. London: Routledge.

Herring, S.C. (2004) ‘Computer-mediated discourse analysis: an approach to researching online behavior’, in Barab, S.A. et al. (eds.), *Designing for virtual communities in the service of learning*. New York: Cambridge University Press. pp.338-376.

Hilton, M.L. and Holmes, D.I. (1993) ‘An assessment of cumulative sum charts for authorship attribution’, *Literary and Linguistic Computing*, 8(2), pp.73-80.

Hodson, J. (2014) *Dialect in film and literature*. London: Macmillan International Higher Education.

Holmes, J. (2013) *Women, men and politeness*. Abingdon: Routledge.

Hoover, D.L. (2003) ‘Frequent collocations and authorial style’, *Literary and Linguistic Computing*, 18(3), pp.261-286.

Hota, S. et al. (2006) ‘Performing gender: automatic stylistic analysis of Shakespeare’s characters’ *Proceedings of Digital Humanities.* Paris. (pp.100-104).

Jockers, M.L., and Thalken, R. (2020) *Text analysis with R*. New York: Springer International Publishing.

Johnson, A. and Wright, D. (2014) ‘Identifying idiolect in forensic authorship attribution: an n-gram textbite approach’ *Language and Law/ Linguagem e Direito*, 1(1), pp.37-69.

Johnstone, B. (1996) *The linguistic individual: self-expression in language and linguistics*. Oxford: Oxford University Press.

Johnstone, B. (2013) *Speaking Pittsburghese: the story of a dialect*. Oxford: Oxford University Press.

Johnstone, B. et al. (2006) ‘Mobility, indexicality, and the enregisterment of “Pittsburghese”’, *Journal of English Linguistics*, 34(2), pp.77-104.

Joseph, J. (2004) *Language and identity: national, ethnic, religious*. London: Palgrave MacMillan.

Jucker, A.H. (1993) ‘The discourse marker *well*: a relevance-theoretical account. *Journal of Pragmatics*, 19(5), pp.435-452.

Juola, P. (2008) *Authorship attribution* Boston, Delft: Now Publishers. Vol. 3.

Juola, P., and Vescovi, D. (2010) ‘Analysing stylometric approaches to author obfuscation’, in Peterson, G. and Shenoi, S. (eds.) *Advances in digital forensics*. Berlin, Heidelberg: Springer, pp.115-125.

Juola, P. and Vescovi, D. (2010) ‘Empirical evaluation of authorship obfuscation using JGAAP’ *Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security* (pp.14-18). Available at: https://doi.org/10.1145/1866423.1866427

Kitson, H. D. (1921) *The mind of the buyer*. New York: Macmillan.

Knights, L.C. (1964) ‘How many children had Lady Macbeth?’, *Explorations*. New York: New York University Press, pp.15-54.

Koppel, M., et al. (2009) ‘Computational methods in authorship attribution’ *Journal of the American Society for Information Science and Technology*, 60(1), pp.9-26.

Koppel, M. et al.. (2012) ‘The “fundamental problem” of authorship attribution’, *English Studies*, 93(3), 284-291.

Koskenniemi, I. (1962) *Studies in the vocabulary of English drama, 1550-1600, excluding Shakespeare and Ben Jonson*. Finland: Annales Universitatis Turkuensis. Vol. 84.

Kozloff, S. (2000) *Overhearing film dialogue*. California: University of California Press.

Kredens, K., et al. (2019) ‘Toward linguistic explanation of idiolectal variation–understanding the black box, *14th Biennial Conference of the International Association of Forensic Linguists.* Melbourne. 1-5 July.

Labov, W. (1966) *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labov W. (1972) *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. and Fanshel, D. (1977) *Therapeutic discourse: psychotherapy as conversation*. New York: Academic Press.

Lakoff, R. (1973) ‘Language and woman's place’, *Language in Society*, 2(1), pp.45-79.

Larner, S. (2014) ‘A preliminary investigation into the use of fixed formulaic sequences as a marker of authorship’, *International Journal of Speech, Language and the Law*, 21(1).

Lawler, S. (2000) *Mothering the self: mothers, daughters, subjects*. London: Routledge.

Leech, G.N., and Short, M. (2007) *Style in fiction: a linguistic introduction to English fictional prose*. London: Pearson Education.

Locher, M. (2006) ‘Polite behavior within relational work: the discursive approach to politeness’, *Multilingua*. 25(3), pp.249–267.

Locher, M. and Watts, R.J. (2005) ‘Politeness theory and relational work.’ *Journal of Politeness Research,* (1), pp.9-33**.**

Love, H. (2002) *Attributing authorship: an introduction*. Cambridge: Cambridge University Press.

Macafee, C. (1983) *Glasgow*. Amsterdam: John Benjamins.

Macaulay, R.K. (2005) *Talk that counts: age, gender, and social class differences in discourse*. Oxford: Oxford University Press.

MacLeod, N. (2020) ‘Assuming identities online: authorship synthesis in undercover investigations’, in Coulthard, M. et al. (eds.) 2nd ed. *The Routledge handbook of forensic linguistics*. London: Routledge, pp.159-173.

Macleod, N. and Grant, T. (2012) ‘Whose Tweet? Authorship analysis of micro-blogs and other short-form messages’ in Tomblin, S. et al. (eds.) *Proceedings of the International Association of Forensic Linguists’ tenth biennial conference.* Aston University, Birmingham, United Kingdom, 11/07/11, pp.210-224.

MacLeod, N., and Grant, T. (2017) ‘“go on cam but dnt be dirty”: linguistic levels of identity assumption in undercover online operations against child sex abusers’, *Language and law/Linguagem e direito*, 4(2), pp.157-175.

McIntyre, D. (2010) ‘Dialogue and characterisation in Quentin Tarantino’s *Reservoir Dogs*: a corpus stylistic analysis’, in McIntyre, D. and Busse, B. (eds.) *Language and style*. London: Palgrave MacMillan, pp.162-183.

McIntyre, D. (2015a) ‘Characterisation’, in Stockwell, P., and Whiteley, S. (eds.) *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp.149-164.

McIntyre, D (2015b) ‘Dialogue: credibility versus realism in fictional speech’, in Sotirova, V. (ed.) *The Bloomsbury companion to stylistics.* London: Bloomsbury, pp. 430-443.

McIntyre, D. and Bousfield, D. (2017) ‘(Im)politeness in fictional texts’, in Culpeper et al. (eds.) *The Palgrave handbook of linguistic (im)politeness.* London: Palgrave MacMillan, pp.759-783.

McMenamin, G.R. (1993) *Forensic stylistics.* Amsterdam: Elsevier.

McMenamin, G.R. (2002) *Forensic linguistics: advances in forensic stylistics*. Boca Raton: CRC Press.

McMenamin, G.R. (2020) 'Forensic stylistics: theory and practice of forensic stylistics', in Coulthard, M. and Johnson, A. (eds.) *The Routledge handbook of forensic linguistics*. Abingdon: Routledge, pp.539-557.

Milroy, L. (1980) Language and social networks. Oxford: Blackwell.

Milroy, J., and Milroy, L. (1978) 'Belfast: Change and Variation in an urban vernacular', in Trudgill, P. (ed.) *Sociolinguistic patterns in British English*. London: Arnold, pp.19-36.

Mills, S. (1995) *Feminist stylistics*. London: Routledge.

Montini, D., and Ranzato, I. (2021) *The dialects of British English in fictional texts*. London: Routledge.

Morin, C., et al. (2020) 'Dialect syntax in Construction Grammar: theoretical benefits of a constructionist approach to double modals in English', *Belgian Journal of Linguistics*, 34(1), pp.248-258.

Mosteller, F., and Wallace, D. (1964) *Inference and disputed authorship: the Federalist*. Boston: Addison-Wesley.

Müller, C.A. (2010) 'James Kelman's literary language', PhD thesis, Flinders University, Adelaide.

Müller, C.A. (2011) *A Glasgow voice: James Kelman's literary language*. Newcastle: Cambridge Scholars Publishing.

Munro, M. (2013) *The complete patter*. Edinburgh: Birlinn.

Nini, A. (2014) *Authorship profiling in a forensic context.* PhD thesis. Aston University, Available at: https://research.aston.ac.uk/en/studentTheses/authorship-profiling-in-a-forensic-context (Accessed: 14 July 2021).

Nini, A. (2018), 'An authorship analysis of the Jack the Ripper letters.' *Digital Scholarship in the Humanities*, 33(3), pp.621-636.

Norrick, N.R. (2009) ‘Interjections as pragmatic markers’, *Journal of Pragmatics*, 41(5), pp.866-891.

Ohmann, R. (1964) ‘Generative grammars and the concept of literary style’, *Word*, *20*(3), pp.423-439.

Overdorf, R. and Greenstadt, R. (2016) ‘Blogs, Twitter feeds, and Reddit comments: cross-domain authorship attribution’, *Proc. Priv. Enhancing Technol.* 3, pp.155-171.

Page, N. (1973) *Speech in the English novel*. London: Longman

Pust, L., (1998) ‘I cannae see it’: negation in Scottish English and dialect data from the British National Corpus’. *AAA: Arbeiten aus Anglistik und Amerikanistik*, *23*(1), pp.17-30.

Pfister, M. (1991) *The theory and analysis of drama*. Cambridge: Cambridge University Press.

Rickford, J.R. (1986) ‘The need for new approaches to social class analysis in sociolinguistics’, *Language and Communication*, 6(3), pp.215-221.

Quaglio, P. (2009) *Television dialogue: the sitcom Friends vs. natural conversation.* Amsterdam: John Benjamins Publishing. (Vol. 36).

Rao, J.R. and Rohatgi, P. (2000) ‘Can pseudonymity really guarantee privacy?’ *USENIX Security Symposium* Denver, Colorado. 14-16 August, pp.85-96.

Rey, J. M. (2001) ‘Changing gender roles in popular culture: dialogue in Star Trek episodes from 1966 to 1993’, in Conrad, S. & Biber, D. (eds.) *Variation in English: multi-dimensional studies*. London: Routledge, pp.139-156.

Richardson, K. (2010) *Television dramatic dialogue: a sociolinguistic study*. Oxford: Oxford University Press.

Rimmon-Kenan, S. (1983) *Narrative fiction: contemporary poetics.* London: Routledge.

River City, *Episode 622,* 2008. BBC Scotland. 16/04/2008.

Rudman, J. (1998) ‘Non-traditional authorship attribution studies in the Historia Augusta: some caveats’, *Literary and Linguistic Computing*, 13(3), pp.151-157.

Rudman, J. (2012) ‘The state of non-traditional authorship attribution studies: some problems and solutions’, *English Studies*, 93(3), pp.259-274.

Rudman, J. (2016) ‘Non-traditional authorship attribution studies of William Shakespeare’s canon: some caveats.’ *Journal of Early Modern Studies*, 5, pp.307-328.

Ruzich, C. and Blake, J. (2015) ‘Ain't nothing like the real thing: dialect, race, and identity in Stockett's novel *The Help*’. *The Journal of Popular Culture*, 48(3), 534-547.

Sacks, H. (1984) 'Notes on methodology’ in Atkinson, J. and Heritage, J. (eds.) *Structure of social action: studies in conversation analysis.* Cambridge: Cambridge University Press, pp.21-27.

Sanderson, I. (2006) ‘Separating *The Archers* from fiction’, *The Guardian*. 7 November. Available at: https://www.theguardian.com/culture/tvandradioblog/2006/nov/07/realityradioseparatingthea. (Accessed: June 2021).

Schiffrin, D. (1987) *Discourse markers*. Cambridge: Cambridge University Press.

Schiffrin, D. (1994) *Approaches to discourse*. Oxford: Blackwell.

Schneider, K.P., and Barron, A. (2010) ‘Variational pragmatics: variation and change’ *Pragmatic Perspectives*, 6, pp.239-267.

Schourup, L. (1985) *Common discourse particles in English conversation.* New York: Routledge.

Short, M. (1996) *Exploring the language of poems, plays and prose*. London: Longman.

Sidnell, J. (2011) *Conversation analysis: an introduction*. Oxford: Blackwell-Wiley. Vol. 45.

Sigelman, L. and Jacoby, W. (1996) ‘The not-so-simple art of imitation: pastiche, literary style, and Raymond Chandler’ *Computers and the Humanities*, 30(1), pp.11-28.

Simpson, P. (1989) ‘Politeness phenomena in Ionesco’s *The Lesson*’, in Carter, R., and Simpson, P. (eds.) *Language, discourse and literature*. London: Routledge, pp.177-198.

Sinclair, J., and Coulthard, M. (1975) *Towards an analysis of discourse*. Oxford: Oxford University Press.

Somers, H.D, and Tweedie, F. (2003) ‘Authorship attribution and pastiche’, *Computers and the Humanities*, 37: 407–429.

Stamatatos, E. (2009) ‘A survey of modern authorship attribution methods’, *Journal of the American Society for Information Science and Technology*, 60(3), pp.538-556.

Stamatatos, E. (2012) ‘On the robustness of authorship attribution based on character n-gram features’, *Journal of Law and Policy*, 21, pp.421-439.

Stockett, K. (2009) *The Help.* London: Penguin.

Stockwell, P. (2020) ‘Literary dialect as social deixis’, *Language and Literature*, 29(4), pp.358-372.

Stokoe, E., and Edwards, D. (2008) ‘Did you have permission to smash your neighbour's door?’ Silly questions and their answers in police-suspect interrogations’, *Discourse Studies*, 10(1), pp.89-111.

Stuart-Smith, J., et al. (2007) ‘‘Talkin’ Jockney’? variation and change in Glaswegian accent’, *Journal of Sociolinguistics*, 11(2), pp.221-260.

Stubbs, M. (1993) ‘British traditions in text analysis’, in Baker, M. et al. (eds.) *Text and technology: in honour of John Sinclair*. Amsterdam: John Benjamins, pp.1-33.

Thompson, C. (2011) *Writing soap: how to write popular continuing drama*. London: Aber.

Toolan, M. (1998) 'The give and take of talk, and Caryl Churchill's *Cloud Nine*', in Culpeper, J. et al. (eds.) *Exploring the language of drama*. Abingdon: Routledge, pp.152-170.

Trudgill, P. (1974) Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, 3(2), pp.215-246.

Tulloch, G. (1997) *The Scots language in Australia*. Edinburgh: Edinburgh University Press.

Ullman, S. (1973) *Meaning and style*. Oxford: Blackwell.

Verdonk, P. (2002) *Stylistics*. Oxford: Oxford University Press.

Wallis, M., and Shepherd, S. (1998) *Studying plays*. London: Hodder.

Wardhaugh, R. and Fuller, J.M. (2015) *An introduction to sociolinguistics*. Oxford: Wiley.

Wolfram, W. (1974) 'The relationship of white southern speech to vernacular black English', *Language*, 50(3), pp.498-527.

Wells, J.C. (1982) *Accents of English*. Cambridge: Cambridge University Press. Vol. 1.

Wierzbicka, A. (1996) *Semantics: primes and universals*. Oxford: Oxford University Press.

Winter, E. (1996) 'The statistics of analysing very short texts in a criminal context', in Kniffka, H. (ed.) *Recent developments in forensic linguistics*. Frankfurt: Peter Lang, pp.141–79.

Woman and Home (2016) 'We meet the writer behind *The Archers'*, 20 September. Available at: The Writer Behind The Archers | Woman & Home (womanandhome.com). (Accessed June 2021).

Wright, D. (2014) Stylistics versus statistics: a corpus linguistic approach to combining techniques in forensic authorship analysis using Enron emails. PhD Thesis. University of Leeds. Available at: https://etheses.whiterose.ac.uk/8278/ (Accessed 14 July 2022).

Wright, D. (2017) ‘Using word n-grams to identify authors and idiolects: a corpus approach to a forensic linguistic problem’, *International Journal of Corpus Linguistics*, 22(2), pp.212-241.

10. Appendices

Appendix 1 **Sample script, *The Archers***

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 2a Quantitative Results summary spreadsheet

Number of Turns

	David	Pat	Ruth	Jen	Lilian	Tom	Brian	Kenton	Eliz	Helen	Lynda	Pip	Tony	Jill	Susan	Fallon	Jim	Eddie	Brenda	Jazzer
Writer 1	1297	1020	759	649	622	858	981	651	1113	601	502	616	1024	579	319	368	465	545	537	655
Writer 2	582	897	599	493	279	631	526	332	356	491	371	609	394	353	562	265	80	225	267	75
Writer 3	1120	382	954	731	794	524	665	948	453	404	856	467	243	341	468	580	744	609	388	407
Writer 4	901	724	696	276	614	498	362	418	399	467	314	588	429	493	198	572	291	286	238	329
Writer 5	1045	904	709	968	716	926	680	444	742	810	424	532	614	618	758	296	537	174	389	368
Writer 6	598	642	734	804	950	491	627	695	390	372	669	362	367	283	270	572	311	401	410	450

Number of Words: \b[A-Za-z0-9]+\b

	David	Pat	Ruth	Jen	Lilian	Tom	Brian	Kenton	Eliz	Helen	Lynda	Pip	Tony	Jill	Susan	Fallon	Jim	Eddie	Brenda	Jazzer
Writer 1	11243	9157	6781	5765	5361	7664	10532	5903	10919	5283	5552	5662	9278	5730	3335	3073	4522	5181	4427	6480
Writer 2	5452	8775	5731	4765	2629	6540	5822	3342	3472	5109	3983	5679	3169	3310	6291	2695	887	2044	2304	695
Writer 3	11965	4496	10001	7325	8496	5450	8424	11835	5667	4922	10795	5170	2598	4017	5944	5633	8868	6001	4302	4034
Writer 4	9180	6993	5919	2396	5793	4864	4315	5222	3491	4153	3745	4682	3828	4831	1682	4936	2921	2413	2517	3058
Writer 5	10862	9169	7425	10206	7474	9634	7748	5707	8448	8633	5264	5367	6191	6806	7891	3371	6582	1830	4258	3866
Writer 6	5278	5983	6319	7020	8248	4304	6256	6657	3398	3150	7095	3034	2953	2327	2523	4888	2900	3887	3308	3637

Number of Digits & Graphemes: [A-Za-z0-9]

	David	Pat	Ruth	Jen	Lilian	Tom	Brian	Kenton	Eliz	Helen	Lynda	Pip	Tony	Jill	Susan	Fallon	Jim	Eddie	Brenda	Jazzer
Writer 1	40393	33316	24697	21662	19850	28025	40527	21853	40842	19164	22135	19964	33069	21444	12466	11100	18067	18835	16345	23848
Writer 2	20012	32901	21089	17934	9855	24706	22022	12326	12971	18429	15906	20127	11636	12550	23275	9735	3455	7377	8561	2442
Writer 3	42905	16419	35630	26711	30298	19483	31677	43208	20595	17282	42124	18016	9074	14854	21305	19663	34116	21472	15513	14356
Writer 4	33813	26230	21814	8916	21217	17912	16707	19243	12839	14823	14697	16791	13966	17566	6275	17562	11725	8793	9309	11197
Writer 5	38906	33446	26682	37378	27061	34807	28884	20903	30347	30334	20096	18450	21944	24973	28359	12010	25156	6479	15195	13731
Writer 6	18889	22223	23002	25693	29488	15251	23343	23949	12488	11058	27287	10499	10496	8531	9076	17352	11093	13874	11509	12948

Average Word Length: calculated by dividing the total number of digits and graphemes in a text by the total number of words.

	David	Pat	Ruth	Jen	Lilian	Tom	Brian	Kenton	Eliz	Helen	Lynda	Pip	Tony	Jill	Susan	Fallon	Jim	Eddie	Brenda	Jazzer
Writer 1	3.592724	3.638309	3.642088	3.757502	3.702667	3.656707	3.847987	3.702016	3.740452	3.627484	3.986852	3.525963	3.564238	3.742408	3.737931	3.612105	3.995356	3.635399	3.692117	3.680247
Writer 2	3.67058	3.749402	3.679812	3.763694	3.748574	3.777676	3.782549	3.688211	3.735887	3.607164	3.993472	3.54411	3.671821	3.791541	3.69973	3.612245	3.895152	3.6091	3.715712	3.513669
Writer 3	3.585875	3.651913	3.562644	3.646553	3.566149	3.574862	3.760328	3.650866	3.634198	3.511174	3.902177	3.48472	3.492687	3.697784	3.584287	3.49068	3.847091	3.57807	3.605997	3.558751
Writer 4	3.683333	3.750894	3.68542	3.721202	3.662524	3.682566	3.871842	3.684987	3.677743	3.569227	3.924433	3.586288	3.64838	3.6361	3.730678	3.557942	4.014036	3.644012	3.698451	3.661543
Writer 5	3.581845	3.647726	3.593535	3.662355	3.620685	3.612933	3.72793	3.662695	3.592211	3.513726	3.817629	3.437675	3.5445	3.669262	3.593841	3.562741	3.821939	3.540437	3.568577	3.551733
Writer 6	3.578818	3.714357	3.640133	3.659972	3.57517	3.543448	3.731298	3.597566	3.675103	3.510476	3.845948	3.460448	3.554352	3.666094	3.597305	3.549918	3.825172	3.569334	3.479141	3.560077

Average Turn Length: calculated by dividing the total number of words in a text by the total number of sentences.

	David	Pat	Ruth	Jen	Lilian	Tom	Brian	Kenton	Eliz	Helen	Lynda	Pip	Tony	Jill	Susan	Fallon	Jim	Eddie	Brenda	Jazzer
Writer 1	8.668466	8.977451	8.934124	8.882897	8.618971	8.932401	10.73598	9.067588	9.810422	8.790349	11.05976	9.191558	9.060547	9.896373	10.45455	8.350543	9.724731	9.506422	8.243948	9.89313
Writer 2	9.367698	9.782609	9.567613	9.665314	9.422939	10.3645	11.06844	10.06627	9.752809	10.4053	10.73585	9.325123	8.043147	9.376771	11.19395	10.16981	11.0875	9.084444	8.629213	9.266667
Writer 3	10.68304	11.76963	10.48323	10.02052	10.70025	10.40076	12.66767	12.48418	12.50993	12.18317	12.61098	11.07066	10.69136	11.78006	12.70085	9.712069	11.91935	9.853859	11.08763	9.911548
Writer 4	10.18868	9.65884	8.50431	8.681159	9.434853	9.767068	11.91989	12.49282	8.749373	8.892934	11.92675	7.962585	8.923077	9.799189	8.494949	8.629371	10.0378	8.437063	10.57563	9.294833
Writer 5	10.39426	10.1427	10.4725	10.54339	10.43855	10.40389	11.39412	12.8536	11.38544	10.65802	12.41509	10.08835	10.08306	11.01294	10.41029	11.38851	12.25698	10.51724	10.94602	10.50543
Writer 6	8.826087	9.319315	8.608992	8.731343	8.682105	8.765784	9.977671	9.578417	8.712821	8.467742	10.60538	8.381215	8.046322	8.222615	9.344444	8.545455	9.324759	9.693267	8.068293	8.082222

Appendix 2b Unique words by author

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 2c Bi-grams results

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 2d Tri-grams results

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 3a Jim corpora

(i) Type-token ratio
(ii) Jim corpora for keyword analysis
(iii) Jim corpora (2010)
(iv) Jim corpora (2015)
(v) David corpora (2010)
(vi) Eddie corpora (2010)

These have been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 4a Writers 1-6 Jazzer corpora

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 4b Writers 1-6 Lynda directing scenes

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 5a Helen & Rob corpora

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

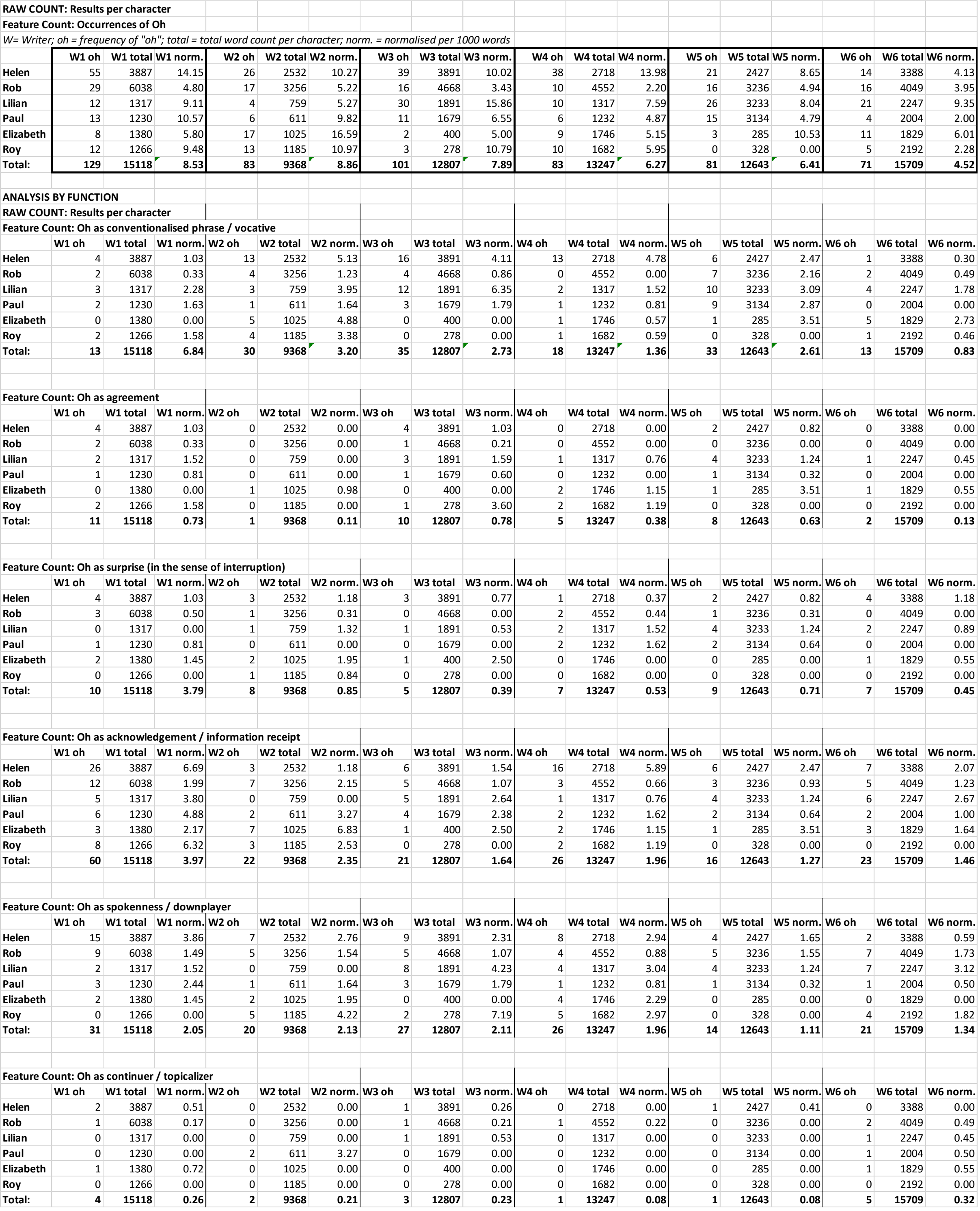
Appendix 5b Lilian & Paul corpora

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 5c Elizabeth & Roy corpora

This has been redacted in accordance with the Non-Disclosure Agreement with the BBC.

Appendix 5d *oh* Results spreadsheet



RAW COUNT: Results per character

Feature Count: Occurrences of Oh

W= Writer; oh = frequency of "oh"; total = total word count per character; norm. = normalised per 1000 words

	W1 oh	W1 total	W1 norm.	W2 oh	W2 total	W2 norm.	W3 oh	W3 total	W3 norm.	W4 oh	W4 total	W4 norm.	W5 oh	W5 total	W5 norm.	W6 oh	W6 total	W6 norm.
Helen	55	3887	14.15	26	2532	10.27	39	3891	10.02	38	2718	13.98	21	2427	8.65	14	3388	4.13
Rob	29	6038	4.80	17	3256	5.22	16	4668	3.43	10	4552	2.20	16	3236	4.94	16	4049	3.95
Lilian	12	1317	9.11	4	759	5.27	30	1891	15.86	10	1317	7.59	26	3233	8.04	21	2247	9.35
Paul	13	1230	10.57	6	611	9.82	11	1679	6.55	6	1232	4.87	15	3134	4.79	4	2004	2.00
Elizabeth	8	1380	5.80	17	1025	16.59	2	400	5.00	9	1746	5.15	3	285	10.53	11	1829	6.01
Roy	12	1266	9.48	13	1185	10.97	3	278	10.79	10	1682	5.95	0	328	0.00	5	2192	2.28
Total:	**129**	**15118**	**8.53**	**83**	**9368**	**8.86**	**101**	**12807**	**7.89**	**83**	**13247**	**6.27**	**81**	**12643**	**6.41**	**71**	**15709**	**4.52**

ANALYSIS BY FUNCTION

RAW COUNT: Results per character

Feature Count: Oh as conventionalised phrase / vocative

	W1 oh	W1 total	W1 norm.	W2 oh	W2 total	W2 norm.	W3 oh	W3 total	W3 norm.	W4 oh	W4 total	W4 norm.	W5 oh	W5 total	W5 norm.	W6 oh	W6 total	W6 norm.
Helen	4	3887	1.03	13	2532	5.13	16	3891	4.11	13	2718	4.78	6	2427	2.47	1	3388	0.30
Rob	2	6038	0.33	4	3256	1.23	4	4668	0.86	0	4552	0.00	7	3236	2.16	2	4049	0.49
Lilian	3	1317	2.28	3	759	3.95	12	1891	6.35	2	1317	1.52	10	3233	3.09	4	2247	1.78
Paul	2	1230	1.63	1	611	1.64	3	1679	1.79	1	1232	0.81	9	3134	2.87	0	2004	0.00
Elizabeth	0	1380	0.00	5	1025	4.88	0	400	0.00	1	1746	0.57	1	285	3.51	5	1829	2.73
Roy	2	1266	1.58	4	1185	3.38	0	278	0.00	1	1682	0.59	0	328	0.00	1	2192	0.46
Total:	**13**	**15118**	**6.84**	**30**	**9368**	**3.20**	**35**	**12807**	**2.73**	**18**	**13247**	**1.36**	**33**	**12643**	**2.61**	**13**	**15709**	**0.83**

Feature Count: Oh as agreement

	W1 oh	W1 total	W1 norm.	W2 oh	W2 total	W2 norm.	W3 oh	W3 total	W3 norm.	W4 oh	W4 total	W4 norm.	W5 oh	W5 total	W5 norm.	W6 oh	W6 total	W6 norm.
Helen	4	3887	1.03	0	2532	0.00	4	3891	1.03	0	2718	0.00	2	2427	0.82	0	3388	0.00
Rob	2	6038	0.33	0	3256	0.00	1	4668	0.21	0	4552	0.00	0	3236	0.00	0	4049	0.00
Lilian	2	1317	1.52	0	759	0.00	3	1891	1.59	1	1317	0.76	4	3233	1.24	1	2247	0.45
Paul	1	1230	0.81	0	611	0.00	1	1679	0.60	0	1232	0.00	1	3134	0.32	0	2004	0.00
Elizabeth	0	1380	0.00	1	1025	0.98	0	400	0.00	2	1746	1.15	1	285	3.51	1	1829	0.55
Roy	2	1266	1.58	0	1185	0.00	1	278	3.60	2	1682	1.19	0	328	0.00	0	2192	0.00
Total:	**11**	**15118**	**0.73**	**1**	**9368**	**0.11**	**10**	**12807**	**0.78**	**5**	**13247**	**0.38**	**8**	**12643**	**0.63**	**2**	**15709**	**0.13**

Feature Count: Oh as surprise (in the sense of interruption)

	W1 oh	W1 total	W1 norm.	W2 oh	W2 total	W2 norm.	W3 oh	W3 total	W3 norm.	W4 oh	W4 total	W4 norm.	W5 oh	W5 total	W5 norm.	W6 oh	W6 total	W6 norm.
Helen	4	3887	1.03	3	2532	1.18	3	3891	0.77	1	2718	0.37	2	2427	0.82	4	3388	1.18
Rob	3	6038	0.50	1	3256	0.31	0	4668	0.00	2	4552	0.44	1	3236	0.31	0	4049	0.00
Lilian	0	1317	0.00	1	759	1.32	1	1891	0.53	2	1317	1.52	4	3233	1.24	2	2247	0.89
Paul	1	1230	0.81	0	611	0.00	0	1679	0.00	2	1232	1.62	2	3134	0.64	0	2004	0.00
Elizabeth	2	1380	1.45	2	1025	1.95	1	400	2.50	0	1746	0.00	0	285	0.00	1	1829	0.55
Roy	0	1266	0.00	1	1185	0.84	0	278	0.00	0	1682	0.00	0	328	0.00	0	2192	0.00
Total:	**10**	**15118**	**3.79**	**8**	**9368**	**0.85**	**5**	**12807**	**0.39**	**7**	**13247**	**0.53**	**9**	**12643**	**0.71**	**7**	**15709**	**0.45**

Feature Count: Oh as acknowledgement / information receipt

	W1 oh	W1 total	W1 norm.	W2 oh	W2 total	W2 norm.	W3 oh	W3 total	W3 norm.	W4 oh	W4 total	W4 norm.	W5 oh	W5 total	W5 norm.	W6 oh	W6 total	W6 norm.
Helen	26	3887	6.69	3	2532	1.18	6	3891	1.54	16	2718	5.89	6	2427	2.47	7	3388	2.07
Rob	12	6038	1.99	7	3256	2.15	5	4668	1.07	3	4552	0.66	3	3236	0.93	5	4049	1.23
Lilian	5	1317	3.80	0	759	0.00	5	1891	2.64	1	1317	0.76	4	3233	1.24	6	2247	2.67
Paul	6	1230	4.88	2	611	3.27	4	1679	2.38	2	1232	1.62	2	3134	0.64	2	2004	1.00
Elizabeth	3	1380	2.17	7	1025	6.83	1	400	2.50	2	1746	1.15	1	285	3.51	3	1829	1.64
Roy	8	1266	6.32	3	1185	2.53	0	278	0.00	2	1682	1.19	0	328	0.00	0	2192	0.00
Total:	**60**	**15118**	**3.97**	**22**	**9368**	**2.35**	**21**	**12807**	**1.64**	**26**	**13247**	**1.96**	**16**	**12643**	**1.27**	**23**	**15709**	**1.46**

Feature Count: Oh as spokenness / downplayer

	W1 oh	W1 total	W1 norm.	W2 oh	W2 total	W2 norm.	W3 oh	W3 total	W3 norm.	W4 oh	W4 total	W4 norm.	W5 oh	W5 total	W5 norm.	W6 oh	W6 total	W6 norm.
Helen	15	3887	3.86	7	2532	2.76	9	3891	2.31	8	2718	2.94	4	2427	1.65	2	3388	0.59
Rob	9	6038	1.49	5	3256	1.54	5	4668	1.07	4	4552	0.88	5	3236	1.55	7	4049	1.73
Lilian	2	1317	1.52	0	759	0.00	8	1891	4.23	4	1317	3.04	4	3233	1.24	7	2247	3.12
Paul	3	1230	2.44	1	611	1.64	3	1679	1.79	1	1232	0.81	1	3134	0.32	1	2004	0.50
Elizabeth	2	1380	1.45	2	1025	1.95	0	400	0.00	4	1746	2.29	0	285	0.00	0	1829	0.00
Roy	0	1266	0.00	5	1185	4.22	2	278	7.19	5	1682	2.97	0	328	0.00	4	2192	1.82
Total:	**31**	**15118**	**2.05**	**20**	**9368**	**2.13**	**27**	**12807**	**2.11**	**26**	**13247**	**1.96**	**14**	**12643**	**1.11**	**21**	**15709**	**1.34**

Feature Count: Oh as continuer / topicalizer

	W1 oh	W1 total	W1 norm.	W2 oh	W2 total	W2 norm.	W3 oh	W3 total	W3 norm.	W4 oh	W4 total	W4 norm.	W5 oh	W5 total	W5 norm.	W6 oh	W6 total	W6 norm.
Helen	2	3887	0.51	0	2532	0.00	1	3891	0.26	0	2718	0.00	1	2427	0.41	0	3388	0.00
Rob	1	6038	0.17	0	3256	0.00	1	4668	0.21	1	4552	0.22	0	3236	0.00	2	4049	0.49
Lilian	0	1317	0.00	0	759	0.00	1	1891	0.53	0	1317	0.00	0	3233	0.00	1	2247	0.45
Paul	0	1230	0.00	2	611	3.27	0	1679	0.00	0	1232	0.00	0	3134	0.00	1	2004	0.50
Elizabeth	1	1380	0.72	0	1025	0.00	0	400	0.00	0	1746	0.00	0	285	0.00	1	1829	0.55
Roy	0	1266	0.00	0	1185	0.00	0	278	0.00	0	1682	0.00	0	328	0.00	0	2192	0.00
Total:	**4**	**15118**	**0.26**	**2**	**9368**	**0.21**	**3**	**12807**	**0.23**	**1**	**13247**	**0.08**	**1**	**12643**	**0.08**	**5**	**15709**	**0.32**