INTERMEDIATE PAST BE: A PARADIGM RESHAPED WITH DATA DRAWN FROM HARES

Pro Gradu Thesis
Department of English
University of Helsinki
Simo Ahava
26 April 2010
TABLE OF CONTENTS

1 INTRODUCTION ................................................................................................. 3
   1.1 Foreword ........................................................................................................ 3
   1.2 The positive past BE verb paradigm ................................................................. 4
   1.3 On the problem of description ......................................................................... 7
   1.4 On the problem of analysis ................................................................................ 11
   1.5 Research questions and hypothesis .................................................................... 15
2 BACKGROUND .................................................................................................... 16
   2.1 Key concepts ..................................................................................................... 16
       2.1.1 Standard English ....................................................................................... 18
       2.1.2 Non-standard English .............................................................................. 26
       2.1.3 Corpus linguistics .................................................................................... 34
   2.2 Past BE variation ............................................................................................... 38
   2.3 Philosophy of variation ...................................................................................... 44
3 HELSINKI ARCHIVE OF REGIONAL ENGLISH SPEECH .......................... 49
   3.1 Overview .......................................................................................................... 49
       3.1.1 History ........................................................................................................ 51
       3.1.2 Data collection ........................................................................................... 55
       3.1.3 Orthographic transcription ........................................................................... 59
       3.1.4 XML annotation ........................................................................................ 61
   3.2 The Cambridgeshire sampler ............................................................................. 65
       3.2.1 Introducing Cambridgeshire ....................................................................... 66
       3.2.2 Informant selection ..................................................................................... 69
       3.2.3 Locality selection ....................................................................................... 72
4 IDENTIFYING THE INTERMEDIATE VARIANT ............................................ 74
   4.1 Method .............................................................................................................. 76
   4.2 Informant #1 – cam02 ..................................................................................... 78
       4.2.1 Overall distribution .................................................................................... 78
1 INTRODUCTION

1.1 Foreword

This study has been made in accordance to the following tenet of descriptive linguistics: “[The aim of description is] to give a comprehensive, systematic, objective and precise account of the patterns and use of a specific language or dialect, at a particular point in time” (Crystal 2003: 133; emphasis in the original). The logic behind such an approach is that all a priori grammatical rules pertaining to the language being studied should be considered potentially falsifiable, no matter how detailed or, in contrast, how general they are.

Here, I will take one such aprioristic category of English, the positive past BE verb, and show that spoken language data from a specific regional variety of English do not agree with how the verb is classified in Standard English\(^1\) grammar. Since this study is descriptive, I will refrain from making empirically unfounded claims about the internal mechanisms of this variation, because in doing so the focus would be away from objective description towards a more subjective interpretation.

This study pays homage to the famous verdict by Sapir (1921: 38): “Were a language ever completely ‘grammatical,’ it would be a perfect engine of conceptual expression. Unfortunately, or luckily, no language is tyrannically consistent. All grammars leak.” I intend to look at the philosophical and linguistic undercurrents of what this conceptual expression and variation thereof actually is, before verifying my postulations with data from a corpus of spoken regional English.

\(^1\) ‘Standard English’ (henceforth StE) with capital letters refers to the prestigious variety of English that is taught in the classroom and used in media and publications. Lower case ‘standard’ is used in its traditional meaning, e.g. ‘standard of education’. See also Chapter 2.1.1 for further discussion.
1.2 The positive past BE verb paradigm

The positive past BE verb paradigm (henceforth ppBE) features in an impressive body of research literature from a variety of English-speaking communities (see e.g. Milroy and Milroy 1993; Hazen 1998; Smith and Tagliamonte 1998; Tagliamonte 1998; Anderwald 2002; Britain 2002; Trudgill 2008; Richards 2010; see also Chapter 2.2). These studies focus mainly on the two variants *was* and *were* and how they are distributed in a representative sample of a specific speech community. However, data drawn from the Helsinki Archive of Regional English Speech (HARES), a collection of rural speech gathered in the 1970s and early 1980s, show that this two-way distinction is no longer the only way to approach this particular variable. Indeed, speech samples from a number of informants from the county of Cambridgeshire\(^2\), England, reveal that the preferred variant is actually an intermediate variant, represented phonetically as [\(\text{w} \emptyset\)]\(^3\).

This variant has been researched previously by Richards (2010), whose study in Morley (a suburb of Leeds) identified four different variants in the ppBE, with [\(\text{w} \emptyset\)] and [\(\text{w} \partial\)] treated as grammatically and phonetically distinct from each other and the two StE variants *was* and *were*. I fully concur with Richards’ statement that “careful consideration must be given to where the boundaries lie between the different variants” (ibid.: 79) and the fact that the salience of the binary treatment of *was*-*were* variation can easily lead the researcher to disregard evidence such as that presented in Richards (ibid.) and this study as flukes or as a result of phonetic reduction.

---

\(^2\) See Map 2 (Chapter 3.1).

\(^3\) ‘Intermediate variants’ come from Richards, who described them as occurring “in between the two standard productions, both in terms of perception (community members seem unclear as to whether these productions are intended to represent *was* or *were*) and production (a short, often centralised vowel and no syllable-final fricative)” (2010: 63).
In this paper, I posit that the intermediate variant is not a fluke. I also claim that it should not be treated as a phonetically reduced version of *was* or *were*, but rather as a grammatically and functionally distinct variant with which the ppBE of the speakers’ grammars should be appended. To support this claim, I will show that the variant has partly or completely replaced *was* and *were* in their respective contexts. Also, I will show that the variant is embedded in the grammars of the speakers by, for example, showing how the speakers’ use of the intermediate variant is resilient to the input provided by the interviewer (cf. L. Milroy 1987a: 31-35; 1987b: 178-183). In other words, in order to make generalisations about a phenomenon such as the intermediate past BE variant, both the linguistic factors and the actions and mindscapes of the language users themselves should be taken into account.

To relate the variation in the ppBE to a methodological and theoretical framework, I will introduce two related problems: (1) the **problem of description**, or how to describe spoken language in predominantly text-based media, and (2) the **problem of analysis**, or how to interpret grammatical phenomena of spoken language with tools devised for and refined in the realm of written language and especially StE. These two problems underlie the methodological discussion of spoken corpus compilation and the theoretical problematisation of grammaticality in spoken language, respectively.

As I will show, this paper will also serve as an addition to the millennia-old theoretical and philosophical debate that has surrounded variation and variability. The debate between empirical and theoretical linguistics in the 20th century has its roots in Antiquity, as philosophers such as Heraclitus, Parmenides, Plato and Aristotle played a significant part in developing the main strands of Western philosophy of science (see e.g. Seuren 1998). Similarly, what we know today as rationalism and empiricism
were largely conceived during the Enlightenment, when especially René Descartes
was instrumental in removing empirical data from the methodology of science – a
notion which was aggressively objected by empiricists such as John Locke and David
Hume. In modern times, the clash between empirical and theoretical linguistics
persists especially in Chomsky’s re-evaluation of Cartesian rationalism and e.g.
Weinreich, Labov and Herzog’s (1968) search for empirical foundations in linguistic
theory-making. I will briefly introduce the historical and philosophical basis for this
debate in more detail in Chapter 2.3, with hopes that this paper will be understood as
an addendum to this discussion, as my research questions and findings lean towards
placing importance on observation and empirical data rather than on abstract
reasoning and rationalism as emphasised by the generativists.

Finally, this study is grounded in the notion that linguistic description should
precede any aprioristic preconceptions that the researcher may have about any
specific language phenomenon. I agree with Haspelmath that the linguist should
“avoid fitting observed phenomena into the mould of currently popular categories”
and that “the linguist’s job should be to describe the phenomena in as much detail as
possible, using as few presuppositions as possible” (2007: 125). Furthermore, there is
a seed of truth in Richards’ assertion:

If we do not allow presupposition to cloud our view of what potential variability
is present, then, by allowing the data to speak for itself, we are able to further
illuminate ourselves as to hitherto uncharted waters of sociolinguistic variation.
(2010: 79)

Naturally, observations need to be described in some tried-and-true way, and
linguistic categories such as *verb*, *noun* and even the language-particular *was* and
were, relevant to this study, serve this purpose well, as long as the researcher is working with a language variety where these categories are firmly attested. However, as this study will show, even language-particular categories, firmly entrenched in the grammar of the standard variety, such as the past BE paradigm of English, can be reduced or expanded depending on the constraints that govern each individual speaker’s grammar and language use.

To summarise, this paper is about how people use language in everyday situations. It might differ from the norms dictated by prestigious varieties such as StE or it might not. What is crucial is that the ways in which the speakers’ grammars differ from or are similar to StE, for example, are not just random fluctuations in the speakers’ language use, but are actually highly structured and ordered (see e.g. Weinreich, Labov and Herzog 1968). It is the job of the linguist formulating the description to observe, describe and model this structure without either falling prey to pressures of fitting these observations to pre-established models of language or discarding the evidence altogether in favour of more abstract levels of reasoning.

The two following chapters of this introduction explain the problems of description and analysis with more detail, because as stated above, they serve as the fundamental backbone of and motivation behind this research paper. In Chapter 1.5, I will lay out my research questions with a recapitulation of my hypothesis and thesis statement.

1.3 On the problem of description

The problem of description can be explained through Halliday’s assertion: “spoken language is not meant to be written down, and any visual representation distorts it in
some way or other” (2002: 21). Describing spoken language in writing is a problem that underlies all efforts taken by linguists working in the field or with spoken language corpora.

One of the main concerns with confining the mechanisms of spoken utterances in writing is that both the corpus compiler and the end-user might be fooled to think that an exhaustive description of a speech situation could be achieved by using audio alone (see e.g. Adolphs and Carter 2007: 134). Not only does the oral-textual interface exclude such important features of spoken interaction as e.g. facial expressions and body movements, but it does not come even close to appreciating the complexity of shared knowledge between participants in a conversation. Bloomfield describes this complexity as follows: “the study of speakers’ situations and hearers’ responses…is equivalent to the sum total of all human knowledge” (1933: 74) and “[t]he situations which prompt people to utter speech, include every object and happening in the universe” (ibid.: 139). Even though Bloomfield “apparently overlooks the empirical significance of the fact that ordinary speakers handle meaning with fairly modest stores of knowledge” (Beaugrande 1991: 62), the notion of an infinitely complex history behind every speech situation is realistic and can be used to infinitely problematise the description of speech situations.

In this paper, I will tackle the problem of description by introducing the Helsinki Archive of Regional English Speech (HARES). Transcription protocols (that is, guidelines for writing down what the transcriber hears) and annotation schemata (that is, what metadata or extra data are included in the transcriptions) for HARES have been designed with a singular goal in mind: the audio must remain the primary data. However, HARES development has seen its fair share of problems, most of which are very common to spoken corpus compilers (see e.g. Allen et al. 2007: 22). A
common denominator to these problems is that even with modern advances in technology, spoken corpus usage is still transcription-dependent (cf. Moore 1999; Halliday 2002; Adolphs and Carter 2007). This means that in order to gain access to the actual data, the audio, corpus users must use an interface that is based on text input.

Furthermore, since the transcriptions are written texts, it would mean that the researcher will most usually work in the order shown in Figure 1. Even though the listener has access to the audio, the actual order of work will, due to the fact that search and query interfaces are based on written language, be from written transcriptions to audio and back. Nevertheless, including the actual audio data in the corpus increases verifiability of data and reduces the responsibility of the transcriber to arrive at conclusions and formulate analyses that might be misinterpretations or that might mislead the end-users.

Indeed, it is this "transcriber's interpretative power" (Moore 1997: 352) that has proven to be most harmful to research on spoken language. Moore (ibid.) goes on to say that "once [the transcriber] transforms the dialogue into a written transcript, nobody bothers to consult the audiotapes". Though this is probably true in many cases, advances in technology allow us to integrate audio and text so that locating the actual data is no longer a strenuous effort but can be done concurrently with the corpus browsing.

Since transcriptions remain today the preferred way of approaching the audio, it is necessary to adopt a single, uniform set of transcription protocols. For HARES, the protocols adopted are akin to those used in the Newcastle Electronic Corpus of Tyneside English (NECTE; see e.g. Beal et al. 2007). The NECTE Orthographic Transcription Protocol (OTP) is based on the principle that regional spellings should
Figure 1. The (traditional) order of work with spoken language corpora

be translated to StE, in order to do away with the use of ‘eye-dialect’ which more often than not leads to “unintelligibility and caricature” (Preston 2000: 617). However, since regional speech includes a plethora of features not identified in StE (Labov 1972a: 4), this principle does not come without its share of problems.

Even though adopting the NECTE OTP for this project might seem like an easy decision, it is not. It is a lamentable fact that no consensus exists as to "what form transcripts should take" (Moore 1999: 353). Because the objective with HARES has been to treat the audio as primary data, we are comfortable with leaning towards StE orthography in our transcription work. Thus the role of annotation (which includes time-aligning the text with the audio) is of increased importance, and we can now take further steps to remove "the prop provided by writing", which de Saussure and

---

4 “[R]espellings which reflect no phonetic facts whatsoever, such as 'sez' for 'says' or 'wuz' for 'was'.” (Preston 2000: 615).

5 I.e. the HARES team (including the present writer; see Chapter 3.1).
his contemporaries found to be "deceptive, [but] still preferable" (1916: 32) in speech representation.

In Chapter 3, I will introduce HARES and the compilation and annotation principles that have been adopted in the project. The intermediate positive past BE variant will be used as an example of the problems the transcriber and spoken corpus compiler will have to ultimately resolve. Indeed, the difficulty of introducing a new item into the grammar of the speaker, and thus into the corpus where the speech is conserved, binds the problems of description and analysis together. Without reaching a well-founded conclusion about the grammaticality and distinctiveness of the intermediate variant, there is no justification to transcribe it in a way that would highlight its special quality.

1.4 On the problem of analysis

How does one, then, arrive at the conclusion that a specific phenomenon is grammatically distinct? Furthermore, what does ‘grammatical’ mean in the context of spoken language? If we resort to published grammar books (such as Biber et al. 1999), no matter how far we look, we will never find the intermediate variant described in them, since it is not a commonly accepted item in the inventory of StE grammar.

This is the problem of analysis. If we approach the data with pre-established categories, our only task being to describe the researched variety with a set of concepts that have been well established in the standard language, we accept the underlying notion of universality of these categories (Haspelmath 2007: 121) and fall short of making a refined and intricate description. If we accept the universality of
categories, we also accept the fact that we are idealising our object of study, since the representations we find attested in our actual data can be reduced to their underlying, universal syntactic representations (Chomsky 1965: 28-29; see also Firth 1968: 16). The level of abstraction that this idealisation introduces is useful when designing typologies and comparing different languages, but it does so by limiting the detail one can put into description.

Naturally, working on varieties of English has less striking implications, since most of the underlying syntactic categories are shared by StE and non-standard varieties (Nevalainen & Tieken-Boon van Ostade 2006: 291). However, accepting this conclusion as aprioristic doctrine remains the bane of empirical studies of English dialects. Consider the following scenario: a linguist who accepts that language description should be based on pre-established categories one day happens to find the variant [wɔ] extremely frequent in his data. Because he considers English to have two well-established variants in the ppBE, *was* and *were*, he immediately starts thinking of which of these two is this phonetically reduced version an allomorph. If he remains true to his approach, he would base his conclusion on category-defining rules and consider the variant a reduction of *were* in sentences with, for example, a plural subject and a form of *was* in sentences with a singular subject. All non-grammatical sequences he would attribute to errors relating to performance (see Chomsky 1965: 4; Weinreich, Labov and Herzog 1968: 125; Halliday 1973: 53) and thus discard as irrelevant.

But what if the speaker whose speech the linguist is studying frequently uses the full form of *were* both with plural and singular subjects? Shouldn’t this observation require a revision of the category-defining rules? The linguist remains adamant. No, he says, this kind of heterogeneity cannot be thought of “as a subject of
systematic description, but as a kind of tolerable imprecision of performance” (Weinreich, Labov and Herzog 1968: 121; see also Labov 1972b: 200). As should by now be obvious, discarding these features of performance as ungrammatical or unfitting to the underlying category system blunts the edge of linguistic description.

In examples (1a-c), the various interpretations of the intermediate variant are described. (1a) is what the traditional linguist-grammian would consider when coming across the intermediate variant. He is driven by the need to categorise phenomena according to his own native-speaker intuitions which, in turn, are based on the rules and constraints of StE grammar. His “judgements,” as Adger and Trousdale put it, “are informed by explicit teaching or implicit reinforcement of norms” (2007: 265). In example (1b), the dialectologist, an expert on the area of Cambridgeshire in particular, has categorised the intermediate variant as a non-standard use of the past BE verb. Her intuitions are based on knowledge of the regional variety spoken in the area and the speaker’s tendency to use non-standard language, in turn based on quantification of the data at hand (see e.g. L. Milroy 1987a: 10). Finally, in example (1c), we have the descriptivist, who has opted to make no judgement as to what pre-established category the variant belongs to, but has decided to craft a new spelling for it. He does this because he is unsatisfied with both the traditional approach and the dialectologist’s approach to the problem. Further research, he thinks, will resolve what the ancestor of this variant is (in terms of language change) and how its distribution is similar to or different from the other variants in the past BE verb paradigm.

1. a) He was there
   b) He were there
c) He *wa* there

The kind of conservatism displayed by the linguist in the hypothetical scenario above and in example (1a) is what dialectology (the systematic study of dialects; see also Chapter 2.1.2) has rallied against. Chambers and Trudgill give a caricature of the (early) division between general linguistics and dialectology:

> At its worst, there has been a kind of mindless friction between the two groups, with the dialectologists scorning linguists as ‘abstractionists’ who deal in ‘hocus-pocus’ rather than real language data, and the linguists dismissing dialectologists as ‘mere butterfly collectors’ who get so entangled in the bushes that they cannot see the trees, let alone the forest. (1980: 17)

Though this rift has become smaller in the more moderate climate of the late 20th century (ibid.), the fact remains that descriptive and general linguistics part ways when it comes to the question of how to deal with empirical findings such as the intermediate variant described in this study. Even though idealisation carries benefits, especially because it precedes the crafting of typologies from language-particular inventories, I stand by Haspelmath when he notes that working with universals seems “to distract descriptive linguists from their more urgent business, that of describing languages in a way that is as complete as possible” (2007: 125).

The problem of analysis, then, is closely related to linguistic description. We could, for example, discard structural analyses by adopting a purely Boasian6 approach, or we could discard empirical evidence by adopting a universalist

---

6 “Boas…and his heirs found the categories in [Native America] languages to diverge so radically from the Standard average European languages that they taught their students not to make any assumptions about the categories in terms of which the language should be described.” (Haspelmath 2007: 122).
approach\textsuperscript{7}. The theoretical stance I take towards the problem of analysis in this study is in neither extreme, but rather somewhere in the middle. I shall argue that by adopting a pre-established category (the positive past BE verb) but redefining it with the empirical evidence at hand, a sufficiently realistic description of this particular phenomenon will be achieved.

However, I lean towards the importance of exhaustive description in my methods (Chapter 4), as I will analyse the variant from a number of perspectives. I shall acknowledge the frequency of the variant in various linguistic contexts (see e.g. Milroy and Milroy 1993: 19), its occurrence in different sentence stress positions and the effect of previous discourse on its emergence. It is my firm belief, and my results agree, that \[wp\] is not \textit{was} or \textit{were}, but a separate, fully functional variant in the ppBE. By avoiding any overt generalisations on the basis of my findings, thus restricting them in space and time to only the selected speakers and their utterances, I will show that the problem of analysis, in this study, is satisfactorily resolved.

1.5 Research questions and hypothesis

In my paper, I ask the following:

- Is there an intermediate variant of the positive past BE variable?
- How should grammatical ‘anomalies’ such as these (from a StE point-of-view) be taken into account when transcribing and annotating a spoken language corpus? (The \textbf{problem of description}.)

\textsuperscript{7} It must be noted here that neither of these extremes can in reality be adopted. A descriptive approach would still have to be based on observation, which, in turn, “is always selective,” as Popper writes. “It needs a chosen object, a definite task, an interest, a point of view, a problem.” (2002: 61.) Similarly, linguists looking for and believing in universal, pre-established categories of language have yet to show conclusive evidence of their existence (see e.g. Haspelmath 2007).
Is this variant a grammatically distinct realisation with which the ppBE of the speakers should be appended, or is it an allomorph of *was* or *were*? (The problem of analysis.)

My hypothesis is that the intermediate variant does exist. As my results will show, it is a frequently observed phenomenon in the speech of a selection of Cambridgeshire informants. To fit this *non-standard* feature of language (see Chapter 2.1.2) into a corpus of spoken language, and thus overcome the problem of description, it is necessary to consider the audio as primary data and craft both the transcription protocols and the annotation schemata so that they are as transparent as possible (see Chapters 3.1.3 and 3.1.4). Finally, to reach a conclusion about the distinctiveness of this variant, it has to be analysed from a variety of perspectives (see Chapter 4): how it is embedded in the grammars of the speakers, what is its linguistic context, how previous discourse and pragmatic features affect its realisation and, importantly, what its actual distribution is in the corpus, that is, how frequently it can be observed.

2 BACKGROUND

2.1 Key concepts

Studies on regional language variation, such as this, have their origins in the need to understand why dialect speech frequently displays features that do not conform to normative and prescriptive rules dictated by a prestigious or standardised variety of the language. Indeed, dialectology has its origins in the late 19th century, when speech analyses of regional German varieties showed how certain sound laws that had been considered exceptionless were, in fact, not so (Chambers and Trudgill 1980: 17). This chapter will illuminate a number of key concepts that each deal with variation in their
own terms, utilising their own terminologies to uncover theoretical and methodological solutions to the problem of describing variability in language and in the world around us.

To understand what the situation is like in the sphere of English, I will introduce *Standard English* (StE), the prestigious variety of English, and *non-standard English*, a collective term for those varieties that differ in some way from StE. I will show how their classification is difficult but necessary, because not only is the literature contradictory when it comes to explaining just when a variety is non-standard enough to be called a dialect and whether or not StE actually exists, but also because an understanding of the one is necessary in order to understand the other (see e.g. Taavitsainen and Melchers 1999; Upton 2000).

As a key methodological concept, I will introduce *corpus linguistics*. The introduction of corpora in the 1960s provided impetus for language studies to harness impressive bodies of data that were easily accessible for research purposes (Svartvik 1992: 8). Defining corpus linguistics is important here, because this study both introduces the compilation of a corpus of English language (see Chapter 3) and uses it as the tool with which to collect the language data within the corpus (see Chapter 4). Though I agree with Chafe that corpora are “an absolutely crucial part of the linguistic enterprise” (1992: 79), they remain tools and should be treated as such. Overwhelming corpus evidence might lead the researcher to draw categorical conclusions that reach beyond the range of data in the corpus itself, and it is because of this that a certain amount of criticism and even scepticism is warranted when using corpora in language studies (see e.g. Chomsky 1965: 15, 202; L. Milroy 1987b: 144).
2.1.1 Standard English

This study is based on language samples taken from speakers who have been traditionally classified as speakers of non-standard English. To open up the plethora of meanings that are associated with *non-standard* (see Chapter 2.1.2), it is necessary to introduce Standard English (StE), as a definition of the former can only be uncovered by defining the latter. Furthermore, for the purposes of this particular study, defining StE and the regional varieties of English is relevant, because appending the past BE paradigm with a new variant, as I have set here to do, shows that StE with its grammatical rules and structures does not govern everyday conversation with an iron fist, but allows for deviations to exist in the form of regional, non-standard varieties.

In this chapter, I will focus on the attitudes that surround StE by introducing both the historical development of the variety and how its increasing prestige resulted in the mid-20th century, especially after Chomsky’s formulation of the generative hypothesis, devaluation of descriptivist studies of non-standard English (see e.g. Trudgill and Watts 2002: 1). The notion that non-standard English is somehow inferior to the prestige varieties is not limited to the sphere of academic discussion, but extends to the general public, too. The “popular mentality which views regional dialects as structurally or expressively inferior to the standard language” (Crystal 2002: 243) has been omnipresent in the past, but recently a heightened interest in local history and language has increased positive interest and respect towards non-standard English. The goal of this discussion is to provide an account of why and how StE achieved its exceptional status in English-speaking cultures today. The following
chapter will focus on the implications the rise of StE has had on the non-standard varieties.

First, consider the following two classifications of StE, provided by Adger and Trousdale (2007: 264-265):

- “One approach is to consider the standard as a variety of English like any other, one which is spoken (at least on occasion) and written by a group of speakers.”
- “The contrasting approach considers the standard as an ideology...a set of beliefs shared to varying degrees by speakers and writers of English.”

As this study is about non-standard speakers and their linguistic independence (apparent in, for example, how they reject StE structures provided by the interviewer\(^8\)), my focus will be on the attitudes that surround StE, and thus I will not introduce the complicated debate of whether StE actually exists as a spoken variety or not\(^9\). This is why, for this study, I consider the second approach more convenient than the first. Also, locating StE especially in the sphere of spoken language is more difficult than one would think. From its beginnings in the 15th century, the standardisation of the English language has been directed towards the written medium (see e.g. Taavitsainen and Melchers 1999: 8). This has resulted in the variety evolving into a normative standard – an ideology (as Adger and Trousdale put it above; see also Crystal 2003: 431) that "exist[s] at a high level of abstraction" (J. Milroy 2000: 11). This abstract quality of StE does not cater to the requirements of spoken language research, because the ephemeral and variable quality of a single utterance of

\(^8\) See Chapter 4.
\(^9\) See e.g. Trudgill (1999b), J. Milroy (2000) or Upton (2000) for more on this particular debate.
spoken language (see e.g. Weinreich, Labov and Herzog 1968 for an account of variability in language) makes it difficult to adopt a hard-and-fast, standardised (and thus homogeneous) system for the description of any spoken language variety. However, this is a failing only on the part of the linguist formulating the grammatical description. The fact that spoken language varieties defy the imposition of a standardised set of norms and rules does not mean that there are no underlying constraints:

Anyone who has taught Linguistics 101 will attest to the fact that one of the most difficult concepts to convey to neophytes is that all languages and language varieties are regulated. ‘The language’ has rules which are fully instantiated in ‘good language’; ‘ordinary language’ chooses to ignore a select few of these, although one might suspect that those ignored at this level are regarded by many speakers as ‘unimportant’ ones. (Preston 2002: 147)

StE is not a monolithic variety with clearly stated principles and underlying rules governing its use and development. Its norms and rules are derived from *de facto* representation, evolving concurrently with the content of e.g. dictionaries and grammar books. Indeed, “English has never had an official regulatory body” (Poplack, Van Herk & Harvie 2002: 88) which would supervise and dictate the terms of its use. However, this does not mean that StE would be on weak footing. On the contrary, StE is the variety taught in the English language classroom, printed in media and literature and approximated in speech by people in the middle and upper classes of society (see e.g. Trudgill 1999b: 18). The regional origins of the variety are often

---

10 Actually, Crystal notes that the non-standard varieties are the ones that have been considered monolithic in traditional accounts and that StE is, in contrast, “manifested in many different varieties” (2002: 243).
overlooked, since StE of today does not have a regional standing but is rather a social variety. The social aspect of StE brings us back to its normative and ideological standing. Because an important facet of language standardisation is the elimination of structural variance\textsuperscript{11} (J. Milroy 2000: 13), an inevitable result is that the standard becomes a variety that not only no one speaks, but that no one can speak (ibid.). Thus the dimensions of StE and non-standard Englishes are rid of the dichotomy that would exist between well-established and distinct varieties of language. In other words, we are reminded of the high-level of abstraction that labelling varieties calls for. Instead of distinct, traceable borders between the different systems, there exists "a complex parameter, with several different, and admittedly subjective ratings along the scale" (Taavitsainen and Melchers 1999: 12).

However, there was a time when StE had a more tangible status than purely ideological. The standardisation of the English language began in the 15\textsuperscript{th} century\textsuperscript{12}, but "maximum" standardisation was not achieved until the 18th century (Milroy and Milroy 1993: 9-10). Standardisation was both "a unifying and levelling force" (Taavitsainen and Melchers 1999: 9) and emphasised the solidification of the writing system by eradicating the various variant spellings that existed in the lexicon (see e.g. Rissanen 2000: 117). The variety that became the source and, in a sense, the target of standardisation was the one which had standing in politics, culture and power:

What the standardisation process has in fact led to is the creation of a relatively focussed variety of the English language which is used as the written medium

\textsuperscript{11} Also, Rissanen (1999: 190) sums up that the process of standardisation "is characterised by a trend towards invariance, deregionalisation, depersonalisation and prestige".

\textsuperscript{12} Rissanen (1999: 190) notes that certain "standardisation tendencies" emerged already in the Old English period, though with little impact on the StE of today, and that the actual standardisation of English has its roots in the "reintroduction of English as the language of administration and literature from the fourteenth century onwards".
and as a medium felt to be appropriate to formal contexts. That this variety had its basis in the East Midland dialect is a matter of historical and geographical accident. If English kings had continued to travel around the country, or if they had settled their court elsewhere in the country, another Middle English dialect might well have been selected. (Nevalainen and Tieken-Boon van Ostade 2006: 287)

Milroy and Milroy (1993: 10) see the emergence of a nation-wide standard as foreshadowing the devaluation of other regional varieties of the language (see also Upton 2000: 72). This is not entirely surprising, since for example StE, considered today as the most prestigious variety of English, has its roots in a variety that already had great value attached to it. Also, the "language of officialdom is, by nature, conservative", and it "marks prestige variants as elements of the standard" (Rissanen 1999: 191). This apparent discrimination of the less prestigious varieties should not, however, be seen as deterministic by nature, because the non-standard varieties that were not part of the standardisation process have continued to persist over the centuries (cf. L. Milroy 1987a: 33), especially when moving further away from the metropolitan area of South-East England.

However, the abstract and ideal quality of StE and the fact that it is equated with the written medium, and thus seen as less susceptible to change than the spoken varieties, and the social standing it has among the upper classes of society mean that it is considered prestigious and, consequently, as the ‘only correct’ variety of English (see e.g. Milroy and Milroy 1993: 14; Wales 2002). To idealise StE brings forth the question of what status all the other, non-standard varieties of English hold (Upton 2000: 71). Non-standard varieties of English have long been burdened by the sole legitimisation of StE (see Wales 2002: 45):
The selection of one variety as the standard variety and the diffusion of this variety through codification and prescription establish it as the canonical variety. This leads to a sense of legitimacy of this variety. The manuals of usage are effectively law codes, and using non-standard forms is analogous to disobeying the statues enshrined in law codes. (J. Milroy 2002: 8)

Defining the 'illegitimate' non-standard varieties has been difficult and loaded with terminology that can be interpreted as pejorative or, even at best, dependent on comparison with StE (e.g. Upton 2000: 75; Larroque 2005: 75). The pejorative connotations arise from describing the non-standard varieties further as something sub-standard (Petyt 1980: 7; Upton 2000: 71). The original thought with such an interpretation is the prescriptivist view that non-standard varieties are "in some way deviations from or corruptions of Standard English" (Trudgill 1999a: 13).

Furthermore, the term non-standard already invites a comparison between the non-standard variety and StE. In its starkest sense, this would mean that studies of regional varieties will always be burdened by the necessity to substantiate the research by comparison between StE and the variety under study. The notion of non-standard varieties as anything but equal and distinct varieties of a language has, thankfully, been in the decline over the last decades, and the general view today, applied throughout the present study too, is that

all varieties of a language are dialectal varieties. Differences are of form, status and function. Thus the stuff of dialectology is the study of any and all varieties, standard and non-standard, which go to make up the language. (Upton 2000: 71)
Upton’s view represents a positive trend that has developed in the field of language studies since the late 20th century. Dialects and, in fact, the entire rural way of life have witnessed a rebirth of interest among the general public (as evident from the popularity of projects such as BBC Voices¹³) and within academia. This is partly because StE has over its long history secured a strong foothold as the sole legitimate variety of English, permitting an unthreatening shift of focus towards non-standard English:

As the status of the standard variety is at this point rather secure, it becomes acceptable...to inquire into the histories of other varieties...[which] thus become legitimate parts of the history of the language as a whole, and, although they are given a subsidiary place, they can be seen as enriching that history. (J. Milroy 2002: 15)

Idealism put aside, it is an undeniable truth that unbalance in the portrayal of StE and non-standard English exists and that it stems especially from the education system of Britain (Taavitsainen and Melchers 1999: 5; see also Wales 2002: 62). The problem with pushing StE forward as the only legitimate variety of English is that it gives an unrealistic picture of how the language is actually used. Though arguments for the centrality of StE in education revolve around consistency, it is potentially harmful especially for the EFL (English as a foreign language) learner not to be introduced to the kind of dialect mixture that native English speakers would intuitively master (ibid.: 7).

¹³http://www.bbc.co.uk/voices/
The centrality of StE becomes apparent in academic research as well. Especially after Chomsky’s vocal support for idealisation and abstraction, “researchers have tended to ignore variation that challenged this idealization” (Adger and Trousdale 2007: 262). Though the Chomskyan approach has been challenged from its conception (see Firth 1968: 130; Weinreich, Labov and Herzog 1968: 125), it has still been embraced, at least partly, by the linguistic community, because it explicitly states abstraction as necessary for crafting a general theory of linguistics (Chomsky 1965: 6, 28). Since all treatments of language necessarily involve varying levels of abstraction, Chomsky’s legitimisation thereof was liberating to the scholarly community that followed in his wake. The level of abstraction that Chomsky proposes is, not surprisingly, the researcher’s own native speaker judgement that is, in turn, based on StE (see e.g. Adger and Trousdale 2007).

To summarise, the difference between StE and the non-standard varieties of English is “closely related to the distinction between written and spoken language” (Kerswill 2006: 34). In its extreme, this dichotomy has fuelled the views of prescriptivists, who hold that any varieties of English that do not adhere to the grammar and rules of StE should be considered somehow vulgar (Peitsara 2004: 66). The problem lies in the fact that StE is the only variety of English that has its own methodology and theory of grammar. Furthermore, its depiction in the educational system brings about an unrealistic model of the actual linguistic situation in native English speaking communities.

Indeed, a truer description of the English language should, by default, include a treatment of non-standard language use, too, since deviation from StE is not an exception to the native speaker – it is the rule.
2.1.2 Non-standard English

As all people have their unique way of living, working, walking and smiling, so do they have a unique way of speaking. Everyone speaks their own personal variety of their language. Language is thus heterogeneous by nature, but on some levels these differences converge to form mutually intelligible varieties. For example, in the case of English, these convergences can have many different guises: regional varieties, social varieties, accents, slangs, jargons and so forth (Peitsara 2004: 66-67). Even though the others play an important role in bringing order to the seemingly infinite amount of idiosyncrasy within the language, it is the regional varieties that have been without doubt the most significant categorisers of similarity within the sphere of English and its speakers, and it can be said that all speakers of English as a native language are speakers of some regional variety (Trudgill 1999a: 2).

In this chapter, I will first introduce dialectology\(^{14}\) in general before moving on to a more particular look at the situation within the sphere of English. As with StE, the many definitions of non-standard, regional variety and dialect are coloured by strong attitudes towards the value of each speaker's personal variety and that of the standard (Upton 2000: 70). It is important to recognise and focus on this idiosyncratic language use (idiolect), because variation on the individual’s level is prioritised in this study. The more one wishes to draw parallels between the idiolects of speakers from e.g. the same geographical region, the more one has to impose some abstract method of categorisation (such as by using terms like dialect or Cambridgeshire speech) on the data.

\(^{14}\) “The systematic study of all forms of dialect” (Crystal 2003: 136).
However, accepting a level of abstraction where a convergence is observed and labelled as a *dialect* or *regional variety* is necessary. Even though "[t]wo individuals of the same generation and locality, speaking precisely the same dialect and moving in the same social circles, are never absolutely at one in their speech habits" (Sapir 1921: 147), we can agree that the speech of the observed individuals has enough similarities that they can be grouped together as long as the research question is designed to accommodate for this kind of idealisation. Ihalainen commented on this ‘agreement principle’ in the following way:

> By calling my informants speakers of the Somerset dialect I simply mean that they were born and bred in Somerset and that their speech shows a number of common features frequently heard in Somerset. Furthermore, those features clearly distinguish my informants' language from Standard English. (1980: 187)

In this study, I feel comfortable enough to use a number of very different speakers to provide speech evidence for my claim that the intermediate variant exists in each speaker's own grammar. I can make this claim (see Chapter 1.5) because my argument is that the deviation from StE is potent enough for the intermediate variant to be classified as a feature of non-standard English use.

The “varietal hierarchy” (Upton 2000: 73) that originates from the machinations behind the standardisation of a prestige dialect is something that dialectology as a discipline revolves around, and it is this uneven valuation of different varieties that dialectology has often rallied against. Historically, the first dialectological effort has
often been attributed to Georg Wenker’s 1876 survey\textsuperscript{15}, which undermined the Neogrammarian hypothesis of exceptionless sound laws\textsuperscript{16} with data collected in the German countryside. Wenker's data revealed a heterogeneity that left no question about the misplaced faith in absolute rule-governed sound laws that the Neogrammarians opted for (Chambers and Trudgill 1980: 17).

The origins of dialectology are thus in dialect geography: the collection of vast quantities of data that provide evidence for differences between speech patterns across geographical regions. Dialect geography also provided comparable material that was of historical value, a means of evaluating the doctrine of the regularity of sound change, and information that might enable us to learn something about the nature of linguistic boundaries. (Shorrocks 2000: 89)

Furthermore, it removed language from the mind of the armchair linguist and placed it where it ought to be: nested in the minds and mouths of the speakers themselves.

At the heart of dialectology is thus the effort to examine speech and mindscapes of speakers from a specific region and to study and document how idiosyncratic features in their speech converge into larger units. Though this is seemingly clear, it must be remembered that convergence is something that stems from the mind of the analyst, not the speaker. As observer-analysts, “we must arbitrarily throw whole masses of experience together as similar enough to warrant being looked upon –

\begin{flushleft}
\textsuperscript{15} Though Vasko (2005: 28) notes that “[t]he first classification of English dialects was in fact published by His Imperial Highness Prince Louis Lucien Bonaparte (1875-1876)”. She continues to remark that the philologist A. J. Ellis collected data on English dialects during this time, too.
\textsuperscript{16} Though, as Shorrocks notes, “[t]here is actually no proof in Wenker’s writings that such a test [i.e. refutation of the Neogrammrian tenet] was in his mind when he founded the \textit{Deutscher Sprachatlas}” (2000: 89).
\end{flushleft}
mistakenly but conveniently – as identical,” Sapir writes, "in spite of great and obvious differences” (1921: 13). Thus it is the linguist who ultimately makes the decision that it is legitimate to make generalisations about the speech patterns found in the data.

A dialect, then, is an abstraction, where speakers and speech patterns in a heterogeneous system are grouped together according to some category of classification. So on the one hand we approach the “Saussurian notion of langue as an object of uniform social understanding”17 (Labov 1972b: 192; emphasis in the original) and on the other we try to retain the exceptional, individual and heterogeneous quality of each utterance studied.

Working with abstractions not only accentuates the dangers of unruly induction18, but also makes defining concepts attached to the process increasingly difficult. The category of classification mentioned above, whether it is dialect, regional variety or some other, is one such concept that eludes a steadfast definition. As Petyt notes,

\[\text{... But somewhere we call a halt to this concentration on differences, and decide that among a certain group there is an important degree of linguistic unity – that its members speak the same dialect. (1980: 12; emphasis added)...}\]

This abstract nature of regional varieties means that we can study them in different scales, ranging from, for example, a larger geographic region (e.g. Devon) to a single locality (e.g. Sampford Peverell) and even to the individual speaker. The deeper one

---

17 See also Saussure (1916: 13, 15).
18 As Popper writes in his evaluation of Hume's critique of induction, “theories can never be inferred from observation statements, or rationally justified by them.” (2002: 56). See also Chomsky (1965: 15) and Beaugrande (1991: 361).
delves into this structure, the more intricate the differences become and thus any comparative analysis becomes burdened by the vast amount of inconsistencies.

Furthermore, the lack of consensus on precisely what the level of convergence has to be for a variety to be called a dialect (Francis 1983: 1) means that dialectological studies have always relied on and been burdened by some artificial categorisation of the speech data. Since it is impossible to find a universal and thorough description for the amount of linguistic markers and historical and social growth required in a speech community for its variety to be called a dialect, it means that when we talk about particular dialects, we are, in a very real way, relying on an abstraction of the true nature of the variety (Petyt 1980: 12-13). For example, when this study describes features of the Cambridgeshire variety, it is, in fact, describing a variety that occurs in the speech of the selected informants within the boundaries of the county and, importantly, only in the corpus sample the data are drawn from.

Especially in English dialectology, the StE variety plays an important role in the analysis of the non-standard regional varieties. StE affects the speakers’ personal way of communicating in the sense that in some situations they strive to follow the rules and norms dictated by this more ‘prestigious’ variety. In languages such as English, with a very strongly established standard form, there might be people who attempt to take the StE variety as their own, either to disguise any particular regional connotation in their speech or to maintain affinity towards higher social classes (Trudgill 1999a: 5).

The connection between StE and the higher social classes comes from the thought of StE as being something of an ideal – something that all English language users should adhere to. There has always been a connection between prestigious language use and social climbing. As Watts observes, "the acquisition and use of
Standard English appeared to guarantee social climbers in the eighteenth century access to the world of politeness" (2002: 155). Furthermore, the prestige of StE makes it easy to discard any and all forms not aligned with its grammatical rules as ‘incorrect’, ‘wrong’ or ‘corrupt’ (Trudgill 1999a: 13). Preston articulates a similar concern in the following way:

[C]ertain linguistic features which occur in the speech of the best-educated, highest-status speakers from some regions or from some ethnic groups are regarded by the general public as nonstandard, although nonlinguists prefer such terms as substandard or simply 'bad English' or 'sloppy speech'. (2002: 134; emphasis in the original)

However, the stark reality is that non-standard English frequently dispels prescriptivisms dictated by grammarians and proponents of StE (see e.g. Labov 1972b: 188; Shorrocks 2000: 95-96), who consider it as the only correct form of the language. So as the speech of English language users is constantly coloured by non-standard forms and structures, does this mean that the speakers are ignorant of the correctness of English? If so, why do the speakers themselves tolerate their own, faulty use of language? To provide an example, consider the following utterance spoken by an informant from Cambridgeshire:

2. we was more happier <pause/> cos do you know we was more contented and we was more happier than they are today (CP, BARTLOW, CAM)\(^{19}\)

\(^{19}\) See Chapter 3.1.3 for a description of the Orthographic Transcription Protocol used in HARES and in this study. See Chapter 3.1.4 for a description of the XML tags used in HARES and in this study. The text in the brackets denotes the initials of the speaker, the village the speaker was interviewed in and the county identification code (CAM = Cambridgeshire, DEV = Devon, ELY = Isle of Ely, ESS = Essex, LAN = Lancaster, SOM = Somerset and SUF = Suffolk).
This passage contains, as evident in StE grammar, incorrect verb agreement (“we was”), incorrect adjective comparison (“more happier”), unnecessary repetition (“we was more happier” – “we was more contented” – “we was more happier”) and, on a phonological level, sound-dropping (or aphaeresis) (“than” reduced to “‘an”, and the dropping of initial /h/ in “happier”). When confronted about his language use, the speaker would probably respect the difference between the forms he uses and those advocated by StE, depending on how well he knows the latter, but it is most probable that he would find little fault in his own language use, since for him it is his standard.

Indeed, at the very heart of linguistic inquiry lies the idea that language is a defining feature of the individual who uses it, and, as put forth by Upton,

Central to the social importance of dialectal speech is the plain fact that it is produced by people, and is used by people as their primary means of communication. It is the vehicle which carries everything from their most humdrum everyday messages to their most intimate confessions. As such, it is very important to them, so basic to their lives that they take it for granted for much of the time, but so essential also that, when others show and interest in it or threaten it, they can be roused to passion in its advocacy or its defence. (2000: 68)

This means that for English speakers it is more than familiar to consider their own variety as their standard of English. This subjective definition might seem irrelevant in a study that accepts idealisation as necessary, but the psychological allegiance that English speakers might consider towards their own variety is necessary to acknowledge. For instance, a speaker might, in an informal interview setting, reject the StE form elicited in a question by the interviewer, e.g. in the following excerpt
cited in Vasko (2005: 169; slightly modified to adhere to HARES transcription protocol):

3. Q: you didn’t have to go to Cambridge yourself very often
   S: I used to go Cambridge always (SC, LITTLE EVERSDEN, CAM)

This passage shows that even though the interviewer provides the grammatically correct prepositional element, “to Cambridge”, in his question, the informant answers using a non-standard English structure.

To summarise, taking the polarity of StE and non-standard Englishes as granted is a faux pas (Penhallurick and Willmott 2000: 8), because instead of clear-cut boundaries between regional varieties and the standard, we find the individual speakers with their own idiosyncratic ways of speaking. When we discuss dialects and regional varieties, or any other forms of convergence, we are relying on idealisation and choosing to see homogeneity in a heterogeneous system. So even though non-standard can be understood literally, i.e. not Standard English (see Chapter 2.1.1), the difficulty of defining the concept of StE and that of dialect proves how abstract the conceptual maze of linguistics actually is. To keep things simple, in this study I consider non-standard as a descriptive feature of the speech pattern I am studying. The intermediate variant in the ppBE of the speakers in this study is not Standard English, because, for example, it does not feature in the grammar book and the speakers themselves opt to use this form more often than those found in StE.
2.1.3 Corpus linguistics

In this chapter, I will introduce corpora and corpus linguistics. Since this paper is also a corpus study (see Chapter 1.3), it is necessary to explain the fundamentals of corpus-driven research. After a brief explanation of what a (computerised) corpus actually is, I will focus on the spoken language corpus, which, as I will show, requires a different approach from traditional, text-based corpora. Finally, I will give some food for thought about the pros and cons of corpus-driven research, with special focus on the pitfalls that working in blind faith with “authentic language” (see e.g. Fillmore 1992: 38; Svartvik 1992: 10) might uncover and how large quantities of empirical data can lead to unwarranted induction and generalisations (cf. Popper 2002: 55-61).

A corpus, according to Crystal, is a collection of LINGUISTIC DATA, either written texts or a TRANSCRIPTION of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a LANGUAGE (corpus linguistics). (2003: 112; emphasis in the original)

A corpus differs from an anthology or other collection of text (or sound) due to its primary application for linguistic research (Francis 1992: 17). Truly, computerised corpora brought about a revolutionary way of approaching vast quantities of linguistic data, enabling researchers and scholars to provide empirical foundations for generalisations that could not be made at earlier times (see e.g. Leech 1992: 106), mainly due to the fact that going over transcriptions and texts manually was far too laborious. Also, having large amounts of data at hand provided linguists the means to
“get the facts right” (Fillmore 1992: 38) by relying on attested language data rather than the intuitive knowledge of the analyst.

Most importantly, computerised corpora made it possible to carry out quantitative research, i.e. statistical calculations of, for example, the distribution of a particular linguistic variable across a speech community (Vasko 2005: 7). Indeed, it is in establishing the “large-scale frequency patterns” (Halliday 2002: 25) that forms the basis for understanding a living language. Leech (1992: 107), too, includes quantitative analysis as one of the four key features of (computer) corpus linguistics. The other three key features (ibid.) are

- Focus on linguistic performance, rather than competence
- Focus on linguistic description, rather than linguistic universals
- Focus on a more empiricist, rather than rationalist view of scientific inquiry

For spoken language corpora, that is, corpora that contain audio and text in the form of written transcriptions, these four key features are instantly identifiable both in corpus compilation and in actual corpus-driven research. The ability to mine multi-modal corpora that incorporate both audio and text has brought about a heightened sense of disparity between the introspective methods of the rationalists and the data-based methods of the empiricists (Halliday 2002: 26).

This verifiability of data that came along with audio-text-integration is of importance to studies focusing on spoken language, because, to quote Halliday, “spoken language isn’t meant to be written down, and any visual representation distorts it in some way or other” (2002: 21). Transcriptions of spoken language are only ever subjective interpretations made by the transcriber. To overcome the problems associated with working on data that someone else has, in essence, pre-
analysed and to minimise the danger of the researcher being misled by the “transcriber’s (possibly incorrect) decisions” (Miller and Weinert 1998: 12), it becomes necessary to include the audio samples in both the spoken language corpus and the research carried out with it. Svartvik has articulated this concern in the following manner:

The greatest risk of all, however, is the distance that may arise between the end user of a standard corpus and the primary textual material – a danger that is particularly imminent in the case of impromptu speech which has been recorded and orthographically transcribed by others than the user and where the actual audio-recording is not readily available or properly consulted. (1992: 10)

Thus, recent efforts in spoken language corpus compilation have included coming up with standards of transcription and annotation that would provide corpus compilers and researchers with a set of tools to be used on all corpora containing spoken language. One such tool, or combination of tools, is XML (Extensible Markup Language) with TEI (Text Encoding Initiative) specifications. This protocol has been successfully adapted in the Newcastle Electronic Corpus of Tyneside English (NECTE; http://www.ncl.ac.uk/necte), and it is because XML and TEI are well-equipped to handle digital information, such as that stored in a corpus, (Allen et al. 2007: 33) that they have been applied in the transcription of audio for the Helsinki Archive of Regional English Speech (HARES), the corpus used in this study.

As a final note, it is healthy to remember that corpora, too, no matter how large and how multi-faceted, are only ever tools that exclude far more than they include. Working with corpora of hundreds of millions of words might invite the analyst to draw conclusions that deceptively reach beyond the scope of the contents. Indeed,
theories and models, especially when based on empirical data, are only valid until one comes across data that refute them (Popper 2002: 260). With something as unconstrained as human language, this might happen every day. Another concern is the opposite. No matter how large a corpus is, one might never come across something that is intuitively clear as being a feature of the language studied. Fillmore agrees, saying that “[i]n the end, there is simply no way to avoid reliance on intuitive knowledge” (1992: 38).

Especially with spoken language corpora, it is also the case that one often stumbles across the word *authentic* used to describe the language data found within. Authenticity as a concept has many interpretations, but as Chafe notes, with spoken language corpora, we can be content that especially spoken conversation “can justifiably be taken as the use of language to which humans are best adapted and thus the one that can tell us most directly about inherent properties of language and the mind” (1992: 89). Even though a spoken language recording is far from comprehensive (and thus fully authentic) when considering all the intricacies of spoken utterances (see e.g. Adolphs and Carter 2007; see also Chapter 1.3), it is an adequate starting point when crafting linguistic descriptions.

To summarise, corpora brought *data* into the spotlight of linguistics. With the empirical foundations provided by vast quantities of oral and written data, linguistic analyses no longer had to rely purely on intuitive knowledge of the analyst and could now be formulated on actual language data. However, hypotheses that are formulated through introspective and aprioristic methods have not, will not and should not disappear from linguistic methodology. Over-confidence on data alone can invite linguists to draw unwarranted conclusions that are based on extrapolations of the data at hand.
Equipped with an understanding that a corpus contains data that have been specifically collected and filtered, that it is designated to testing hypotheses with vast quantities of attested language samples and that it cannot be used as an excuse to arrive at conclusions that cannot be verified with the data, the linguist can use the corpus to greatest satisfaction.

2.2 Past BE variation

In this chapter, I will briefly introduce previous discussion on variation within the ppBE. The purpose here is to emphasise the salience of this variable both in academic discussion and in actual, attested language data. The former is apparent from the sheer volume of research done on the variable and especially its non-standard use in English-speaking communities. The latter becomes clear from the range of data from “virtually every English speaking community to have undergone sociolinguistic observation” (Richards 2010: 62; see also Britain 2002: 17). I will not include a thorough review of all the previous research done on the variable, since the purpose of this chapter is to focus on Cambridgeshire and not the entire linguistic landscape of Great Britain. However, I will pick out a number of relevant papers, mainly Britain (2002), which outlines past BE usage in East Anglia (which borders eastern Cambridgeshire), Vasko (2005) and (forthcoming), where the ppBE in Cambridgeshire dialect grammar is introduced in more detail, and Richards (2010), which introduces the intermediate variant in Morley English.

---

20 See e.g. Milroy and Milroy (1993); Hazen (1998); Smith and Tagliamonte (1998); Tagliamonte (1998); Anderwald (2002); Britain (2002); Trudgill (2008); Richards (2010); Vasko (forthcoming). The previous articles all focus on Present-Day English. For a look at earlier development of the variable, see e.g. Nevalainen (2006).

21 See Chapter 3.2.1 for a more detailed look at whether or not Cambridgeshire should be considered a part of the East Anglia region.
To begin with, a general claim about past BE variation in regional varieties, also exemplified by the results of this study, is that it portrays a typical case of regularisation. In this paper, I consider regularisation to mean the loss of irregular distinction in the paradigm (cf. Britain 2002\textsuperscript{22}, where the phenomenon is called simplification). In the case of ppBE, this would mean that the speaker shows a tendency to use only was or were (see Table 1) in contexts where the past BE verb is expected.

Anderwald notes that the loss of person distinction in the past BE verb paradigm should not be regarded as regularisation, since the term, for Anderwald, implies a completed process, where distinction is altogether abolished (2002: 10-11). She prefers the use of the term generalisation to depict the “system of asymmetries that diverges greatly from the standard” (ibid.) Naturally, I agree with her in that regularisation to was or were is not consistent or 100% anywhere (see also Vasko forthcoming: 3). But as with the study of non-standard patterns or, in fact, any kind of grammatical phenomena of language, it is the tendency of the speakers to use a certain structure over other available structures that is of interest and importance. Thus I am content to use regularisation, as it also has the connotation of smoothing out irregularities in the paradigm.

Vasko (forthcoming) provides a thorough account of was-were variation in Cambridgeshire. Her research plays a significant role in the analysis part of the

\textsuperscript{22} Note also that this kind of phenomenon is defined as dialect levelling in some literature. Though it involves the eradication of irregularities, dialect levelling also implies a social significance in the machinations of the process. “Dialect levelling involves a reduction of marked, socially heavily stigmatised, highly localised, or minority forms in favour of unmarked, less stereotyped, supralocal, majority variants in a dialect mix” (Britain 2002: 35; emphasis in the original). In this study, where I am preoccupied with describing the linguistic system, Britain’s simplification, which he uses to denote the emergence of regularities in formerly irregular systems (ibid: 37), is closer to my use of regularisation.
present study, since the dialect recordings she inspects contain speech from the same
localities and even from some of the speakers that are referred to in this study. She
identifies both was- and were-regularisation manifested in Cambridgeshire (see Map
1). In her analysis, Vasko notes that the distribution of was- and were-regularisation
in the speakers’ grammars is geographical in nature (forthcoming: 1).

Table 1. Was and were regularisation

<table>
<thead>
<tr>
<th>Regularisation to was</th>
<th>Regularisation to were</th>
</tr>
</thead>
<tbody>
<tr>
<td>I was</td>
<td>We were</td>
</tr>
<tr>
<td>You was</td>
<td>You were</td>
</tr>
<tr>
<td>He/she/it was</td>
<td>They were</td>
</tr>
</tbody>
</table>

As Map 1 shows, were-regularisation is typical in the north-west (minus sign), was-
regularisation in the south-east and the east (X-sign) and a mixture of the two systems
(plus sign) is dominant elsewhere.

Vasko’s results give fuel for thought on a number of issues that have to do with
the present study. First of all, they exemplify the typical situation within past BE
variation across varieties of English: The shape of the verb paradigm varies greatly
across speakers and geographical regions, and the “grammatical conditioning”
(Britain 2002: 17) behind the variation can differ even within a single, generally
accepted variety of the language (such as the dialect spoken in Cambridgeshire).
Another issue that Vasko’s data tackle is the existence, description and analysis of
“extensive levelling to were” (Britain 2002: 19), which is corroborated by data from
Cambridgeshire and thus brings forth evidence to this phenomenon, which Britain, too, did identify but on notably low levels.\(^{23}\)

In this study, Vasko’s data are relevant, as I have chosen my informants from both was- and were-areas, and also from the grey area in between, where the speakers display a tendency to shift between the two systems. Note that this study is not intended to refute or corroborate Vasko’s findings, but rather I will show how a different perspective can shed new light on variation within a dialect and partly explain the existence of, for example, Vasko’s three regularisation patterns (shown in Map 1) in the Cambridgeshire data. This understanding is shared by Richards, as her mission is to find “what other options there could be” (2010: 70) beyond the traditional binary distinction of was-were. Richards also notes the benefits of

\(^{23}\) Note that the Cambridgeshire data Britain (2002) refers to and the region he analysed were not the same as those in Vasko (forthcoming).

\(^{24}\) Adapted from Vasko (forthcoming: 3). Here, the minus sign represents data where were-generalisation is found, the X-sign indicates a tendency to regularise to was and the plus sign shows areas where the speakers shift between the two systems. The square denotes the city of Cambridge, and the triangle depicts Elsworth, a locality visited in the Survey of English Dialects (see Orton 1962).
approaching ppBE variation with fresh ideas, since she, as do I, vouches for an analysis of the intermediate variant(s) in the paradigm, where they are considered independently from the “standard representations of was and were” (ibid.).

Richards justifies the existence of the intermediate variant through a historical treatment. The “historical precursors” (2010: 66) she identifies in dialect literature are not very conclusive, but they reveal that past BE forms where the final fricative sound is dropped have been recorded and documented in past dialect studies. It is hardly surprising, then, that the historical antecedent to the intermediate variant is ambiguous:

This poses an interesting question: is [wa] derived from unstressed [we], in which case it is most likely to represent were, or could it be derived from [wə]...in which case it would seem logical to assume [wə] as representative of was? The existing literature does not appear to have any definitive answer to this ambiguity, nor is it necessarily the case that these alternatives are mutually exclusive. (Richards 2010: 67; emphasis in the original)

In this study, the problem is resolved by looking at the variant as a separate and distinct entity within the ppBE paradigm. Doing so discards (but not without cause) the historical question of where the variant came from in favour of a synchronic look at the distribution and quality of the variant in each individual speaker’s grammar. Richards’ resolution is similar, as her interpretation that “the intermediate variants occupy their own space in the past tense be paradigm” is “a novel way of resolving the past tense be problem” (2010: 78; emphasis in the original), and the results in this paper will present new evidence that favours this interpretation.
To summarise, *was*-were variation within the ppBE is one of the most thoroughly studied phenomena in regional varieties of English. Traditionally, research has focused on the StE variants, though with a notable preference to accept was-regularisation as the dominant pattern across varieties of English (see e.g. Smith and Tagliamonte 1998; Britain 2002; Trudgill 2008). However, Vasko’s data from Cambridgeshire (forthcoming) show that were-regularisation is attested, and Britain (2002) confirms this, though by noting that it occurs at notably low levels. Richards (2010) argues that in her data the intermediate variants are distinct from the StE variants, and that analysing the data without recognising the possibility of the new variants as independent entities within the paradigm would lead to a “subjective decision…on the part of the researcher as to where the intermediate variants should be allocated” (15).

In my analysis, I will take Vasko’s findings together with Richards’ approach and formulate a new understanding of the ppBE paradigm. This would substantiate Britain’s testimonial, where he noted that

> A new configuration of verb forms, mostly insensitive to person/number, is diffusing across the Fens, levelling away a wide range both of variable paradigms and phonetically variant forms…and has reallocated two of the dominant forms in the dialect mix to conform to a number of analogical pressures, thereby simplifying the past BE paradigm to two forms: *was* and *weren’t*. (2002: 38)

My thesis differs from Britain’s statement above in that I argue that regularisation in the ppBE is towards the intermediate variant. Furthermore, I will not be studying polarity, even though Richards notes that along with grammatical subject it has been
“considered in virtually every study of variable past tense be” (2010: 67). My decision to leave out the negative past BE variants is based on the simple fact that I have chosen to delimit my analysis to ppBE. Though I recognise the possibility that variability among the negative variants could have implications on the distribution and quality of the positive variants (see e.g. Smith and Tagliamonte 1998; Vasko forthcoming: 5-8), I claim that my research on the positive intermediate variant stands alone and is conclusive enough to establish the variant’s existence.

2.3 Philosophy of variation

This section has a singular purpose: to show that the philosophical and theoretical discussions surrounding variation and the methodological approaches covered in this paper comprise not single, exclusive and isolated revolutions but are in fact continua which begin with the Pre-Socratics and last to the present day. The distinction between what appears and what is, first spelled out by Heraclitus but largely developed by Parmenides (Popper 2002: 214), is present in all ontological meditations of the world as well as in linguistic theory (see e.g. Seuren 1998). In the latter it becomes especially clear when we meditate on how much of the variation in the ppBE can be classified as nothing more than a surface phenomenon (what appears) and how much of it is functionally grounded (what is). This section will outline the major philosophical trends that cover the general ontological meditations and their implications on linguistic theory.

For studies of language, it all boils down to how much we wish to emphasise the unique quality of each utterance and how much we must abstract and idealise the object of our research. If we accept a purely descriptive view, driven on by empiricist
freedom, we have simply *too much data*. On the other hand, if we abstract the object
of our inquiry by, for example, studying the idiolect or by focusing solely on
(hypothetically) universal deep structures, the limitations on what hypotheses we can
make become overbearing.

As stated above, the continuum of theory and methodology that underlie the
decisions I have made in the course of this study extends as far as the Classical period
and receives a major uphaul during the Enlightenment (c. 17\textsuperscript{th}-18\textsuperscript{th} centuries). What
Plato expressed with his dualism\textsuperscript{25} and especially with his ideas about our innate
knowledge of certain truths of the world (Solomon 1997: 136) is reflected in
Descartes’ rationalism during the Enlightenment. In contrast, Aristotle’s attack on
Plato’s dualism and his concept of a single, observable reality is revisited in John
Locke’s refusal to accept the rationalist doctrine. John Locke “rejected the
unsupportable ‘intuitions’ that provided Descartes with his rules and his premises,
and he turned instead to the data of experience as the ultimate source of all
knowledge” (Solomon 1997: 149). At the latest, it is during the Enlightenment and in
the debate between the rationalists and empiricists that the ontological question of
permanence (the innate truths mentioned above, for example) shapes into a deep
methodological dichotomy whose reverberations are still strongly felt in science.

During the 20\textsuperscript{th} century, we see the latest chapter of the empiricist-rationalist-
debate being written. Noam Chomsky, in line with Cartesian rationalism, denounced
the observational method (i.e. the ‘practical discovery procedure’ mentioned below)
as unfit for accessing knowledge of language:

\textsuperscript{25} Plato’s metaphysics bridged the gap between Heraclitus’ concept of an ever-changing world and
Parmenides’ world-view which “insisted that the real world, the eternal and unchanging world, was not
the same as the world of our experience” (Solomon 1997: 68). Plato brought these two together with
his two-world distinction: one for what we experience and observe (the material world) and the other
for what actually exists (the idea world; see e.g. Seuren 1998: 6)
Once we have disclaimed any intention of finding a practical discovery procedure for grammars, certain problems that have been the subject of intense methodological controversy simply do not arise. (1957: 56)

Chomsky’s rejection of empirical data and his support for the existence of innate thought structures, both inherited from Cartesian thinking, were central in establishing the main strands of theoretical linguistics of the 20th century. For Chomsky, the focus of linguistic research had to be linguistic competence (knowledge of language and its structures) rather than performance (use of language and its structures). His mentalistic model of linguistic theory was thus “concerned with finding a mental reality underlying actual behavior” (1965: 4).

Weinreich, Labov and Herzog (1968) gave the empiricists’ answer to Chomsky’s model of linguistic theory. According to them, features of language performance are not “all errorlike vagaries of performance” (125) but feature structure and orderliness that should just as well be in the focus of research as the underlying forms. Where Chomsky’s views are fuelled by the same idealism and intuition that was present in Plato’s dialogues and that formed the basis for Cartesian rationalism, modern day empiricists require that actual data be legitimised as the backbone of linguistic methodology. Weinreich, Labov and Herzog’s pluralistic view of language and language users is thus strongly contrasted with the isolation of the individual mind that e.g. Chomsky so vehemently supports.

For the purposes of this study, the empiricist method works better. The question that thus arises is whether a feature of regional speech, such as the intermediate past BE variant studied here, is to be included as an ordered structure in the grammar of
the speaker or as a deviation of StE grammar, which contains the underlying structures of English use.

The first approach would echo the problem of Heraclitean flux\(^{26}\), since if the speaker’s grammar can freely adopt forms and structures that deviate from what the underlying grammar or set of norms dictates, it would mean that grammars are ever-changing. Weinreich, Labov and Herzog voiced this concern in the following way\(^{27}\): “After all, if a language has to be structured in order to function efficiently, how do people continue to talk while the language changes, that is, while it passes through periods of lessened systematicity?” (1968: 100). I am more than content to adopt the ideology that the grammar of StE and that of the speakers are not mutually exclusive, but interact in many ways depending on, for example, both linguistic and social context, relative status of the speakers and so forth. In other words, variation in language is structured and it is the responsibility of the linguist to find the rules that govern this variation.

In summary, this chapter had the ambitious goal of introducing a continuum of thought that underlies the theoretical and methodological principles present in this paper. What is striking in the development of Western philosophy of science is how the most widely known philosophies seem to rely on extremes in order to get their message through. For example, Descartes’ rationalism was extreme in that it doubted everything except the existence of a thinking subject. Locke’s rebuttal was extreme in that it saw the entirety of our earthly knowledge (save for e.g. mathematical

\(^{26}\) The Pre-Socratic philosopher Heraclitus maintained that the world was in constant flux. In his opinion, for example, it was impossible to step into the same river twice, as the ever-flowing waters of the river have changed it so that it can no longer be regarded as the same the second time (Seuren 1998: 5).

\(^{27}\) Note that though Weinreich, Labov & Herzog (1968) discuss mainly language change, I believe, and this study is based on this belief, that variation in real time poses the same problems for linguistic theory.
constants) to accumulate after birth, and for example David Hume’s development of Enlightenment empiricism was extreme since it forced us to question causality and the realism of the inductive method (see e.g. Solomon 1997: 212-213). Linguistics of the mid-20\textsuperscript{th} century is also based on an extreme: the debate between the empiricism of Weinreich, Labov and Herzog and (Chomskyan) generativism. In this paper, I have avoided taking sides, because I do not consider epistemological truths to reside in any extreme. The fact remains that reality is a concept of which we can ultimately say nothing conclusive at all. What we can do is hypothesise.

Bearing this caveat in mind, I consider the methodology of this paper to be situated more in the empiricist camp, mainly because I use actual observational data to confirm my hypotheses. In a Heraclitean twist, I consider language to be in constant flux. In order to understand this flux, we must employ tools that necessarily abstract and idealise the object of our study. However, we do not need to accept homogeneity as the only legitimate level of abstraction, but in true empiricist spirit we can now unravel the seemingly paradoxical notion of structured (or orderly) heterogeneity. The variation I look at in this study and its implications on how spoken language can or cannot be confined according to a priori categories is an example of a seemingly confusing, heterogeneous mess which simply requires the researcher to spell out a level of abstraction that both the data and the conclusions conform to. It has been the purpose of Chapters 1 and 2 to outline this level in order for the rest of this study, i.e. the actual case study itself, to have a theoretical backdrop to lean on.
3 HELSINKI ARCHIVE OF REGIONAL ENGLISH SPEECH

3.1 Overview

The Helsinki Archive of Regional English Speech (HARES) is an archive of audio-recorded interviews collected in rural England during the 1970s and 1980s\textsuperscript{28}. HARES comprises audio data from the following seven regions of England (see Map 2): Cambridgeshire (1), Devon (2), Essex (3), Isle of Ely (4), Lancaster (5), Somerset (6) and Suffolk (7)\textsuperscript{29}. Total length of digitised audio is around 211 hours (see Table 2). This figure is prone to change, as the audio is still being post-processed and a number of lost tapes have been uncovered since the first digitisation effort.

The interview sessions were originally designed to elicit data for the study of dialect syntax (see Chapters 3.1.1 and 3.1.2). However, with technological advancements in audio digitisation and web programming, the applications of the corpus have multiplied. This is one of the reasons why work on HARES began in the first place.

In addition to the valuable linguistic content, the archive provides a unique insight into local history, cultural customs, folk stories and narratives, all enthusiastically recounted by the informants. (Ahava and Vasko 2009)

Though the corpus has been designed primarily as a linguistic tool, other uses become obvious as one understands how the archive is actually a collection of stories. The

\textsuperscript{28} HARES is a new project (est. 2008), and it is still a work-in-progress. Because of this, there is a shortage of publications on the topic. Since the present writer has worked on the project since its conception, the credibility of any information provided about HARES should thus be established. Work on HARES has been funded by the Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki and the City Centre Campus Online Services, University of Helsinki.

\textsuperscript{29} Cambridgeshire proper (southern Cambridgeshire) and Isle of Ely (northern Cambridgeshire) are treated as separate regions in HARES due to the fact that they were visited by different fieldworkers and because Isle of Ely was historically a county in its own right.
informants were given the freedom to steer the discussion to whichever direction they chose. Interviewer intervention was kept at a minimum (though this was not always successful), and elicitation of linguistic material was done as surreptitiously as possible.

In this chapter, I will introduce HARES by first providing a short summary of its history (see Chapter 3.1.1). Though HARES is a new product and should be considered a corpus in its own right, the data were not collected for HARES per se, but rather for a dialect syntax project in the 1970s, i.e. in the early days of corpus studies and over three decades before HARES work even began. Thus the decisions made during the collection of the material (see Chapter 3.1.2) have a direct consequence on what HARES looks like and what it can be used for.
Table 2. Audio data in HARES

<table>
<thead>
<tr>
<th>Region (code in Map 2)</th>
<th>Audio (hh:mm:ss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambridgeshire (1)</td>
<td>81:07:53</td>
</tr>
<tr>
<td>Devon (2)</td>
<td>11:14:23</td>
</tr>
<tr>
<td>Essex (3)</td>
<td>05:40:05</td>
</tr>
<tr>
<td>Isle of Ely (4)</td>
<td>37:30:06</td>
</tr>
<tr>
<td>Lancashire (5)</td>
<td>06:23:52</td>
</tr>
<tr>
<td>Somerset (6)</td>
<td>38:26:54</td>
</tr>
<tr>
<td>Suffolk (7)</td>
<td>31:13:54</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>211:37:07</strong></td>
</tr>
</tbody>
</table>

However, the HARES team\textsuperscript{30} has developed a new orthographic transcription protocol (OTP; see Chapter 3.1.3) for the new corpus, which differs greatly from the one used in the previous projects. The OTP used in HARES was partly inspired by that used in the Newcastle Electronic Corpus of Tyneside English (NECTE; see e.g. Beal et al. 2007). The major breakaway especially from Helsinki Corpus of British English Dialects (HD) came with the implementation of an XML schema for the annotation of the data (see Chapter 3.1.4). Though this chapter will touch upon the many problems inherent to spoken language corpus compilation, the reader is also directed to revisit Chapters 1.3, 1.4 and 2.1.3 for discussion on the difficulties of working with audio data.

### 3.1.1 History

HARES is built around the same dataset as its predecessor, HD. HARES as well as HD are both development stages of a project whose roots are in the early 1970s in the

\textsuperscript{30} In the beginning of 2010, the HARES team consists of Dr. Anna-Liisa Vasko (coordinator), Simo Ahava (assistant) and Joseph McVeigh (assistant)
Department of English at the University of Helsinki. Founded through the initiative of
the professor and chair of the department, Tauno F. Mustanoja, the dialect project
began as a collaborative effort with Professor Harold Orton of the University of
Leeds. Mustanoja gathered a group of Finnish postgraduate students who all shared
research interest in dialect syntax. The Helsinki Dialect Syntax Group, as it was
called, set out to do interviews in England in order to uncover patterns of

The group members and their fieldwork regions were Ossi Ihalainen, Somerset;
Riitta Kerman, Essex and Lancashire; Leena Pasanen, Suffolk; Ossi Stigell, Devon;
Irmeli Tammivaara-Balaam, Isle of Ely; and Anna-Liisa Vasko, Cambridgeshire. Their fieldwork data and research were intended to later supplement the Survey of
English Dialects (SED; see e.g. Orton 1962), a major dialect research project of the
20th century, but one that focused almost solely on phonological and lexical
phenomena.

SED influence on the Helsinki project was evident especially in the preparatory
stages of fieldwork, as “the principles for the sampling process were to some extent
based on the SED criteria for the selection of localities and informants” (Vasko 2005:
38). However, whereas SED relied on questionnaires to elicit the data, the informal
interview method was favoured in the Helsinki project. This marked a clear departure
from SED, as the informal interview method was used purposefully in lieu of any

31 Riitta Niinivaara did some fieldwork in Suffolk in 1976, and her data is included in HARES even
though she wasn’t a member of the Helsinki Dialect Syntax Group. Riitta Kerman’s data from Essex
and Lancashire were collected for her MA thesis in 1988. However, her informant and locality
selection criteria, as well as the later time period of the audio collection, differed from the other
fieldworkers’ methods so much as to warrant the exclusion of Essex and Lancashire data from the
otherwise rural corpus. Kerman’s data are not currently being processed into HARES, mainly because
they suffer from poor compatibility with other HARES data due to shortage of informants (only twelve
altogether), small number of locations visited (only six altogether) and because Kerman’s fieldwork
took place in urbanised areas.
“direct elicitation methods” (Ihalainen 1981: 27) that were seen as potentially harmful to the analysis of natural language.

The fieldwork was completed, for the most part, by the end of the 1970s. In the beginning of the following decade, Professor Ossi Ihalainen, Mustanoja’s successor as the project coordinator, set out to incorporate the vast amount of transcribed interview data from all the fieldworkers into a computer corpus. The idea behind creating a computerised corpus of the dialect material was supported by another major corpus project at the English Department in the early 1980s: the Helsinki Corpus of English Texts (HC). HC was to comprise two parts: a diachronic and a dialectal. The former would be a corpus of historical English texts, covering a time period of circa 750–1700. The latter would include the material, recorded and transcribed, collected by the Helsinki Dialect Syntax Group. The diachronic corpus, now called just the Helsinki Corpus, was completed and released in 1991\(^{32}\), but the compilation of the dialect corpus was postponed due to Ihalainen’s untimely death in 1993.

However, in 1997 Dr. Kirsti Peitsara took over the dialect project. Under the coordination of Peitsara, the hand-written transcriptions were transferred to text files and digitisation of the first reel-to-reel tapes began. With the dialect corpus taken under the wing of VARIENG and with Dr. Anna-Liisa Vasko, the Cambridgeshire fieldworker, rejoining the project to work on her doctoral dissertation (see Vasko 2005), HD was completed in 2006.

In 2007, digitisation of the rest of the original reel-to-reel tapes was funded by the City Centre Campus Online Services, University of Helsinki, so that the valuable

---

Work on HARES began after the digitisation was complete. It was considered more prudent to create an entirely new corpus, one which would have the audio as the primary data, to accommodate the digitised material. The other option would have been to make extensive modifications to HD so that it could support the use of audio for research, but this would have required a complete overhaul of the corpus, making the creation of a new archive the preferred choice.

Since early 2008, HARES has been a work-in-progress. In May 2009, Anna-Liisa Vasko and Simo Ahava presented HARES for the first time as a poster presentation in the 30th International Computer Archive of Modern and Medieval English (ICAME) conference at Lancaster, UK. This will be followed up in May 2010, when the first product of HARES, the Cambridgeshire sampler (see Chapter 3.2), will be published in time for ICAME 31 at Giessen, Germany, where it will be presented by Simo Ahava and Joseph McVeigh.

Throughout the history of both corpora (HD and HARES), the uniqueness of the audio data has remained the main motivation for work on the projects. The informants give enthusiastic reports about life from the mid-19th century to the time of the recordings (1970s and 1980s). Their expertise in their chosen professions (e.g. threshing, fishing, turf-cutting, mowing, horsekeeping, ditching, cow-milking, ploughing and birding to name but a few) and their love for a simpler way of life are reflected both in the complexity of their verbal expression as well as in how they steer the discussion through unpredictable twists and turns. It is the purpose of the following chapters to show how the HARES team has tapped into these valuable audio resources in order to create an archive that retains the uniqueness of the spoken
testimonials while at the same time harnessing them for a wide variety of different applications.

3.1.2 Data collection

In this chapter, I will explain the basic methods of data collection that the fieldworkers relied on during their work among the informants. As mentioned above, the data were originally meant to supplement the SED, so the selection criteria for informants and localities mirrored those of SED (see Chapters 3.2.2 and 3.2.3 for a rundown of these criteria). However, the data were collected with the informal interview method and not with questionnaires, due to the fact that the data were intended to be used for the study of dialect syntax.

A syntactic study requires longer stretches of speech than, for instance, a lexical study. Therefore, instead of the direct questionnaire method used in the mainly lexical and phonological survey of the SED, the corpus...was collected by the tape-recorded-interview method. (Vasko 2005: 39; emphasis in the original)

It is important to understand the data collection method behind this (or any) corpus, because how the material was gathered has direct implications on how it can be used. For instance, if the HARES data had been collected with a more direct method (such as with an elicitation strategy of some kind), the informants might have tried to avoid using the structures that were clearly dialectal, because they understood how these structures might be stigmatised (see e.g. Ihalainen 1981: 27). The opposite also applies. Some informants might have overused some dialectal feature, because they felt it necessary to parade their expertise and knowledge of the dialect.
The method of choice for HARES data collection, i.e. the informal interview method, required that the interview situations involved an interviewer (usually the fieldworker) and one or more informants. Informality here means that there was no specifically set agenda for the interviews.

The interviews are free in form. The informants, members all of small agricultural communities, were asked to tell about their everyday activities, such as haymaking, harvesting, tending horses, selling and buying cattle, slaughtering, fishing, catching rabbits, digging peat, social activities, etc. (Ihalainen 1976: 608)

During the talk sessions, it was paramount “to make the informant comfortable and to create an atmosphere conducive to spontaneous and free conversation” (Vasko 2005: 49). It is this lack of constraints on the interview sessions that marks the clearest departure from using a questionnaire or directing the informant to provide specific evidence about his or her language use:

If we are to obtain any kind of insight into the structure of everyday spoken language, we need to look at speech where the speaker has selected his own topic which does not emerge as a result of direct questioning. (L. Milroy 1987b: 59)

Another reason for creating an informal setting for the interview was the minimisation of the Observer’s Paradox. The very presence of a microphone, an outsider or even a local with an unfamiliar agenda (such as collecting data for language research) affects the behaviour of the informant (see e.g. L. Milroy 1987b: 59). The relative status of the participants is also something that was carefully
considered when planning the talk sessions (Vasko 2005: 50). For the talk sessions to best reflect a relaxed, ‘everyday’ conversation setting, so-called secondary informants were often present and permitted to participate in the conversation. However, they were usually carefully screened so that their social status would be similar to that of the primary informant.

Finally, Vasko notes that “[b]efore talk sessions involving a tape-recording, I usually went to see the chosen speakers to give them the chance to get to know me” (2005: 48). It was important to establish a relationship with the interviewee, because the eradication of pre-assigned social roles (such as those of the interviewer and interviewee) was considered paramount in establishing a relaxed atmosphere and a relationship of trust between the participants (see e.g. L. Milroy 1987b: 62).

All these steps taken by the interviewers improved the authenticity of ‘everyday language’ recorded by the fieldworkers. Because the fieldworkers’ agenda was to collect speech data that approximated how these dialect speakers would speak in an everyday, relaxed situation with their peers, all these measures to minimise the effect of the Observer’s Paradox are justified. However, as stated in the beginning of this chapter, the data collection method has direct implications on how the data can ultimately be used. So even though the data were intended to reflect the natural course of everyday conversation, was this objective really achieved? First of all, the effect of the Observer’s Paradox is very difficult to counter. Labov notes:

33 People who weren’t specifically chosen for the interview, but were familiar to the primary informant and thus helped make the conversation more “natural” (Vasko 2005: 50). A typical secondary informant was a friend or spouse of the interviewee.
One way of overcoming the paradox is to break through the constraints of the interview situation by various devices which divert attention away from speech, and allow the vernacular to emerge. (1972b: 209)

However, this statement carries a number of problems in its wake. First, as pointed out in Chapter 2.1.2, the concept of regional variety, dialect or Labov’s ‘vernacular’ is notoriously difficult to define. It is simply impossible, without resorting to radical abstraction, to decide when the variety spoken by the interviewee is closest to some ‘pure’ form of his or her dialect. Labov’s notion of an emerging vernacular is thus misguided, because it is the full scale of the speaker’s inventory, StE included, that make up his idiolect. Another problem emerges when the interviewer tries to mask the true intention of the talk session (as in Vasko 2005: 52). Making the interviewee unaware of the true goal of the talk session is only a half-success. The interviewer is also a participant in the conversation. This is something that is often neglected in the literature, because it is the informant’s speech that is being studied. However, the interviewer is always aware of what he or she wishes to elicitate during the talk session, and this is reflected (sometimes more, sometimes less) in the way they react to what the informant says.

To summarise, I would encourage the reader to consider the data referred to in this paper as a close approximation of how the informants would speak in an everyday situation. Not always is this the case, though. The very presence of a fieldworker, a local interviewer or a microphone evokes the Observer’s Paradox. At times, this is apparent in the speech of the informants themselves. When the interviewer has taken elaborate steps to mask the true agenda of the talk session, the informant might pay little attention to their speech, but the same can hardly be said
for the interviewer. Every question asked by the interviewer is an act of elicitation.
Whether it’s a follow-up question to something said by the informant that the
interviewer considers valuable for his or her research or whether it’s something as
innocuous as a typical ‘how old are you’, the interviewer is always aware of the value
of the reply for the study of the speaker’s language.

3.1.3 Orthographic transcription
In order to maintain consistency across all the different regional varieties and
idiolects represented in the corpus, a protocol for transcription had to be established.
For this, we looked at the Newcastle Electronic Corpus of Tyneside English (NECTE;
see e.g. Beal et al. 2007), since a) both NECTE and our project were established as
collections of dialect interviews, b) both have been compiled with the study of
morphosyntactic variation in mind and c) both aim to promote the use of the original
audio files as primary data. For HARES, this was a significant departure from HD,
because the latter had depended on a rather narrow transcription schema. In the
following examples, the difference between an original HD transcription and a
modified HARES transcription is shown:

4. HD: You know, he come 'ome that Feast Sunday night. An' I often used to
stand' and look up them stars.
5. HARES: you know he come home that Feast Sunday night <pause/> and I
often used to stand and look up them stars

(ET, WILLINGHAM, CAM)
Even though Vasko writes that “the use of so-called ‘eye-dialect’ is avoided” (2005: 59), apostrophes to denote dropped word-initial or word-final sounds, for example, seem to exist solely because the transcriber believes “that they somehow capture the authenticity of the utterances so transcribed” (Preston 2000: 616) or because they are considered as typical features of the dialect in question\(^\text{34}\). Whether or not the use of e.g. apostrophes is justified on the grounds that they mark the distinctiveness of the dialect is irrelevant when audio files are provided as research material. In fact, I argue that having a number of variant spellings for a single lexical item (such as *home* and ‘ome) complicates the transcription work needlessly and places a disproportionate amount of importance on thorough documentation which would supply the researcher enough information on the spelling conventions. With audio files as primary material, it is sufficient to adopt a single, predominantly StE spelling for the lexical item on the grounds that even if the end-user is studying variant phonetic representations of the item, they would still have to go through all occurrences of the word, regardless of how it is spelled in the corpus (cf. Beal et al. 2007: 2.1).

In Table 3 below, I present the Orthographic Transcription Protocol (OTP) used in HARES, with short comments on why any specific convention was settled upon (see Ahava and McVeigh forthcoming for a more thorough description of the OTP). Choosing a protocol that mimics StE orthography made work on the transcriptions considerably easier. All the decisions described above have to do with diminishing the role of the transcriber as an interpreter and thus as a direct consequence on any future research made on the transcribed material. I solemnly believe that the OTP the

\(^{34}\) See Peitsara (Unpublished) for a look at HD’s transcription principles. See also Ihalainen (1980: 187) for pre-HD transcription examples.
HARES team crafted allows for the most transparent use of transcriptions and provides the most direct route to the audio.

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>StE orthography</td>
<td>Because audio is the primary material, a consistent orthography is ensured.</td>
</tr>
<tr>
<td>- No eye-dialect</td>
<td></td>
</tr>
<tr>
<td>- Variant spelling avoided</td>
<td></td>
</tr>
<tr>
<td>Dialect vocabulary</td>
<td>If the dialect word exists in an established dictionary (e.g. dockey for lunch) it is used as such. However, if no dictionary entry is found for some dialect word, the HARES team establishes a new spelling for it and uses it consistently.</td>
</tr>
<tr>
<td>- Dictionaries for existing spellings</td>
<td></td>
</tr>
<tr>
<td>- New spellings for previously unrecorded items</td>
<td></td>
</tr>
<tr>
<td>Dialect grammar</td>
<td>Grammatical features are transcribed as they are heard on the audio. However, some functional words are given a new spelling due to their exceptional quality and pervasiveness throughout the recordings (for example, warn’t as a third negative past BE variant and een’t as an alternative to isn’t and ain’t).</td>
</tr>
<tr>
<td>- Non-standard grammar is preserved</td>
<td></td>
</tr>
<tr>
<td>- Sometimes with new spellings</td>
<td></td>
</tr>
<tr>
<td>Other features</td>
<td>Frequently occurring vocalisations such as mm and uh are retained. Hesitations and false starts (e.g. he we- we- went home) are transcribed as well.</td>
</tr>
<tr>
<td>- Interjections retained</td>
<td></td>
</tr>
<tr>
<td>- Hesitations and false starts retained</td>
<td></td>
</tr>
<tr>
<td>Other orthographic principles</td>
<td>Punctuation and sentence-initial capitalisation is nonexistent, mainly because they are not features of spoken language. However, for clarity’s sake, we have chosen to retain other capitalisation (e.g. London, I, Sunday).</td>
</tr>
<tr>
<td>- No punctuation or sentence-initial capitalisation</td>
<td></td>
</tr>
<tr>
<td>- Other capitalisation according to StE</td>
<td></td>
</tr>
</tbody>
</table>

3.1.4 XML annotation

HARES is annotated with XML or eXtensible Markup Language. It can be summarised as follows:
XML (http://www.w3.org/XML/) aims to encourage the creation of information resources that are independent both of the specific characteristics of the computer platforms on which they reside (Macintosh versus Windows, for example), and of the software applications used to interpret them. To this end, XML provides a standard for structuring documents and document collections. (Allen et al. 2007: 33)

Platform and software independency is the main reason why XML was adopted as the standard for HARES. Another reason was that NECTE had been annotated according to the XML standard (see e.g. Beal et al. 2007). To be ‘well-formed’, any XML file has to be validated against a Document Type Declaration (DTD), which contains information about the tags used in the XML file and general document structuring. The DTD chosen for HARES (and NECTE) is Text Encoding Initiative (TEI; http://www.tei-c.org/). TEI hosts a comprehensive tag set especially for the annotation of digital corpora. Even though actual corpus query software for spoken language corpora such as HARES are still few and far between, the TEI Consortium is constantly developing their recommendations for document structuring and the community is vast, so the near future might see a number of applications emerging, all suitable for browsing the HARES data. Finally, an important reason behind choosing XML and TEI for data annotation was because of the observation that they are “emerging as world standards for the encoding of digital information” (Kretzschmar et al. 2006: 195). This notion of a ‘world standard’ underlies the multiple applications of XML today and especially in the near future.

XML enables the use of tiered annotation, since there is no limit to the number of levels of tagging that might be included in the annotation schema. Whereas until
now we have been working solely on the most important level, the orthographic transcription level, future tiers such as phonetic transcription and part-of-speech tagging are entirely possible (see Beal et al. 2007 for a rundown of NECTE’s amalgamation of phonetic, grammatical and orthographic levels). Another perk with using tiered annotation is that data description and presentation become more objective. As explained in the previous chapter, reverting to StE as the model for the OTP diminishes the role of the transcriber as an interpreter. Similarly, using XML tags to annotate the data further abates the potentially harmful effect of transcriber judgment, because now particularly difficult cases can be marked as such. For example, the intermediate variant in the ppBE, studied in this paper, is one of the more difficult items to transcribe, because any spelling the transcriber chooses becomes burdened by the difficult choice of determining to what inventory the variant belongs. With XML, we can transcribe the variant orthographically as whatever would be expected according to StE grammar and then tag it separately as an ‘item of interest’. For example:

6. (Not tagged) they were/was/? going in the army
7. (Tagged) they <w type="INT"> were </w> going in the army (EW, SWAFFHAM PRIOR, CAM)

The main problem with using XML is that it is not exactly user-friendly. Even though reading and understanding XML files becomes easier once the user familiarises with XML syntax, the tags present a visual obstruction for the corpus user (see Kretzschmar et al. 2006: 1997). Of course, with most modern corpus software tags can be excluded from concordance searches, but this would naturally nullify all the work put in during annotation, and the tags provide a lot of useful meta-level
information about the text and the informants. After the Cambridgeshire sampler (see the following chapter) is released, we plan on crafting an XSLT style sheet that allows for conversion of the raw XML data to a more user-friendly format, such as HTML (Hyper Text Markup Language). This way it can be viewed with any web browser.

In Table 4 below I present the XML tag set used in HARES (only the body tags) with the tag description and reference to the relative chapter in TEI P5 guidelines (http://www.tei-c.org/Guidelines/P5/)

35. Tags are always between angle brackets and can contain various attributes, as in (8).

8. `<u xml:id="q15" who="ier1cam07"> ... </u>
9. `<seg synch="#s13-1"> I was </seg> ...
   `<seg xml:id="#s13-1"> really </seg>

An XML element consists of a start tag and an end tag and everything between them. The element can also be empty, such as `<u/>` (which is the same as `<u>` `</u>`). In HARES, we have assigned each utterance a unique ID, so that they can be directly referred to if needed. Also, overlapping speech segments such as (9) have unique IDs that link the two overlapping segments together.

To summarise, XML has already established itself as a standard of sorts for annotating corpora. We chose XML for HARES because of this reason and also because of XML’s adaptability to corpus compilers’ different needs. Extensibility of the markup language ensures that we can add new tags and elements that provide

35 See Appendix A for an annotated transcription excerpt from HARES.
Table 4. Body tags used in HARES XML

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Guideline (P5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;u&gt; ... &lt;/u&gt;</td>
<td>An uninterrupted stretch of speech by a single speaker.</td>
<td>8.3</td>
</tr>
<tr>
<td>&lt;anchor/&gt;</td>
<td>Synchronises the text with the audio in 10 second intervals.</td>
<td>16.3</td>
</tr>
<tr>
<td>&lt;seg&gt; ... &lt;/seg&gt;</td>
<td>Used to 1) synchronise overlapping elements and 2) describe any non-Standard spellings (such as those resulting from hesitation).</td>
<td>(1) 16.3 (2) 3.4</td>
</tr>
<tr>
<td>&lt;vocal&gt; &lt;desc&gt; ... &lt;/desc&gt; &lt;/vocal&gt;</td>
<td>Used to describe any non-lexical vocalisations (such as laughter, coughing).</td>
<td>8.3</td>
</tr>
<tr>
<td>&lt;gap/&gt;</td>
<td>Denotes a gap in recording.</td>
<td>3.4</td>
</tr>
<tr>
<td>&lt;unclear/&gt;</td>
<td>An unclear segment in transcription.</td>
<td>3.4</td>
</tr>
<tr>
<td>&lt;pause/&gt;</td>
<td>A pause in speech (length not specified).</td>
<td>8.3</td>
</tr>
</tbody>
</table>

detailed and unique information about HARES only. The creation of style sheets for converting the XML files to other formats is relatively painless, so in the future HARES can be appended with files in e.g. HTML and plain text format.

3.2 The Cambridgeshire sampler

The Cambridgeshire sampler is a collection of transcribed and annotated interviews from Cambridgeshire. The sampler is due for release in May 2010. The sampler will contain 20 recordings (see Appendix B) from around the county. Development of the sampler has been a pioneer effort for HARES compilation, because the OTP and the XML schema have been shaped on the basis of the sampler’s compilation principles.

The Cambridgeshire recordings were chosen as the source of the first sampler for two
main reasons: 1) Cambridgeshire was the most widely covered area in the dialect project and it is the only one with extensive documentation (see Vasko 2005) and 2) the original fieldworker, Anna-Liisa Vasko, is the only fieldworker still working on the data, so she has been able to provide crucial information about the informants and the recordings.

In this chapter, I will first introduce Cambridgeshire as a traditional dialect area. In her doctoral dissertation, Vasko noted that “Cambridgeshire has been more or less a blank spot on the linguistic map” (2005: 36), which was one of the reasons why she decided to ‘fill in’ this spot by doing her fieldwork in the county. The variation in ppBE in Cambridgeshire was already introduced in Chapter 2.2, so my introduction of the area will be more superficial and largely based on the introductory chapters of Vasko (2005). Next, in Chapters 3.2.2 and 3.2.3, respectively, I will give a summary of the informant and locality selection criteria for both the Cambridgeshire recordings in general and the sampler in particular. The general criteria were adopted by the other fieldworkers as well, though Vasko does have significantly more female speakers as primary informants in her recordings.

3.2.1 Introducing Cambridgeshire

Cambridgeshire county is located on the outskirts of East Anglia, a cultural and historical district in eastern England. Though the “inclusion of Cambridgeshire in East Anglia...depends on the way the area is defined” (Vasko 2005: 3), historically the county is separate from East Anglia. Even though the region is often difficult to clearly define, the inclusion of Cambridgeshire into East Anglia is of importance at least to the people themselves, as some of the interviewees display a sense of pride in
being identified as East Anglians (ibid.: 4). The county is most often associated with its urban and academic centre, the town of Cambridge\textsuperscript{36}, but among dialectologists it is today regarded as a significant transitional area because of the surrounding Fens (see Britain 2001: 221). Indeed, Britain (2002: 17) describes the Fens as a particularly interesting region for the study of dialects, because they function as a natural boundary between the Midland and East Anglian dialects.

Cambridgeshire speech has been, more or less, infrequently visited by scholars and academics. Vasko summarises the general attitude towards the Cambridgeshire dialect in the following three assertions:

1) Cambridgeshire speech is not much different from StE. This statement was historically justified by scholars who saw the dialect as a relic of Mercian speech, which had been a major influence on StE. Additionally, later scholars wrote that StE was the language spoken in the Fenland (“including the Cambridgeshire fenland area”). Very recently, some academics have continued in this vein, positing that Cambridgeshire speech is still very close to StE, relying on phonological data as evidence for these claims. (Vasko 2005: 18-22)

2) Cambridgeshire speech is not different from its neighbouring counties (Vasko 2005: 23-25). Cambridgeshire is surrounded by counties that have been generally considered far more ‘dialectal’ or ‘provincial’ than it and that have garnered much more interest from scholars and dialectologists alike. Norfolk, Suffolk and Huntingdonshire, for example, have had a great influence on the speech of Cambridgeshire up to a point where sometimes it is safer to speak of a dialect mixture in the county than a clearly separate variety of speech. However, this observation is refuted by Vasko (ibid.: 268).

---

\textsuperscript{36} Vasko (2005: 10) notes that “[i]t is not surprising that the earliest works paying attention to the speech in Cambridgeshire centred around the University, the institution which has made the name of the County so famous.”
3) Cambridgeshire speech is similar to cockney. Though scholars have claimed that Cambridgeshire speech rings with the cockney sound, it has been lately established that the direction of influence is, in fact, from Cambridgeshire to London and not the other way around. (Vasko 2005: 25-26)

All three assertions can be attributed to a single observation: there is not enough information about Cambridgeshire speech (Vasko 2005: 35-36). The county has been often neglected because of a larger interest in its neighbours (especially Norfolk and Suffolk). The existing evidence is often “too brief” or “insufficiently localized” to be of use in describing intricacies of the variety (ibid.).

Vasko (2005) is thus an important source in determining qualities of the Cambridgeshire dialect. Her dissertation looks into the grammatical peculiarities of Cambridgeshire speech, with a special focus on prepositional locative expressions. In Table 5 below, I have summarised Vasko’s findings about the main characteristics of Cambridgeshire speech. What can be deciphered from the table is that the dialect shares a lot of features that are traditionally considered ‘dialectal’. Particularly variability in all its manifestations provides a sharp image of Cambridgeshire speech especially in terms of how it differs from StE grammar. This is an observation that I can subscribe to on the basis of my work on the corpus itself as well as the research that I present in this study.

All in all, Cambridgeshire provides a fruitful target for research on and surveys of regional speech partly because it has been neglected in previous surveys (see above) but also because the rural, easygoing manner of Cambridgeshire folk (that becomes apparent when one listens to the HARES recordings) provides easy access to information and stories that the people willingly share with the outside world. Whether they talk of the ‘good old times’ or whether they go on and on about the
latest neighbourhood gossip the stories they recount with enthusiasm are of special value linguistically, culturally and historically.

3.2.2 Informant selection

As mentioned before, the dialect project originated as a continuation of the Survey of English Dialects (SED). Thus, the selection criteria for informants and localities in the dialect project mirrored those of SED. Typically, the informants were similar to what Chambers and Trudgill describe as NORMs: “nonmobile, older, rural males” (1980: 29). However, this term was coined after the fieldwork and there were many
exceptions to this rule, for example Riitta Kerman’s Essex and Lancashire data (young informants of both sexes) and the many female informants in Vasko’s Cambridgeshire and Tamiiavaara-Balaam’s Isle of Ely recordings.

The fieldworkers found their informants through the help of local assistants (when available) and via post offices, pubs or village shops (Vasko 2005: 44). Especially the local assistants were helpful because they knew the older people and could point the fieldworkers in the right direction. As word spread of the fieldworkers, older people in the villages even anticipated their visits:

Towards the end of the data-collecting period, in three of the villages there were queues of old villagers willing to talk about the old days. It seemed as if there was a competition between them. A person was not regarded as an expert on village life or on old words or sayings unless he or she were interviewed. (Vasko 2005: 44)

Finding the ‘best’ informant is of course a difficulty. An outstanding problem in dialect studies, one that has been touched upon multiple times in this paper, is trying to define when a spoken variety approaches the ‘purest’ form of everyday, dialectal speech. Vasko observed that some informants were experts on village history, clearly rural in their habits, but they still spoke a variety that resembled StE (2005: 44). Gladly, these informants were included in the recordings, because their stories unearth an enormous amount of evidence on rural life in the old days.

HARES informants were selected on the basis of their status as ‘experts’ on the local dialect. A typical informant was an elderly male, who was a retired farm-labourer and who had lived in the same region throughout his life. For the
Cambridgeshire recordings, justification for informant selection is provided by Vasko:

- **Lifelong residence**: “To guarantee that the informants’ speech included the features typical of their communities, speakers with uninterrupted residence in the village were selected.” (Vasko 2005: 45) Naturally, almost all the male speakers were absent during World War I, and some informants had moved from one village to another in Cambridgeshire. Nevertheless, having lived in the same region all their lives is a matter of pride to some informants as they feel that they are truly experts on local life. Some informants are so set in their ways that they have never left the village for even the smallest errand. Local residency is extended also to the parents of the informants and preferably to spouses, too.

- **Social class and sex**: “The informants were chosen from the lower end of the socio-economic scale, as it is generally agreed by linguists that non-standard regional features occur most consistently in the speech of working-class people.” (Vasko 2005: 45) Since the villages Vasko visited were in the rural parts of the county, finding farmers, horsekeepers, land-workers and other rural workers was not difficult. Female informants were usually housewives. In the Cambridgeshire recordings, the proportion of female informants is higher than in other contemporary projects (such as SED). Even though the traditional (that is, as late as the early 1970s) consensus had been that male informants were more rustic in their speech than women, Vasko (ibid.: 46) notes that the speech of women received increased attention in linguistics as the discipline of sociolinguistics continued to gain a stronger foothold.

- **Age**: “The informants were selected from the oldest age-group, since older speakers are generally considered to be more conservative in their speech, with young people being the innovators.” (Vasko 2005: 46) The informants were, without exception, elderly people. In the corpus, the youngest interviewees are in their 60s, while the oldest are nearly a hundred.

Vasko goes on to note that
However, the value of the Cambridgeshire Corpus may not be in its size, but in the careful selection of the informants – persons who were born and bred in Cambridgeshire and whose family background guaranteed a long inheritance of what one might call ‘genuine’ Cambridgeshire speech – and in its dense coverage of localities. (2005: 261-262)

For the Cambridgeshire sampler, the informants were included if their speech was sampled in the forthcoming Cambridgeshire grammar (Vasko forthcoming). Even though the selection could have been more equally distributed around the county and the different sociolinguistic categories (sex, age, occupation, etc.), we maintain that it gives an accurate look into the differences and similarities that exist within the elusive notion of a single, uniform ‘dialect’. From an individualistic perspective, the sampler contains twenty or so greatly different life stories, each unique in their representation in the speech of the informant.

I chose the four informants in this study because of the emergence of the intermediate variant in their grammars. As this study is about variation in individual grammars, I saw no need to resort to random sampling or a larger sample of the corpus. My method is largely based on qualitative assessment, with which I support my hypothesis. The four informants and the peculiarities in their speech are presented in Chapter 4.

3.2.3 Locality selection

For the fieldworkers, the main criterion for selecting localities within a region was that they had to be what might be considered rural towns. Before entering the region, the fieldworker would design a network of localities that he or she would visit.
Vasko, for example, “visited practically every village in Cambridgeshire” (2005: 42) looking for suitable informants. She was also keen on visiting villages in all corners of the county, so that the informants she interviewed would truly represent the county as a whole.

As for the Cambridgeshire sampler, the informants were chosen first, since their speech would be analysed in the forthcoming Cambridgeshire grammar. What naturally follows is that the localities were selected on the same grounds. Even though some villages are overrepresented (e.g. Willingham), the overall distribution is quite comprehensive (see Map 3 in Appendix B). Since the sampler contains only a small portion of the Cambridgeshire recordings, future efforts will fill in the map and will, at some point, contain all 46 localities visited by Vasko.

In the present study, criteria for selecting localities have been of secondary importance. My data come from the informants, and since my thesis revolves around the speech of individuals and variation within their grammars, geographical variables are not decisive factors in my analysis. However, I tried to avoid choosing multiple informants from a single village with Willingham as the exception to the rule. This was because I agree that variability does manifest differently in different regions, so in order to get a most colourful look at the variation within the speakers’ grammars, choosing different localities was an obvious choice. Furthermore, I deliberately chose the three localities (Willingham, Burwell, Harlton), because the ppBE regularisation patterns of the informants were different enough to warrant a reanalysis of Vasko’s (Forthcoming; see also Chapter 2.2) conclusions.
4 IDENTIFYING THE INTERMEDIATE VARIANT

My thesis statement, introduced in the first chapter, was that the positive past BE verb paradigm of certain Cambridgeshire speakers should be appended with a new intermediate variant [wə]\(^{37}\). This claim is based on observation, as the variant occurs frequently in the speech of a number of speakers, all interviewed during the 1970s in Cambridgeshire. In this chapter, I will present empirical data to support my claim. Table 6 below contains brief profile descriptions of the four speakers I chose for this study (consult also Table 20 in Appendix B).

The speakers I selected represent a fairly colourful picture of the HARES Cambridgeshire sampler. First of all, I have one speaker from the lowest age group (actually, cam02 was the youngest speaker among all the Cambridgeshire informants) and two from the highest. I have also sampled the speech of one of the two women in the Cambridgeshire sampler. Geographically, two of the informants are from Willingham in northern Cambridgeshire, one from Burwell in the northeast and one from Harlton in the central southwest.

My method is presented in Chapter 4.1. In the sections following that, I introduce each informant as a separate case study. In order to assess qualitatively the speech of the informants, dedicating a chapter for each was necessary. The ppBE of each informant is presented first by looking at the overall distribution of the variants in the paradigm. Next, the linguistic context is evaluated, with special focus on non-

---

\(^{37}\) I recognise Lass’ (2009) and Flemming’s (2009) concerns about reducing the neutral vowel to a single schwa-representation. However, even though I admit that there are notable differences in the vowel qualities of the schwas in my data, I do not consider it necessary to splinter the intermediate variant into several, context-sensitive allomorphs on the basis of the quality of the neutral vowel in the nucleus. Since my research question is clearly limited to the investigation of a possible intermediate variant that manifests as [wa], looking into the various representations of this particular variant is redundant and needlessly complicates the analysis.
Table 6. Informants chosen for the present study

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Sex</th>
<th>Locality</th>
<th>Length</th>
<th>Turns</th>
<th>Words</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>cam02</td>
<td>60-70</td>
<td>M</td>
<td>Willingham</td>
<td>40:03</td>
<td>629</td>
<td>6092</td>
<td>“A story teller. Very natural, fairly broad. Interviewed in his barn. False teeth, but didn’t have them when interviewed.”</td>
</tr>
<tr>
<td>cam04</td>
<td>70-80</td>
<td>F</td>
<td>Willingham</td>
<td>36:33</td>
<td>382</td>
<td>3914</td>
<td>“Almost deaf, false teeth. Interviewed at her home with her husband. Dialect broad.”</td>
</tr>
<tr>
<td>cam09</td>
<td>90+</td>
<td>M</td>
<td>Burwell</td>
<td>47:41</td>
<td>390</td>
<td>5201</td>
<td>“Proud of his expertise. Poor hearing and sight. A very good dialect speaker. Interviewed at his home.”</td>
</tr>
<tr>
<td>cam10</td>
<td>90+</td>
<td>M</td>
<td>Harlton</td>
<td>51:50</td>
<td>357</td>
<td>3522</td>
<td>“Good health, not very good hearing. Kind and talkative. Recorded with his friend. A very good dialect speaker.”</td>
</tr>
</tbody>
</table>

38 Turns indicates the number of uninterrupted passages the speaker has.
39 Word count includes all distinct word-like segments from the speaker.
40 Comments are from Vasko (2005: 325-337 and unpublished field notes).
standard occurrence of the variants. Thirdly, I take a look at the stress patterns of each utterance, paying special attention to the occurrence of the intermediate variant. Finally, I take a look at the discourse context, looking for any evidence of e.g. resilience to interviewer input.

4.1 Method

Richards (2010) had identified two intermediate variants in her research on Morley English. She treats [wɻ] and [wə] as two separate entries in the ppBE, supporting her claim with acoustic and morpho-phonological evidence. However, I shall focus on [wə] alone for two reasons: First, [wə] is more interesting in that it occurs across the board in a variety of contexts. Furthermore, its ancestor is difficult to define, as noted also by Richards (ibid.: 71-72). Secondly, I agree with Richards that “[wɻ] is almost certainly derived from [wɻz]” (ibid.), which in the name of looking for a regularisation pattern towards the intermediate variant (as I have set to do) means that [wɻ] is not the kind of neutral variant that I am looking for, but a rather clear allomorph of was in specific contexts (namely in unstressed position).

I identified the variants in the ppBE by ear, using the Transcriber software (http://trans.sf.net/). A more thorough and conclusive method of acoustic analysis was passed by mainly because of the quality of the recordings and the restrictions of the data collection method: much of the audio is of mediocre quality, making a computer analysis obsolete due to background noise or other distractions. The data were

---

41 ‘Non-standard’ here means that the variant present in the data is not what would be expected according to SiE grammar. For example, I were there is non-standard, because was would be expected with singular subjects.
collected via the informal interview method, which means that wordlists or other methods of determining phonological inventories of the speakers were not available. Finally, distinguishing all the past BE variants was not such a difficult task, as long as one listened carefully. In Table 7 below, I present the variants along with a comment on how I identified each one.

**Table 7. Past BE variants identified in the study**

<table>
<thead>
<tr>
<th>Past BE variant</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>[wɔ]</td>
<td>The intermediate variant. Central schwa vowel sound and no fricative release</td>
</tr>
<tr>
<td>[wɔɾ]</td>
<td>The intermediate variant occurring from time to time before a vowel (indicating rhoticity)</td>
</tr>
<tr>
<td>[wɔ]</td>
<td>Richards’ (2010) second intermediate variant. In this study, it is treated as an allomorph of was</td>
</tr>
<tr>
<td>[wɔz]</td>
<td>Standard was</td>
</tr>
<tr>
<td>[wɔ:(r)]</td>
<td>A positive variant that is clearly not was or were. Occurs in stressed position only</td>
</tr>
<tr>
<td>[wɛ:(r)]</td>
<td>Standard were</td>
</tr>
<tr>
<td>[wɔznt]</td>
<td>Standard wasn’t</td>
</tr>
<tr>
<td>[wɔ:(r)nt]</td>
<td>A negative variant that is clearly not wasn’t or weren’t</td>
</tr>
<tr>
<td>[wɔnt]</td>
<td>Same as [wɔ:(r)nt] above, but with a short vowel and no /r/-sound</td>
</tr>
<tr>
<td>[wɛ:(r)nt]</td>
<td>Standard weren’t</td>
</tr>
</tbody>
</table>

I chose to disregard all past BE variants that occurred in two specific environments:

1) All positive, unstressed variants that occurred before an /sl/-sound. This is because there is a “likelihood of assimilation” (Richards 2010: 72) in these particular cases. For example: “cos I [wɔ] sort of master” or “cos I [wɔs] sort of master” (ET, WILLINGHAM, CAM).
2) All instances where the variant occurs during overlapping speech. This is simply because it is, in almost all places, practically impossible to distinguish vowel quality or whether or not there is an audible fricative sound if more than one person is speaking.

Apart from the two exceptions mentioned above, I went through every single instance of the past BE verb, both positive and negative, in each interview chosen for this study. As my results will show, the evidence is conclusive, as there is a strong tendency to favour the intermediate variant in unstressed contexts among the speakers I chose for the present study. Finally, even though polarity was not considered a contributing factor to my thesis, I have included the distribution of the negative variants for two reasons: 1) the negative variants complete the picture of variation in the past BE verb paradigm and 2) they provide some interesting results for future research (see below).

4.2 Informant #1 – cam02

Cam02 is a male speaker from Willingham. He is the youngest informant in the Cambridgeshire recordings. He is a fast and relaxed speaker, and he enthusiastically shares information about work on the land and anecdotes from his youth. He was interviewed by the fieldworker’s local assistant.

4.2.1 Overall distribution

All occurrences of the positive and negative past BE variants in cam02’s speech are presented in Table 8 below. Cam02 displays a strong preference towards using the

42 See Table 21 in Appendix C for the full data table.
intermediate variant, as it occupies 65.7% of all occurrences of the positive variants. The hypothetically rhoticised [wɔː] is strongly represented too, amounting to 11.4% of the variants. Standard [wɔdz] is nonexistent, though the shortened [wɔ] is encountered 6 times (5.7%). In Vasko (forthcoming), Willingham is presented as a locality where the informants regularise to were, and the percentage (10.5%) of [wɛ:(r)] together with the lower percentage for was and its shortened allomorph would indicate that this is the case. However, the neutral [wɛ:(r)] occurs 7 times (6.7%), giving fuel for possible future research on this variant.

Table 8. Positive and negative past BE variants in cam02

<table>
<thead>
<tr>
<th>Positive variants</th>
<th>[wɔ]</th>
<th>[wɔr]</th>
<th>[wɔ]</th>
<th>[wɔdz]</th>
<th>[wɛ:(r)]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>69</td>
<td>12</td>
<td>6</td>
<td>0</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>% of total</td>
<td>65.7</td>
<td>11.4</td>
<td>5.7</td>
<td>0.0</td>
<td>6.7</td>
<td>10.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative variants</th>
<th>[wɔdznt]</th>
<th>[wɛ:(r)nt]</th>
<th>[wɛnt]</th>
<th>[wɛ:(r)nt]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>% of total</td>
<td>0.0</td>
<td>74.0</td>
<td>0.0</td>
<td>26.0</td>
<td>100%</td>
</tr>
</tbody>
</table>

Only two of the four negative variants were encountered: [wɛ:(r)nt] with a 74% occurrence rate and [wɛ:(r)nt] with 26%. Even though polarity is not under scrutiny in this study, the distinct [wɛ:(r)nt] is interesting, especially since it is favoured by the speaker as the negative past BE variant. Further research should be devoted to this phenomenon.
4.2.2 Linguistic context

Non-standard occurrence of the positive variants is presented in Table 9. The table is split in three parts, with the first indicating occurrence of tokens where standard *was* would be expected, the second where standard *were* should occur and the third indicating occurrence in segments preceding a vowel.

<table>
<thead>
<tr>
<th></th>
<th>[wɔ]</th>
<th>[wɔɾ]</th>
<th>[wɔɾ]</th>
<th>[wɔɻ]</th>
<th>[wɻ(r)]</th>
<th>[wɛ(r)]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>was</em>-contexts (e.g. he _ my father’s uncle)</td>
<td>50</td>
<td>8</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>8</td>
<td>77</td>
</tr>
<tr>
<td><em>were</em>-contexts (e.g. they _ too big)</td>
<td>19</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>pre-vocalic</td>
<td>5</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>33</td>
</tr>
</tbody>
</table>

This table reveals that the speaker displays a tendency towards non-standard use of ppBE variants. The stressed [wɛ:(r)] and [wɻ:(r)] occur a total of 13 times in *was*-position (e.g. *he were my father’s uncle*) and only 5 times where StE grammar would require *were*. The only instances of the (hypothetic) phonetically reduced [wɻ] all occur in utterances where they would be expected. [wɔ] and [wɔɾ] occur in both contexts, revealing a general tendency to use the intermediate variant regardless of context. Note that all [wɔɾ]-tokens manifest before a segment that begins with a vowel, providing further evidence for the variant being the rhoticised version of [wɔ]. However, in five cases [wɔ], too, occurs before a vowel, strengthening the case of an independent intermediate variant.

---

43 [wɻ:(r)] is treated as a non-standard occurrence regardless of context due to the fact that its function and grammatical distinctiveness are as of yet undetermined.
4.2.3 Stress pattern

Table 10 contains information about how often the variants occur in stressed positions. Naturally, [wə] and [wɔr] are nonexistent here, “given the lack of stressed schwa elsewhere in English” (Richards 2010: 74).

Table 10. Stressed and unstressed occurrences of the ppBE variants in cam02

<table>
<thead>
<tr>
<th></th>
<th>(stressed/unstressed)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>[wə]</td>
<td>0/69</td>
<td>18/87</td>
</tr>
<tr>
<td>[wɔr]</td>
<td>0/12</td>
<td></td>
</tr>
<tr>
<td>[wɔ]</td>
<td>0/6</td>
<td></td>
</tr>
<tr>
<td>[wɔz]</td>
<td>0/0</td>
<td></td>
</tr>
<tr>
<td>[wɔ:(r)]</td>
<td>7/0</td>
<td></td>
</tr>
<tr>
<td>[wɛ:(r)]</td>
<td>11/0</td>
<td></td>
</tr>
</tbody>
</table>

These results are hardly surprising. The shortened variants [wə], [wɔr] and [wɔ] only occur in unstressed positions44, and [wɔ:(r)] together with [wɛ:(r)] are favoured in stressed positions.

4.2.4 Discourse context

In cam02, the informant is talkative and lively, leaving little room for the interviewer to interrupt. The surrounding discourse provided little evidence for how strong the intermediate variant is in the informant’s grammar. The only case where resilience to interviewer input could be found was the following:

8. Q: no <pause/> that was your grandfather
   S: that [wə] me grandpa that

44 Note the difference to Richards (2010) here, as she found that “[wə] is strongly favoured in stressed contexts” (74).
Here the informant refuses the variant offered by the interviewer and uses the preferred intermediate variant instead.

4.3 Informant #2 – *cam04*\(^{45}\)

Cam04 is a female speaker from Willingham. She was interviewed with her husband. She is one of the two female primary informants included in the Cambridgeshire sampler of HARES. She has false teeth, which might affect how she releases the fricative sounds at the end of words. However, the intermediate variant is still very pervasive in her speech, as Table 11 below reveals.

### 4.3.1 Overall distribution

<table>
<thead>
<tr>
<th></th>
<th>Positive variants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[wə]</td>
</tr>
<tr>
<td>n</td>
<td>42</td>
</tr>
<tr>
<td>% of total</td>
<td>48.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Negative variants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[wɹznt]</td>
</tr>
<tr>
<td>n</td>
<td>0</td>
</tr>
<tr>
<td>% of total</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

Cam04 favours the intermediate variant in her speech, with 48.8% for the /r/-less variant and 15.1% for [wər]. Similarly to cam02, the other Willingham informant in this study, she seems to regularise to [wɹ:(r)], with 19.8% of tokens attributed to the variant. [wɹ] and [wɹz] are not entirely underrepresented, as they occur 11.6% and

---

\(^{45}\) See Table 22 in Appendix C for the full data table.
4.7% of the time, respectively. Interestingly, cam04 does not use \([\text{wD:(r)}]\) a single time (remember the 6.7% occurrence rate in cam02 data). Distribution of the negative variants shows a firm tendency to use \([\text{wE:(r)nt}]\) (75.0%) with not a single token of \([\text{wDZnt}]\) present and only 5 occurrences of \([\text{wD:(r)nt}]\) and \([\text{wDnt}]\) combined in the data.

### 4.3.2 Linguistic context

Cam04 favours the intermediate variant regardless of context, with similar results in both environments. Where was is required, she uses over half of the time the intermediate variant (both /\(r/\)-less and rhotic), and where were is expected, she favours the intermediate almost categorically.

<table>
<thead>
<tr>
<th></th>
<th>([\text{w@}])</th>
<th>([\text{w@r}])</th>
<th>([\text{wD}])</th>
<th>([\text{wDZ}])</th>
<th>([\text{wD:(r)}])</th>
<th>([\text{wE:(r)}])</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>was-contexts</td>
<td>20</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>9</td>
<td>51</td>
</tr>
<tr>
<td>were-contexts</td>
<td>22</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>35</td>
</tr>
<tr>
<td>pre-vocalic</td>
<td>7</td>
<td>12</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>23</td>
</tr>
</tbody>
</table>

Her use of was and its shortened allomorph is standard, whereas were occurs in both standard and non-standard contexts in her speech. As for the rhoticised \([\text{w@r}]\), all 12 tokens occur before a vowel sound in the following segment. However, 7 of the 23 pre-vocalic tokens are \([\text{w@}]\). This is a similar result as with cam02, where the /\(r/\)-less
intermediate variant was present in pre-vocalic position as well, providing evidence for its salience as a distinct variant.

4.3.3 Stress pattern

In Table 13 below, all occurrences of the ppBE variants are categorised with respect to whether they occur in stressed or unstressed position.

Table 13. Stressed and unstressed occurrences of the ppBE variants in cam04

<table>
<thead>
<tr>
<th>(stressed/unstressed)</th>
<th>[wɔ]</th>
<th>[wɔr]</th>
<th>[wɔ:]</th>
<th>[wɔz]</th>
<th>[wɔ: (r)]</th>
<th>[wɛ: (r)]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/42</td>
<td>0/13</td>
<td>0/10</td>
<td>3/1</td>
<td>0/0</td>
<td>17/0</td>
<td></td>
<td>20/66</td>
</tr>
</tbody>
</table>

The only irregularity above is the single occurrence of [wɔz] in unstressed position which is otherwise dominated by the intermediate variant. However, the explanation can be found in the preceding discourse (see Chapter 4.3.4 below).

4.3.4 Discourse context

Most of cam04’s interview is shaped by dialogue with her husband and the interviewer. Even though she does not show a single case where the intermediate variant is maintained regardless of what the interviewer says, in the following examples are a number of situations where her use of the past BE variant was different from or exceptionally similar to the one offered.

9. Q: not too bad was it
   S: no it weren’t
10. Q: and that was it  
   S: that was it  

11. S2: one and tuppence [wəːnt] it  
   S: no it weren’t

Examples (9) and (11) demonstrate how the informant’s grammar has taken regularisation to weren’t in negative contexts as a rule. She declines both the positive was and the negative [wəːnt] in the dialogue. Example (10) shows how sometimes resilience to interviewer input is not strong enough. Here she mimics the interviewer by producing was in a unique context (unstressed position). Since this occurs only once in the data, it is presumable that this ‘anomaly’ is the result of interviewer input.

4.4 Informant #3 – cam09⁴⁶

Cam09 is an elderly male informant from Burwell in north-east Cambridgeshire. He is a slow and deliberate speaker, who nonetheless uses many non-standard structures in his speech. In Vasko (forthcoming), Burwell was listed as a locality where both was- and were-regularisation occurs. However, my data show that there is a strong tendency to prefer was in all contexts. The intermediate variant is not as strongly represented here as with the other informants, but a case can be made for its salience nevertheless.

4.4.1 Overall distribution

Table 14 below shows the overall distribution of the past BE variants in cam09.

⁴⁶ See Table 23 in Appendix C for the full data table.
Table 14. Positive and negative past BE variants in cam09

<table>
<thead>
<tr>
<th></th>
<th>Positive variants</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[wɔ]</td>
<td>[wɔːr]</td>
<td>[wɔ]</td>
<td>[wɔːz]</td>
<td>[wɔː:(r)]</td>
<td>[wɛː:(r)]</td>
</tr>
<tr>
<td>n</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>41</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>% of total</td>
<td>32.4</td>
<td>0.0</td>
<td>0.0</td>
<td>57.7</td>
<td>2.8</td>
<td>7.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Negative variants</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[wɔːznt]</td>
<td>[wɔːː:(r)nt]</td>
<td>[wɔnt]</td>
<td>[wɛːnt]</td>
<td>[wɛː:(r)nt]</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>4</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>% of total</td>
<td>28.6</td>
<td>71.4</td>
<td>0.0</td>
<td>0.0</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Cam09 uses [wɔːz] in 57.7% of all utterances where the positive past BE variant occurs. Nearly a third of the cases (32.4%) have the intermediate variant in them. [wɛː:(r)] (7.0%) and [wɔː:(r)] (2.8%) occur only rarely, and the rhoticised [wɔːr] and reduced [wɔ] are nonexistent in the data. As for the negative variants, cam09 shows a tendency to regularise to the neutral [wɔːː:(r)nt] (71.4%) with wasn’t as the only other negative variant present in the data (28.6% of the cases). Contrary to Vasko’s (forthcoming) data in which Burwell appears to be a locality where regularisation is towards both was and were, cam09 clearly prefers was.

4.4.2 Linguistic context

Was occurs in over half of the cases where it would be expected according to StE grammar (see Table 15). However, was is also the most frequent non-standard variant, occurring in were contexts 10 out of 19 times. The intermediate variant is strongly represented in the data, occurring in just under a third of all the cases in both contexts. In addition to these, there are a couple of non-standard occurrences of were (along with 3 tokens in were-context) and two occurrences of the neutral [wɔː:(r)]
variant in *was* context. Contrary to the two Willingham informants (cam02 and cam04), cam09 does not use the rhoticised [wɔr] in unstressed segments that precede a vowel. Instead, he uses [wɷz] and [wɷ].

**Table 15. Linguistic context of ppBE variants in cam09**

<table>
<thead>
<tr>
<th>[wɷ]</th>
<th>[wɷr]</th>
<th>[wɷ]</th>
<th>[wɷz]</th>
<th>[wɷ:(r)]</th>
<th>[wɷ:(r)]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>was-contexts (e.g. he _ my father’s uncle)</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>were-contexts (e.g. they _ too big)</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>pre-vocalic</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

### 4.4.3 Stress pattern

The stress patterns are unsurprising on all accounts except with *was*. The intermediate variant occupies only unstressed positions, and [wɷ:(r)] and [wɷ:(r)] occur only as stressed. *Was*, on the other hand, occurs in both positions, manifesting in stressed position 12 times (out of a total of 19) and as unstressed 29 times (out of a total 52). These results show that cam09 is a typical non-standard speaker, who favours *was* in all contexts.

**Table 16. Stressed and unstressed occurrences of the ppBE variants in cam09**

<table>
<thead>
<tr>
<th>(stressed/unstressed)</th>
<th>[wɷ]</th>
<th>[wɷr]</th>
<th>[wɷ]</th>
<th>[wɷz]</th>
<th>[wɷ:(r)]</th>
<th>[wɷ:(r)]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/23</td>
<td>0/0</td>
<td>0/0</td>
<td>12/29</td>
<td>2/0</td>
<td>5/0</td>
<td>19/52</td>
<td></td>
</tr>
</tbody>
</table>
4.4.4 Discourse context

Previous discourse does not reveal anything conclusive about the intermediate variant, but it again highlights how these informants might decline a StE feature provided by the interviewer and favour the non-standard use of the verb. The following examples show how cam09 displays resilience to interviewer input.

12. Q: you weren’t scared
   S: no <pause/> was you
13. Q: that was a big house
   S: lovely house that [wɔːr]

In (12) the informant does not use the standard *were*, even though it is provided by the interviewer (though in the negative). In (13) the informant uses the [wɔːr] variant in an emphatic position at the end of the utterance, even though the interviewer provides *was* in her question.

4.5 Informant #4 – *cam10*\(^{47}\)

Cam10 is an elderly male informant from Harlton in central south-west Cambridgeshire. He was interviewed together with a friend who is very talkative and dominates the discussion in nearly half of all informant turns. Because of this, data for cam10 are scarce. He has the lowest amount of past BE tokens in his recording, even though the recording session itself was the longest out of the four.

\(^{47}\) See Table 24 in Appendix C for the full data table.
4.5.1 Overall distribution

Table 17. Positive and negative past BE variants in cam10

<table>
<thead>
<tr>
<th>Positive variants</th>
<th>[wə]</th>
<th>[wər]</th>
<th>[wʊ]</th>
<th>[wʊz]</th>
<th>[wʊ:(r)]</th>
<th>[wɛ:(r)]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>17</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>% of total</td>
<td>73.9</td>
<td>0.0</td>
<td>17.4</td>
<td>4.3</td>
<td>4.3</td>
<td>0.0</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Negative variants</th>
<th>[wʊznt]</th>
<th>[wʊ:(r)nt]</th>
<th>[wʊnt]</th>
<th>[wɛ:(r)nt]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>% of total</td>
<td>0.0</td>
<td>100.0</td>
<td>0.0</td>
<td>0.0</td>
<td>100%</td>
</tr>
</tbody>
</table>

As stated above, data were scarce. Only 23 positive variants were found all in all. However, out of these the intermediate variant [wə] dominates, occurring 73.9% out of all the positive tokens. [wʊ] is the next highest, with 4 tokens out of 23 (17.4%), though three of these occur in stressed position (see below). [wʊz] and [wʊ:(r)] have only one token each, leaving little room for interpretation. The negative variants are almost nonexistent, with only two cases of [wʊ:(r)nt] found.

4.5.2 Linguistic context

Table 18. Linguistic context of ppBE variants in cam10

<table>
<thead>
<tr>
<th>[wə]</th>
<th>[wər]</th>
<th>[wʊ]</th>
<th>[wʊz]</th>
<th>[wʊ:(r)]</th>
<th>[wɛ:(r)]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>was-contexts (e.g. he _ my father’s uncle)</td>
<td>16</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>were-contexts (e.g. they _ too big)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>pre-vocalic</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
There was only one context where *were* is expected (see Table 18), and the intermediate variant occupies it. In *was* contexts, the intermediate variant occurs 73% of the time, with [wɔ] and [wɔz] occurring a total of 5 times. [wɔ:(r)] is represented with a single token in the *was*-context. In unstressed segments preceding a vowel, only a single token was found, and that was the intermediate [wə].

### 4.5.3 Stress pattern

**Table 19.** Stressed and unstressed occurrences of the ppBE variants in cam10

<table>
<thead>
<tr>
<th></th>
<th>(stressed/unstressed)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>[wə]</td>
<td>0/17</td>
<td>5/18</td>
</tr>
<tr>
<td>[wər]</td>
<td>0/0</td>
<td></td>
</tr>
<tr>
<td>[wɔ]</td>
<td>3/1</td>
<td></td>
</tr>
<tr>
<td>[wɔz]</td>
<td>1/0</td>
<td></td>
</tr>
<tr>
<td>[wɔ:(r)]</td>
<td>1/0</td>
<td></td>
</tr>
<tr>
<td>[wɛ:(r)]</td>
<td>0/0</td>
<td></td>
</tr>
</tbody>
</table>

Stress pattern reveals only one anomaly. [wɔ] occurs three times (out of a total 4) in stressed position. In fact, this is the only informant, who uses stressed [wɔ], a feature which Richards (2010) found dominant in her data from Morley. All the other occurrences of ppBE are unsurprising, with all [wə] tokens occurring in unstressed position and all [wɔz] and [wɔ:(r)] in stressed position.

### 4.5.4 Discourse context

Cam10 was interviewed together with his friend. Even though his friend talks a lot, the dialogue between the interviewer and cam10 prove more interesting in terms of discourse context analysis. The following two examples show how cam10 declines the StE variant provided by the interviewer in a question and uses his preferred variant, the intermediate [wə].
14. Q: was she born in Haslingfield too
   S: no she [wɔ] born in Hauxton
15. Q: so nothing was wasted
   S: no nothing [wɔ] wasted

Even though the interviewer uses was in both questions, the informant replies with the intermediate variant.

5 DISCUSSION

The empirical evidence to support the notion of an independent intermediate variant in the positive past BE paradigm of my four speakers’ grammars was presented in Chapter 4. Even though the speakers were different in many terms: age, sex, occupation, geographic region and use of the past BE variants, some general conclusions can be drawn from the data.

First, my hypothesis that the intermediate variant is salient enough to be treated as a distinct entry into the grammars of the speakers was well-founded. It was the preferred variant in all cases except for cam09, who seemed to prefer the standard was more. If [wɔr] is to be treated as the rhoticised version of the intermediate variant, then especially cam02 and cam04, the two Willingham informants, show even higher percentages in the use of the intermediate variant (77% and 64% respectively). As my data show that [wɔr] occurs solely before a segment that begins with a vowel sound, the hypothesis that it is the rhoticised version of [wɔ] is supported by the empirical evidence I provide. The distribution is not complementary, as pre-vocalic [wɔ] was encountered in all the informants’ data. This would indicate that [wɔ] is embedded strongly enough in the speakers’ grammars to override any
constraints that would normally have the speaker use a rhoticised variant pre-vocally. It is also possible that this phenomenon would simply mean that [wɔ] is not (yet) as strong a dialectal feature as [wɔ] among the speakers. A third possible explanation is that in the examples where the /r/-less variant occurs in unstressed segments preceding a vowel, the vowel sounds are amalgamated in the flow of speech as in the following example.

16. a) Separate sounds: we [wɔ] a-bout to go
   b) Assimilated sounds: we [wɔ]-bout to go

The intermediate variant occurred in both was and were contexts in the speech of all the speakers analysed for this study. This is strong evidence in favour of its treatment as a distinct variant in the verb paradigm, countering the interpretation that it is simply an allomorph of were. In the analysis of linguistic context, cam09 was the odd one out, as he seemed to prefer was-regularisation in both contexts. Similarly to what Richards found, this would support the case that “[wɔ] is more likely to be derived from were, than from was” (2010: 75), even though for example cam04’s data showed that [wɔ] and were can coexist as well (though the percentages for were in her data were less dramatic than was in cam09).

A look at the stress patterns proved unsurprising. The intermediate variant occurs only in unstressed position. This owes to the fact that the schwa vowel simply does not occur in stressed position in the English language. Full form of was occurred once in cam04’s data, but as I showed, this was probably due to interviewer input. In her Morley data, Richards found [wɔ] to occur in stressed position (2010: 74), but in my results only cam10 had [wɔ] tokens in stressed segments. Among the other
speakers, stressed segments were occupied by either the full forms of *was* and *were*. Cam02’s data, however, did reveal an interesting trend: stressed [wɒ:(r)]. Though dominant only in cam02’s recording, this variant demands a closer look in future research. Perhaps it is the stressed intermediate variant, with [wɒ:(r)nt] and [wɒnt] as its negative.

Finally, discourse context showed, with just one exception, how the speakers might decline an StE variant provided by the interviewer in favour of a dialectal structure of their own. In only one case (cam04), the speaker adapted her speech to that of the interviewer, using *was* uniquely in unstressed position. Even though the examples I found were only few, they provide some insight as to how certain dialectal structures are firmly embedded in the grammars of the speakers.

5.1 Comparison to Vasko (forthcoming)

As stated in Chapter 2.2, Vasko’s research plays a prominent role in the analysis of my data, because she referred to the same localities and even some of the same speakers that I analyse in the present study. Though she analysed interviews that are not in the present study (and vice versa), she observed that the regularisation patterns could be categorised geographically:

A locality-by-locality analysis further shows a clear geographical distribution in *was/were* usage. Levelled *were* (*I were, you were, he/she/it were, we were, they were*) is attested in the southwest, whereas levelled *was* (*I was, you was, he/she/it was, we was, they was*) is notable in three regions: the southeast, east and far north. With speakers from the centre area this levelling is variable, usually not following the standard person-number distinction. (Forthcoming: 3; emphasis in the original)
Cam02 and cam04 were from Willingham, which Vasko listed as a were region (see Map 1). My data agree with this insofar as the lower percentages of was among these two speakers are concerned. However, the intermediate variant dominates in the data of these two speakers. Perhaps the intermediate variant was interpreted as a were-token in Vasko’s data (an understandable conclusion, though refuted by my analysis in the present study).

Cam09 was from Burwell, a locality where Vasko found the informants to switch between was and were regularisation. This is partly corroborated by my data, as was-regularisation is dominant in my data and regularisation to the intermediate variant is robust, too. However, were is present in cam09’s speech, especially in stressed positions, complicating the picture further. Indeed, in cam09’s case, the variation in ppBE seems to be at its most unrestrained, though some patterns of regularity do emerge.

Cam10 was from Harlton, which Vasko reports as a similar locality to Burwell, in that the speakers regularise to both was and were systems. However, my data, though scarce, showed a clear preference to use the intermediate variant across the board. Even discourse context showed that the variant is embedded in the speaker’s grammar.

As I stated in Chapter 2.2, the purpose of my research was not to refute Vasko’s data, as the true nature of the intermediate variant is still to be determined. However, my data clearly show that the intermediate variant is robust among the speakers I analysed. It has not replaced were (though were occurs at lower levels than in Vasko’s study), which would favour the interpretation that the intermediate variant is a separate entry into the speakers’ ppBE.
5.2 Comparison to Richards (2010)

Richards’ research question was similar to mine. She was looking for a pattern of emergence for the intermediate variant and hoping to establish its salience as a separate variant in the past BE paradigm. However, she found two intermediate variants: [wɔ] and [wɛ]. She attributed the existence of the latter to a “derivational ambiguity” (2010: 72) that made it difficult to determine what the full forms of these intermediate variants are.

Richards observed that [wɛ] was favoured in stressed positions (2010: 74). Only cam10 in my data had [wɛ] in stressed position, and even then only with three tokens. However, my data revealed [wɛ:(r)] (especially in cam02) in stressed positions. Perhaps this could be used to complete Richards’ analysis, by introducing [wɛ:(r)] as the stressed intermediate variant [wɛ]. My data are not conclusive enough to warrant such an interpretation, but future research into [wɛ:(r)] might be more revealing.

Richards notes in her discussion (2010: 79) how multiple interpretations of the intermediate variant are possible. First of all, it has not been disproved here that the intermediate variant is a derivative of were (or even was), meaning that it should be treated as an allomorph. However, my data show that the intermediate variant occurs in a variety of contexts and shows much more variance than the StE representations of was and were. Secondly, I agree with Richards (ibid.) that there is room for controversy in our interpretation of the intermediate variant. But when one looks at the data I present in the previous chapter and focuses especially on the salience of the intermediate variant, there should be no doubt that my results should, at the very
least, provide novel ways of approaching the past BE variable. Richards offers the interpretation that the intermediate variant is “an ambiguous form of the verb...some sort of catch-all option” (ibid.). On the basis of my data, this solution is a plausible one.

6 CONCLUSION
At the heart of this paper lies the elusive concept of variation: how speakers can simultaneously share a common knowledge of a unified language system and develop it in idiosyncratic and unique ways. Studies of variation do not come without their fair share of controversy. Since they necessarily incorporate the attention to detail that a descriptivist approach demands together with a level of abstraction that is likewise required by all studies of linguistic phenomena, the results are often as debatable as they are revelatory.

It is thus the researcher’s job to describe the phenomenon as lucidly as possible, without drawing conclusions that extend beyond the scope of the data at hand. Neither must the researcher overanalyse the phenomenon by proposing grammatical or functional fluctuations where neither are to be found. The art of descriptive research balances between theory and methodology: The former provides the framework for the research and fuels the imagination, motivating the researcher to formulate hypotheses that are grounded in intuition and observation. The latter encompasses all aspects of the data itself, with special reference to collection and sampling – what is excluded and what is included.

In this study, I have combined theory and methodology to show how difficult it is to pass judgment on features of spoken language in an effort to describe them. My
case study, the intermediate variant in the positive past BE verb paradigm, proves particularly problematic, since the salience of the binary distinction of *was* and *were* makes introducing new entries into the paradigm a cumbersome task. However, my findings support my hypothesis. Among the speakers I chose to analyse for this study, there is a strong tendency to use the intermediate variant across all contexts and regardless of e.g. interviewer feedback. The speakers have adopted the variant as their preferred choice, and that is why it requires lengthy discussion in papers such as this and Richards (2010).

The problem of analysis, i.e. the difficulty of determining grammatical distinctiveness of spoken language phenomena, is ever-present in this study. However, nowhere is it as obstructive as during corpus compilation work. The intermediate variant is only a minuscule drop in the ocean of spoken language structures that require informed transcriber judgment. Every single passage that the transcriber interprets to text should require lengthy research in order to determine if it deserves the special status of being called a distinct ‘word’ in the grammar and lexicon of the speaker. Naturally, time constraints and common sense dispel any effort to actually engage in such a demanding task, but that does not mean that the problem should not be taken seriously. For example, Richards’ (2010) and my findings could potentially falsify or at least demand correction of a very large body of previous research on the past BE variable. The variant has been easy enough to overlook, considering that it occurs in unstressed positions and can often be confused as a phonetically reduced version of *were*, an observation that I have hopefully brought to question with my results.

My methods were not perfect, however. Further research is called for in regard to the acoustic qualities of the intermediate variant (even though Richards 2010
already did this) and especially the [wɔː(r)] and [wɔː(r)nt] variants. The main problem with acoustic analysis is the quality of the audio data. With modern audio processing software\(^{48}\), most of the glitches can be eradicated and the audio quality can be brought up by several notches with simple equalisation and noise removal filters. After this, software such as Praat (http://www.fon.hum.uva.nl/praat/) can be used for the analysis itself. However, normalising the audio file of each individual interview is a major undertaking and cannot be prioritised at the present stage of HARES compilation. In the future, nevertheless, we will look into this and hopefully find some way of integrating acoustic analysis and corpus browsing in the same toolkit\(^{49}\).

As we incorporate more data into the Cambridgeshire section of HARES, the picture will become even more complete. Studying the speakers of the neighbouring counties\(^{50}\) should provide further evidence to the phenomenon, too. After the Cambridgeshire sampler is complete, work will commence on a sampler of one of the other counties. By continuing in this manner, in a few years’ time we will have samplers from all the seven counties in HARES, giving researchers the tools and the method to investigate phenomena such as the intermediate variant across counties, speech communities and individuals.

In the present study, it was the language of the individual that was of paramount importance. I did not aim to draw any conclusions on a larger scale, nor did I plan to make any overt generalisations about the future of the intermediate variant. As I stated early in this paper, my research is limited in time and space to the recordings

\(^{48}\) Audacity (http://audacity.sourceforge.net/) is the freeware audio editor that has been used with HARES audio files.

\(^{49}\) To see an example of this in practise, visit the ONZE Miner tool website (http://onzeminer.sourceforge.net/).

\(^{50}\) Mainly Suffolk and Essex, as HARES includes data from these two counties.
themselves. HARES has not been appended with data after the 1980s, which means that I have little knowledge of the current situation in Cambridgeshire (as there have been no other dialect projects in the region). It would be interesting and useful to visit the county again, looking for evidence of the intermediate variant or any of its further manifestations.

On a final note, the problems of description and analysis, both introduced in the first chapter of this paper, can never be fully resolved. They are the reason why no single, universal protocol for transcribing spoken language exists, why countless papers are published each year, debating the intricacies of some grammatical phenomenon and why papers such as the present study are always in demand. As Saussure observed, reality of language is never truly accessible (1916: 110). In order for us to be able to discuss it, observe it and categorise it, we reduce language to a level of abstraction, rendering it to “associations which bear the stamp of collective approval” (ibid.: 15). The problems of description and analysis highlight the complexity and disparateness that sometimes exist between these two levels of observation and underlying reality (the paradigmatic question of what appears and what is). What remains for the linguist to do is to formulate a hypothesis that is based on observation but grounded in intuition, follow through with a methodology that highlights the uniqueness of the research but simultaneously acknowledges its place in a more general description of the language and finish with results that substantiate the hypothesis but concede that it is only one interpretation of the phenomenon.
REFERENCES


Beal, Joan, Karen Corrigan, Nicholas Smith and Paul Rayson. 2007. Writing the vernacular: Transcribing and tagging the *Newcastle Electronic Corpus of*


Smith, Jennifer and Sali Tagliamonte. 1998. ‘We were all thegither…I think we was all thegither’: was regularization in Buckie English. In *World Englishes 17*(2). 105-126.


APPENDIX A. Annotated transcription excerpt with comments

All non-XML text is **bolded** in this transcript. Comments can be found in the footnotes of each page.

<TEI id="cam07">

<teiHeader type="text">
<fileDesc>
<titleStmt>
<title>HARES interview cam07</title>
<respStmt>
<resp>Transcribed from original audio by</resp>
<orgName>HARES team</orgName>
</respStmt>
</titleStmt>
<publisher></publisher>
</publicationStmt>
<sourceDesc>
<recordingStmt>
<recording type="audio" dur="PT54M36S">i
<respStmt>
<resp>Interview recorded by</resp>
<persName>Anna-Liisa Vasko</persName>
</respStmt>
</recording>
</recordingStmt>
</sourceDesc>
</fileDesc>
<profileDesc>
<particDesc>
<person id="ier1cam07" role="interviewer" sex="f">
<persName>Anna-Liisa Vasko</persName>
</person>
<person id="ier2cam07" role="interviewer" sex="f">
<persName>M</persName>
</person>
<person id="ieelcam07" role="interviewee" age="70-80" sex="m">
<persName>E</persName>
</person>
</particDesc>
</profileDesc>
</teiHeader>

\(^{i}\) Each text is assigned a unique ID
\(^{ii}\) Duration of the audio file (54 minutes 36 seconds)
\(^{iii}\) The interviews were almost always recorded by the fieldworkers themselves
\(^{iv}\) Each participant is assigned a unique ID, with prefix ier- for interviewers and iee- for interviewees
\(^{v}\) Only the initials of the informant’s name are provided
Swaffham Prior

Farmer

Left school at 11

G W

n/a

n/a

n/a

n/a

6 August 1975

know as I can tell you all what they told me I'd a cos they had that's a long while back we know but there

mm

yes

vi Place and date of the interview
vii This is where the actual interview text begins. Each interview text is assigned a unique ID, with the suffix -ortho referring to the level of representation (orthographic transcription)
viii Each utterance is assigned a unique ID with the prefix s- for informant speech and q- for interviewer speech. The who-attribute refers back to the participant description metadata in the header
ix Each anchor is assigned a unique ID, with the last four digits increasing in intervals of 10 to denote elapsed time in seconds
but you know a long while back because you were born in nineteen hundred

ought three

mm

oh yeah

yes

and you were born here

born in the next village Reach

Reach uh huh

yes

yes that's where I was born

and did you go to school there

yes I went till I was eleven

mm

and then they took me away to work on the farms like during the war cos they were so shorthanded with the people you see cos they was going in the army

x The attribute synch links two overlapping segments together. Synch is followed by the linking symbol # and then the ID of the overlapping segment which can be found in the following utterance. The overlapping segment is assigned a unique id, with the same prefix as the utterance but suffixed by a hyphen and a number which is 1 for the first overlapping segment in the utterance, 2 for the second and so on.
yeah

nineteen fourteen war

yeah

well they took me away from school and I had to go to work on the land then and help with the stock

and that was in Reach

yes

mm

yes that was in Reach

and did you come to this village after Reach then

when yes when I got married

[END OF EXCERPT]
### APPENDIX B. The Cambridgeshire sampler content and map

**Table 20. Cambridgeshire sampler content**

<table>
<thead>
<tr>
<th>ID</th>
<th>Sex</th>
<th>Age</th>
<th>Occupation</th>
<th>Locality</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>cam01</td>
<td>m</td>
<td>n/a</td>
<td>farm work</td>
<td>Willingham</td>
<td>22/07/1977</td>
</tr>
<tr>
<td>cam02</td>
<td>m</td>
<td>60-70</td>
<td>land work</td>
<td>Willingham</td>
<td>21/06/1974</td>
</tr>
<tr>
<td>cam03</td>
<td>m</td>
<td>70-80</td>
<td>land work</td>
<td>Rampton</td>
<td>19/06/1974</td>
</tr>
<tr>
<td>cam04</td>
<td>f</td>
<td>70-80</td>
<td>housewife</td>
<td>Willingham</td>
<td>16/06/1974</td>
</tr>
<tr>
<td>cam05</td>
<td>m</td>
<td>90+</td>
<td>horse keeping</td>
<td>Rampton</td>
<td>22/06/1974</td>
</tr>
<tr>
<td>cam06</td>
<td>m</td>
<td>80-90</td>
<td>farm work, driving, shop keeping</td>
<td>Waterbeach</td>
<td>20/06/1974</td>
</tr>
<tr>
<td>cam07</td>
<td>m</td>
<td>70-80</td>
<td>farm work, milking cows, turf-cutting</td>
<td>Swaffham Prior</td>
<td>06/08/1975</td>
</tr>
<tr>
<td>cam08</td>
<td>m</td>
<td>80-90</td>
<td>farm work</td>
<td>Willingham</td>
<td>23/06/1974</td>
</tr>
<tr>
<td>cam09</td>
<td>m</td>
<td>90+</td>
<td>farm work</td>
<td>Burwell</td>
<td>07/08/1975</td>
</tr>
<tr>
<td>cam10</td>
<td>m</td>
<td>90+</td>
<td>farm work</td>
<td>Harlton</td>
<td>15/07/1975</td>
</tr>
<tr>
<td>cam11</td>
<td>m</td>
<td>70-80</td>
<td>farm work, orchard work</td>
<td>Newton</td>
<td>30/07/1975</td>
</tr>
<tr>
<td>cam12</td>
<td>m</td>
<td>80-90</td>
<td>farm work</td>
<td>Harston</td>
<td>16/07/1975</td>
</tr>
<tr>
<td>cam13</td>
<td>m</td>
<td>80-90</td>
<td>farm work, coal delivery</td>
<td>Bassingbourn</td>
<td>18/07/1975</td>
</tr>
<tr>
<td>cam14</td>
<td>m</td>
<td>80-90</td>
<td>farm work</td>
<td>Castle Camps</td>
<td>05/07/1975</td>
</tr>
<tr>
<td>cam15</td>
<td>m</td>
<td>70-80</td>
<td>farm work</td>
<td>Bartlow</td>
<td>05/07/1974</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10/07/1974</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13/07/1974</td>
</tr>
<tr>
<td>cam16</td>
<td>f</td>
<td>90+</td>
<td>housewife, short service work</td>
<td>Waterbeach</td>
<td>19/06/1974</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27/06/1974</td>
</tr>
<tr>
<td>cam17</td>
<td>m</td>
<td>70-80</td>
<td>land work</td>
<td>Fulbourn</td>
<td>18/07/1975</td>
</tr>
<tr>
<td>cam18</td>
<td>m</td>
<td>80-90</td>
<td>land work, road work</td>
<td>Little Eversden</td>
<td>23/07/1975</td>
</tr>
<tr>
<td>cam19</td>
<td>m</td>
<td>70-80</td>
<td>farm work</td>
<td>West Wickham</td>
<td>17/07/1975</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22/07/1975</td>
</tr>
<tr>
<td>cam20</td>
<td></td>
<td>90+ &amp;</td>
<td>farm work &amp; horse keeping</td>
<td>Rampton &amp; Over</td>
<td>26/09/1974</td>
</tr>
<tr>
<td></td>
<td>m</td>
<td>90+</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

\(^1\) Some recordings contain interviews from separate dates.

\(^2\) This interview had two primary informants.
Map 3. Localities included in the Cambridgeshire sampler

1 Over 6 Swaffham Prior 11 Bassingbourn
2 Willingham 7 Fulbourn 12 Little Eversden
3 Rampton 8 West Wickham 13 Harlton
4 Waterbeach 9 Castle Camps 14 Harston
5 Burwell 10 Bartlow 15 Newton
APPENDIX C. Informant data

Table 21. Cam02 data

<table>
<thead>
<tr>
<th>Grammatical subject</th>
<th>[wə]</th>
<th>[war]</th>
<th>[wə]</th>
<th>[wəz]</th>
<th>[wə:(r)]</th>
<th>[wəznt]</th>
<th>[wə:(r)nt]</th>
<th>[wənt]</th>
<th>[wə:(r)nt]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Was contexts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>he/she/it</td>
<td>18</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>4</td>
<td>-</td>
<td>8</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Sing. NP</td>
<td>25</td>
<td>5</td>
<td>6</td>
<td>-</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Sing. existential</td>
<td>6</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>12</td>
</tr>
<tr>
<td><strong>Were contexts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>we</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>they</td>
<td>13</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>18</td>
</tr>
<tr>
<td>Plural NP</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Plural existential</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>69</td>
<td>12</td>
<td>6</td>
<td>-</td>
<td>7</td>
<td>11</td>
<td>-</td>
<td>17</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>Stress</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>11</td>
<td>-</td>
<td>17</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>No stress</td>
<td>69</td>
<td>12</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87</td>
</tr>
<tr>
<td>Pre-vocalic</td>
<td>5</td>
<td>12</td>
<td>-</td>
<td>1</td>
<td>5</td>
<td>-</td>
<td>10</td>
<td>-</td>
<td>4</td>
<td>37</td>
</tr>
</tbody>
</table>
Table 22. Cam04 data

<table>
<thead>
<tr>
<th></th>
<th>[wɔ]</th>
<th>[wər]</th>
<th>[wɔ]</th>
<th>[wɒz]</th>
<th>[wɛ:(r)]</th>
<th>[wɛznt]</th>
<th>[wɛ:(r)nt]</th>
<th>[wʊnt]</th>
<th>[wɛ:(r)nt]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grammatical subject</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Was contexts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>15</td>
</tr>
<tr>
<td>he/she/it</td>
<td>12</td>
<td>2</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>Sing. NP</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>18</td>
</tr>
<tr>
<td>Sing. existential</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td><strong>Were contexts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>we</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>they</td>
<td>6</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>Plural NP</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>Plural existential</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>42</td>
<td>13</td>
<td>10</td>
<td>4</td>
<td>-</td>
<td>17</td>
<td>-</td>
<td>3</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td><strong>Stress</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>No stress</td>
<td>42</td>
<td>13</td>
<td>10</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>66</td>
</tr>
<tr>
<td>Pre-vocalic</td>
<td>7</td>
<td>12</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>24</td>
</tr>
</tbody>
</table>
Table 23. Cam09 data

<table>
<thead>
<tr>
<th></th>
<th>[wə]</th>
<th>[wər]</th>
<th>[wʊ]</th>
<th>[wʊz]</th>
<th>[wʊ:(r)]</th>
<th>[wʊznt]</th>
<th>[wʊ:(r)nt]</th>
<th>[wʊnt]</th>
<th>[wʊ:(r)nt]</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grammatical subject</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Was contexts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>8</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>12</td>
</tr>
<tr>
<td>he/she/it</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>10</td>
</tr>
<tr>
<td>Sing. NP</td>
<td>13</td>
<td>-</td>
<td>-</td>
<td>17</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>34</td>
</tr>
<tr>
<td>Sing. existential</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td><strong>Were contexts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>you</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>8</td>
</tr>
<tr>
<td>we</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>they</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>8</td>
</tr>
<tr>
<td>Plural NP</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>Plural existential</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>23</td>
<td>-</td>
<td>-</td>
<td>41</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>10</td>
<td>-</td>
<td>85</td>
</tr>
<tr>
<td><strong>Stress</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>12</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>10</td>
<td>-</td>
<td>33</td>
</tr>
<tr>
<td><strong>No stress</strong></td>
<td>23</td>
<td>-</td>
<td>-</td>
<td>29</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>52</td>
</tr>
<tr>
<td><strong>Pre-vocalic</strong></td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>23</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>[wɔ]</td>
<td>[wər]</td>
<td>[wɔ]</td>
<td>[wɔ:z]</td>
<td>[wɔ:(r)]</td>
<td>[wɔ:(r)nt]</td>
<td>[wɔnt]</td>
<td>[wɛ:(r)nt]</td>
<td>TOTAL</td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
<td>--------</td>
<td>----------</td>
<td>-----------</td>
<td>--------</td>
<td>-----------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td><strong>Grammatical subject</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Was contexts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>I</em></td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><em>he/she/it</em></td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><em>Sing. NP</em></td>
<td>10</td>
<td>-</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td><em>Sing. existential</em></td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>Were contexts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>you</em></td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><em>we</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><em>they</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><em>Plural NP</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><em>Plural existential</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>17</td>
<td>-</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td><strong>Stress</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>No stress</em></td>
<td>17</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td><strong>Pre-vocalic</strong></td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>