Ampersand 5 (2018) 1–17

Contents lists available at ScienceDirect

Ampersand

journal homepage: www.elsevier.com/locate/amper

The English Dialects App: The creation of a crowdsourced dialect corpus

Adrian Leemann[a,*], Marie-José Kolly[b], David Britain[c]

[a] Department of Linguistics and English Language, Lancaster University, County South, Lancaster, LA1 4YL, United Kingdom
[b] Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zürich, Switzerland
[c] Department of English, University of Bern, Länggassstrasse 49, 3012 Bern, Switzerland

A B S T R A C T

In this paper, we present the English Dialects App (EDA) and the English Dialects App Corpus (EDAC). EDA is a free iOS and Android app, launched in January 2016 that features a dialect quiz and dialect recordings. For the quiz, users indicate which variants of 26 words they use and the application guesses their local dialect; for the recordings, users can record a short text. The result is EDAC which includes metadata on mobility, ethnicity, age, educational level, and gender. More than 47,000 users from across the UK have indicated dialect variants for these 26 words, and more than 3,500 users have provided audio recordings. Unavoidably, EDAC does not successfully reflect distributions of age, ethnicity, qualification levels, and other parameters found for the UK population given that smartphone-based research reaches a specific stratum of the population. Yet there are also clear benefits to the sampling strategy used – benefits and pitfalls are discussed in this article. Future analyses will provide the most comprehensive understanding of English regional dialect variation since the work of the traditional dialectologists. We showcase two such analyses in this article. EDAC should, we demonstrate, be of interest to researchers in dialectology but also in forensic phonetics.

1. Introduction

The most recent nationwide dialect corpus for England dates back at least half a century and is based on a geographically broad but socially restricted sample of largely non-mobile, older, rural, male speakers (NORMs) (cf. [1]). The age of this corpus and its restricted speaker profile motivated the collection of a contemporary corpus of dialect and acoustic-phonetic data from across England. This new corpus, enabling us to update our knowledge of regional dialect distribution, has, furthermore, a number of important applications beyond dialectology and sociolinguistics, for example in forensic phonetics. In this paper we present the corpus and the collection protocol under which it was created, and demonstrate the potential of its dialectological and forensic applications.

The collection of large multi-locality corpora in dialectology has a venerable tradition that goes back to the 19th century. Traditional studies from that period relied mostly on questionnaires to elicit dialect material in lexis and phonology. Georg Wenker, for example, documented dialects in the late 19th century by distributing some 50,000 questionnaires with 40 test sentences to teachers across Germany. They were asked to transliterate 40 sentences into the local dialect of the community. Despite its age and methodological advances in the past century, the material collected continues to be used in contemporary dialectological research [2]. Towards the end of the 19th century, Jules Gilliéron sent out fieldworker Edmond Edmont to cycle across France to conduct hundreds of interviews between 1886 and 1900 [3]. This type of fieldwork by Wenker and Edmont typically resulted in linguistic atlases and lay the ground for future work on dialects in Switzerland, Italy, and Spain [4].

In England, meanwhile, the most significant advances in charting the nation's dialects were made by Alexander Ellis (see, especially [5], but also [6]). Like Wenker, Ellis sent out dialect transliteration tasks to people (usually clergy) – principally two short reading passages (one a story, the other a list of sentences) –, and, like Gilliéron, he was fortunate to have a trained phonetician, Thomas Hallam, to travel around the country collecting data and checking the transliterations. In all, data was collected from 1145 places across those parts of the British Isles in which English was the vernacular language in the mid-19th century. No atlas emerged from this endeavor, instead two maps of the islands' main dialect regions (but see [7]). With the exception of Kurath and Lowman [8], based on data collected in 1930, regional dialect documentation only reemerged onto the agenda in England once again after World War

Abbreviations: API, application programming interface; BKA, German Federal Criminal Police Office; BNC, British National Corpus; EDA, English Dialects App; EDAC, English Dialects App Corpus; FRED, Freiburg English Dialect corpus; ICE, International Corpus of English; NORMs, non-mobile, older, rural, male speakers; ONS, Office for National Statistics; SED, Survey of English Dialects

* Corresponding author.
E-mail address: a.leemann@lancaster.ac.uk (A. Leemann).

https://doi.org/10.1016/j.amper.2017.11.001
Received 10 August 2017; Accepted 21 November 2017
Available online 24 November 2017

II. Orton and Dieth [9] and a large team of fieldworkers collected data (on-the-spot phonetic transcriptions of answers to questions, fill-in-the-gap exercises, and so on, in a very long questionnaire) between 1950 and 1961 in 313 localities across England – the Survey of English Dialects (SED). To preserve or at least record 'the traditional types of vernacular English' ([9]; p. 14; see also [4]), fieldworkers interviewed mostly NORMs. The impact of the SED on the dialectology of England has been immense. Dialectologists and variationists have drawn upon the data, for example, to provide a historical backdrop for contemporary research (e.g. [10]) as well as to help ascertain the likely dialects spoken by 19th century colonial emigrants (e.g. [11]). Several atlas publications emerged from the SED (e.g. [12]), and because the data collection protocols were so systematically and carefully followed, the data have lent themselves to later computation using dialectometric techniques (e.g. [13]). Despite this, such traditional approaches to dialectology, anchored in the countryside, were criticized for their almost total abandon of the varieties spoken in urban areas. The advent of sociolinguistic approaches to dialectology in the 1960s saw radical changes in data collection methods, especially with respect to the nature of the sample (a wider spectrum of natives from the community were eligible for investigation), the type of data collected – relatively informal conversations within sociolinguistic 'interviews' – and the locations of investigation, a shift from rural areas and geographical coverage to the study of single urban locations. For a considerable period at that time, studies of geographical variation were few [1].

Today, the paper and pen techniques of the traditional dialectologists are gradually being replaced by large scale, computer-aided or mobile device-aided surveys. Since the advent and wide availability of computers, large-scale dialectological analysis has been aided by the construction of machine-readable dialect corpora, though it could be argued that none of these provide the systematic coverage of the Survey of English Dialects. Data in the British National Corpus (BNC), which contains 10 million words of spoken English, was categorized into 28 different dialects, but the BNC has limited value for dialectologists since it was not accompanied by audio [14]. A more recent attempt has been made to collect sound recordings from across the UK in the spoken corpus of BNC2014 [15]. The International Corpus of English (ICE) aims to chart different national varieties of English around the world. ICE, however, is not explicitly set up to investigate local or regional dialect *within* the country. The Freiburg English Dialect corpus (FRED [16]) is a 2.5-million-word collection of transcribed oral history recordings (mostly of NORMs) collated from holdings across Great Britain. Further, there is the Helsinki Dialect Corpus [17], another collection of NORMs, mostly from rural East Anglia and the South-west of England. Of these, only FRED comes close to nationwide coverage, however; but its sample was only slightly more recent than that of the SED: "we were looking for material preferably from the 1970s and 1980s, recording older speakers, or from the 1990s, if these recorded very old speakers" ([16]; p. 36).

Computer- or mobile device-based crowdsourced data collection efforts in the UK are in their infancy. MacKenzie et al. [18] studied phonological, lexical, and syntactic variation from around 5000 speakers across England who completed an online questionnaire disseminated by undergraduate students as part of a class module. Vaux's [19] Cambridge Online Survey of World Englishes crowdsources lexical, phonetic, and morphosyntactic variation in World Englishes (including British English) and has been online for more than a decade. The regional data collected is illustrated on maps on the website of the survey, but – to the best of our knowledge – has not been analyzed further or published yet. The maps show particularly high response rates in urban areas of the Southeast (around London) and the Northwest (around Manchester, Liverpool, and Leeds). Social media data, too, are starting to be used to examine regional variation. A first pilot study conducted by Willis [20] used a ten-day corpus of Welsh tweets to examine the pronoun *chdi*. Willis reports a similar distribution to that found by large-scale traditional surveys – however, with much smaller expenditure of time and money (see also [21,22]). Most recently, Grieve et al. [23] studied two billion words written by one million tweeters. Most tweets were geocoded with longitude and latitude information. They examined 35 lexical items and their 115 regional variants and compared regional distributions to the BBC Voices corpus [24]. They found that the regional variation reported in the Twitter corpus (collected in 2014) largely aligns with the variation found in the BBC voices data (collected in 2004/2005).

But, overall, while research activity on *individual* varieties of British English is undoubtedly healthy, we still know relatively little about contemporary variation at the regional and national level and about how these individual studies mesh together into a supralocal picture of the dialect landscape of the country. In this paper, we present the core functionalities of a free iOS and Android application – English Dialects App (hereafter EDA) – that was developed to generate a contemporary corpus of the English of England. We restricted the app to England because the app's dialect prediction mechanism (i.e. the gamification-approach that motivated users to provide us with their dialect data, see section 2) relied on there being a systematic historical corpus, with consistent coverage of the same variables from the same time period, that could be used as a comparative baseline. This does not exist for the British Isles as a whole – while surveys have been conducted of the linguistic varieties spoken in Wales, Scotland and Ireland, they were conducted at different times, with different methods, and different sets of variables. Using these corpora would distort the dialect prediction mechanism (it would be based on different kinds of data in different places), and systematic comparison would therefore not be possible. Further, in designing the app, we had to avoid overburdening users. Including all, especially Scottish and Irish variants of many of the variables, along with the English ones, would have made the app unwieldy, with too many variants for many of the variables. As will be demonstrated below, however, speakers outside England, too, participated and provided spoken data.

EDA's main functions are, firstly, to locate local and regional dialect characteristics via a quiz which 'predicts' users' dialects based on their responses and, secondly, to gather and make available nationwide audio data via users' uploading of self-recorded readings of a short story. Following the motto 'There's no data like more data' (cf. [25]), EDA has automatically collected dialect data from more than 47,000 speakers coming from more than 4900 localities across the UK and more than 3500 speakers from the UK have participated in the audio recording functionality over the course of less than one and a half years. Using an app as a dialect guessing tool is not new: it caught the public's interest in German-speaking Switzerland in early 2013 [26] and in the United States in late 2013, when the New York Times published a web-app – the 'Dialect quiz' [27] to predict a user's American English dialect. The US quiz consists of 25 questions such as, 'What do you call it when rain falls while the sun is still shining?'. The user provides their answer and proceeds to the next question. In the end, dialect location predictions are displayed. Posted on the Times's website within the last 10 days of 2013, this quiz became the year's most popular piece of content [28].

The first part of our paper is devoted to presenting the core functionalities of the English Dialects App: the dialect guessing quiz and the dialect recordings. The second part presents descriptive statistics of the outcome of this automated data collection, the English Dialects App Corpus (EDAC). We then discuss the use of EDAC for research and showcase some early results on language change and population statistics of acoustic parameters for the UK. It will become clear that the development of such a database is key to the discovery and quantification of linguistic phenomena on a hitherto unprecedented scale. We end the article with a discussion of the challenges and benefits of EDAC.

2. English Dialects App (EDA)

EDA's core functionalities were driven by the incorporation of