sciendo

ICAME Journal, Volume 45, 2021, DOI: 10.2478/icame-2021-0004

Better data for more researchers – using the audio features of BNCweb

Sebastian Hoffmann and Sabine Arndt-Lappe, Trier University

Abstract

In spite of the wide agreement among linguists as to the significance of spoken language data, actual speech data have not formed the basis of empirical work on English as much as one would think. The present paper is intended to contribute to changing this situation, on a theoretical and on a practical level. On a theoretical level, we discuss different research traditions within (English) linguistics. Whereas speech data have become increasingly important in various linguistic disciplines, major corpora of English developed within the corpus-linguistic community, carefully sampled to be representative of language usage, are usually restricted to orthographic transcriptions of spoken language. As a result, phonological phenomena have remained conspicuously understudied within traditional corpus linguistics. At the same time, work with current speech corpora often requires a considerable level of specialist knowledge and tailor-made solutions. On a practical level, we present a new feature of BNCweb (Hoffmann et al. 2008), a user-friendly interface to the British National Corpus, which gives users access to audio and phonemic transcriptions of more than five million words of spontaneous speech. With the help of a pilot study on the variability of intrusive r *we illustrate the scope of the new possibilities.*

1 Introduction

The aim of this paper is threefold: First, on a rather basic level, the paper is intended to provide an overview of the functionality of a new feature of BNCweb (Hoffmann et al. 2008), a user-friendly interface to the 100-million word British National Corpus (BNC, Burnard 2007). This feature gives users access to the audio and the phonemic transcriptions of more than five million words of spontaneous speech. Secondly, and we believe much more importantly, this paper aims to foster an increased level of interaction between what appear to be two fairly distinct schools of linguistics. In a deliberate oversimplification,

we will be referring to these two schools as 'corpus linguists' and 'speech-oriented linguists', respectively; see further explanations below. We aim to show that the new feature of BNCweb has the potential to be interesting and relevant for both groups that we are hoping to address. Finally, the paper is also a call for the creation of additional tools and resources that further develop the possibilities offered by the intended rapprochement of the two groups.

1.1 'Corpus linguists' and 'speech-oriented linguists'

The impetus for writing this paper stems from discussions between the first and the second author that revealed clear differences in our understanding of the label 'corpus linguist'. In a nutshell, the first author's conceptualisation of the term seemed – at least initially – to exclude the second author from being a referent of this label while the second author felt rather puzzled by this exclusion. Eventually, we realised that the source of our misunderstanding relates to the research communities we belong to, and the kind of concepts and practices we take for granted as a result of this association.

The first author is part of a research community that has employed electronic corpora, i.e. 'collection[s] of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language' (Sinclair 1996), for almost half a century and that has seen the field of corpus linguistics develop from a niche approach to being mainstream. The longest-running annual corpus-linguistic conference, ICAME (i.e. the International Archive of Modern and Medieval English), will meet in 2021 for the 42nd time, and the kind of research that would be presented at a typical ICAME is nowadays carried out by a much larger group of people than could ever attend a single conference.

The point we would like to raise here is that within this research community, i.e. what the first author would – perhaps too restrictively – refer to as 'corpus linguists', research in phonology and morpho-phonology is clearly a lacuna. Thus, in the main programmes of recent ICAME conferences, few papers have dealt with this topic area, thereby continuing the trend set by earlier meetings.[1] It simply isn't a topic that appears to be deemed relevant by much of the ICAME community.

At the same time, there is a varied, and growing, number of researchers who – like the second author – extensively work with speech data, i.e. acoustic data and transcriptions of such data, from corpora such as the Buckeye Corpus (Pitt et al. 2007), the Switchboard Corpus (Godfrey and Holliman 1997) and the ONZE corpus (Gordon et al. 2007). Of course these researchers are also corpus linguists; however, they are clearly part of a different research community (or

even research communities) and they typically also employ different tools to query and analyse their data. Thus, with the greater availability of pre-processed speech data that came with technological advancement (e.g. the development of automatic forced alignment), speech corpora have increasingly become important sources of evidence for researchers working, especially, in psycholinguistics, sociolinguistics, phonology and its interfaces (esp. morphophonology and the syntax-phonology interface), and computational linguistics. In these disciplines data resources that fall into the category 'corpora' have increasingly been employed to complement experimental data and data from databases such as CELEX (Baayen and Piepenbrock 1995). The advent of such data resources has also played an important role in the development of somewhat more recent linguistic disciplines such as sociophonetics (e.g. Thomas 2010; Baranowski 2013), Corpus Phonology (e.g. Durand et al. 2014; Durand 2018) or Laboratory Phonology (e.g. Cohn et al. 2011).

Our paper is intended to encourage and enable interaction between these two very different research communities, who will in what follows be distinguished by the labels 'corpus linguists' and 'speech-oriented linguists', even though the former may carry a different semantics for many readers than the one just introduced, and even though the latter label may be unduly vague and at the same time suggestive of a degree of homogeneity that readers may feel is unfitting, given that we use the label 'speech-oriented linguists' for researchers working in different disciplines (psycholinguistics, sociociolinguistics, phonology, phonetics, and others). This increased level of interaction is likely to reveal that there is much more common ground than suggested by our oversimplified characterisation of the two research communities.

1.2 Corpora of spoken data – and their use by the two research communities

The relationship between the first author's type of corpus linguists and phonology has always been a fairly tenuous one. While much of corpus-linguistic research has been conducted on written data, there are plenty of spoken resources, too, and few – if any – corpus linguists would discount the importance of segmental and suprasegmental phonological phenomena in understanding the nature of spoken conversation at all levels of linguistic description (e.g. the interface between intonation and pragmatics and the connection between phonetic reduction and information status, to name just two obvious contexts). However, few spoken corpora exist in a format that makes it possible to give justice to this importance in quantitative analyses of speech. In fact, the vast majority of spoken corpora are available in the form of written transcriptions only, which are necessarily a reduced representation of the original speech events.

The compilers of the pioneering London Lund Corpus (LLC, Svartvik and Quirk 1980) were very much aware of this reductive nature of the transcription process. They therefore tried to minimise its impact by providing a sophisticated set of annotations for prosodic features. However, this in turn made it much more difficult to search the data with 20th-century concordancers, which were not designed to handle this level of annotation.[2] Perhaps also as a result of this, many of the spoken corpora released in the 1990s and the first decade of the 21st century did not include this type of annotation. Thus, the original release of the seminal British National Corpus in 1994, with its more than ten million words of spoken data, included basic orthographic transcriptions only, and the annotation of speech-specific phenomena was restricted to some dysfluencies (e.g. truncated words) and indications of speaker overlap. Access to the original audio tapes was only possible by paying a visit to the British Library. Another example is the International Corpus of English (ICE, Greenbaum 1996), which has a growing number of national one-million word subcorpora (60 per cent of which are from spoken interactions), but whose transcriptions for the most part contain no annotations for phonological (segmental or prosodic) features.[3] Furthermore, the British subcorpus (ICE-GB) and ICE-Nigeria are the only components for which audio data is available.[4] Even if a corpus was distributed with its corresponding audio recordings, most concordancers available to the corpus-linguistic community lacked – and still lack – the ability to work with aligned audio data.[5]

In the meantime, such technical limitations should in theory no longer have a limiting effect. The advent of SGML, and its more recent offspring XML, has made it possible to store complex levels of corpus annotation and there is a range of tools that can make use of these annotations. Furthermore, aligning audio – or multimodal – data with transcriptions no longer poses particular technical difficulties. Nevertheless, the degree to which aspects of pronunciation, stress and intonation are annotated in spoken corpora continues to be fairly limited and there are still relatively few corpus resources that give access to audio data and the corresponding transcriptions. The newly released spoken component of BNC 2014 (Love et al. 2017) is a case in point: transcriptions are again purely orthographic and no audio data is available.[6]

A number of corpus resources exist that have been specifically compiled for the analysis of segmental and suprasegmental speech phenomena. A prominent example is the Buckeye Corpus (Pitt et al. 2007), which provides access to high-quality recordings of 40 speakers of American English (about 300,000 words), and which is orthographically transcribed and phonetically labelled. Using the tool SpeechSearcher (Dehdari et al. 2009), it is possible to search the corpus for

word and phone segments in sophisticated ways (e.g. using regular expressions) and to restrict these searches to a particular (type of) speaker. Results can be exported for further analysis with specialised software (e.g. Praat; Boersma and Weenink 2020). However, a quantitative analysis of these results will often require a considerable level of specialist knowledge (e.g. the ability to script) to access the full richness of the information contained in the data.

A second example is the Switchboard Corpus (Godfrey et al. 1992), which contains about 2,400 spontaneous telephone conversations between 543 speakers of American English recorded in 1991. The participants, who did not know each other personally, chose from a range of predetermined topics that were discussed on average for six and a half minutes, resulting in a corpus of approximately 3 million words. The corpus exists in several versions with various types of annotation, including a phone-level transcription and annotations of prosody, focus/contrast, syntactic structure, and word-class. More recently, a subset of the recordings containing just over 830,000 words was released that combines all levels of annotation in the NXT-format Switchboard Corpus (Calhoun et al. 2010; see also http://groups.inf.ed.ac.uk/switchboard/index.html), giving users unprecedented opportunities to investigate how different linguistic levels of description interact in the corpus. However, while a basic graphical user interface and command line tools are provided, considerable technical expertise is again required to harvest the full potential of the corpus. Other spoken corpora have been used to analyse phonological phenomena, but access to these sources is often restricted. This is for example the case for the ONZE corpora (Gordon et al. 2007), which can only be queried locally at the University of Canterbury, New Zealand.[7]

In recent years, dramatic progress in the area of automatic speech recognition (ASR) and the improved accuracy of forced alignment tools have made it much easier to compile large repositories of spoken data that contain segmental and suprasegmental phonological information. The automatic detection of gestures, head movements and gaze further expand the researcher's options and today the largely automated corpus creation of multi-modal data is within reach (see e.g. Steen et al. 2018; Uhrig 2018). However, as was the case for the Buckeye and Switchboard corpora, considerable technical skills are are currently still required to make full use of the resources.

The Buckeye and Switchboard corpora, and similar speech corpora, have rarely been used by the research community we have labelled 'corpus linguists'. This can perhaps best be explained by the fact that they are specialised corpora that have limited application in answering the type of questions that require data from a variety of contexts and speakers. At the same time, 'speech-oriented lin-

guists' have not employed the kind of data traditionally used by 'corpus linguists' from the ICAME community and beyond, for the simple reason that these corpora typically do not contain the kind of annotation (e.g. phonetic transcription, time alignment) that is essential for this type of work. This is, we believe, an unfortunate situation, because even though the research communities are fairly distinct, they share many interests in the same type of questions. To name only two prominent areas: members of both groups carry out research on the nature of the mental lexicon, and both are interested in sociolinguistic variation; many other areas of joint interest could no doubt be found. Therefore, combining their data sources, tools and methods to a greater extent than is currently the case would, we believe, be beneficial to both communities.

In what follows, we will present a new feature of BNCweb, a tool that has so far predominantly been used by the 'corpus linguistic' community. This new feature provides access to large amounts of authentic and spontaneous speech data in a way that we believe will be of interest to both 'speech-oriented linguists' and 'corpus linguists'. As we will show on the basis of a pilot study of intrusive *r*, it significantly extends the opportunities for research on natural conversational speech as represented in the British National Corpus and therefore invites further cooperation between the two research communities.

2 The BNC, BNCweb and the Audio BNC

The 100-million word British National Corpus was compiled to be a balanced reference corpus of late 20th century British English. More than 25 years after its initial release in 1994, the BNC continues to be the most widely used reference corpus of the variety. About ten per cent of the corpus – 10,409,858 words to be precise – make up the spoken part, which is further divided into the context-governed component (with four domains: Educational and informative, Business, Institutional and Leisure, approx. 6.2 million words) and the demographically sampled component (approx. 4.2 million words), which comprises spontaneous conversations recorded by a set of respondents who were chosen to be representative of a cross-section of British society in the early 1990s. For many speakers, a whole range of demographic data is annotated in the corpus (e.g. sex, age and social class of speaker), and further information is often available about the type of interaction (dialogue, 85% or monologue, 15%) and the region in which the material was collected (North, South and Midlands). As mentioned above, the corpus was initially released in the form of orthographic transcriptions only.

BNCweb is a user-friendly web-based interface that was created to search and process the data contained in the BNC. It provides easy access to a wide

range of functions that make it possible to linguistically analyse the results of corpus queries (e.g. collocations, distribution across metatextual categories, sorting of query results). Originally developed at the University of Zurich (see Lehmann et al. 2000), the functionality of BNCweb was further extended by the first author and Stefan Evert in the first decade of the 21st century (see Hoffmann and Evert 2006, Hoffmann et al. 2008). Since then, the tool saw little development, given that CQPweb, a more modern version of the interface, had been created by Andrew Hardie, which is not restricted to the use with the BNC but which can search any corpus that meets the necessary format requirements (see Hardie 2012). The feature described in the present paper is the first new functionality added in a long time. BNCweb is available for free via a web server at Lancaster University, but requires registration before use (http://bncweb.lancs.ac.uk/bncwebSignup/user/register.php). Readers are invited to replicate the individual steps of analysis described in our paper in order to get a first-hand experience of the new functionality.

The original audio tapes of the spoken component of the BNC were stored at the British Library Sound Archive. In 2009-2010, approximately 1,400 hours of recordings were digitized in cooperation with the Oxford Phonetics Lab and anonymized in accordance with the permission agreement with the contributors. In a final step, the digitized recordings and the BNC transcriptions were matched using an automated forced alignment process, which has resulted in a sizeable portion of the spoken component – approximately 5.4 million words – that is both aligned and phonemically transcribed. In the remainder of this paper, this resource will be referred to as the Audio BNC; for further information, see Coleman et al. (2011) and Coleman et al. (2012). Both the audio files and the corresponding alignment information, in the form of Praat TextGrid files, are available from http://www.phon.ox.ac.uk/AudioBNC. Together, they form the basis for the new audio feature in BNCweb.

3 Intrusive r

As mentioned at the end of Section 1, the new functionality of BNCweb will be introduced with the help of a pilot study to demonstrate its applicability to both 'corpus linguists' and 'speech-oriented linguists'. The feature chosen for this purpose is a form of *r*-sandhi, a connected speech phenomenon in which a syllable-final <r> appears before vowel-initial words (and morphemes) to avoid hiatus in non-rhotic varieties of English. Although this is not universally acknowledged (see e.g. Trudgill 1974), previous research has shown that the use of *r*-sandhi is variable (cf. e.g. Bauer 1984; Foulkes 1997; Hannisdal 2006;

Mompeán and Mompeán-Guillamón 2009; Pavlik 2011, 2016 on British English). There are two types of *r*-sandhi: non-etymological intrusive *r* as in example (1), where an /r/ is inserted although the word *idea* does not end in the grapheme <r>, and etymological linking *r*, as in example (2), where the final <r> in *hear* is pronounced as a result of hiatus avoidance.

(1) Now [pause] the *idea [r] is* that er they put a couple of monitors and monitor your brainwaves. (BNC FLY:608)

(2) You can *hear [r] it* yeah, but the only way to hear it (BNC HMD:1396)

In what follows, we will be exclusively concerned with non-etymological intrusive *r*. In addition, we will only consider cases that involve two orthographic words and will therefore disregard intra-word instances of *r*-sandhi (e.g. *draw[r]ing*). The following criteria can be used to define potential instances of intrusive *r* in a spoken corpus:

The first word:
- ends in /ɑː/, /ɔː/, /ə/, /ɪə/, /ʊə/ or /aɪə/
- the written form does not end in *-r* or *-re*
- must not be an interjection / filled pause or similar vocalisation
- further lexical exclusions: *the, a*, reduced, i.e. [ə]-final forms of *to, into, onto*

The second word:
- starts with a vowel
- must not be a filled pause (*er* or *erm*)

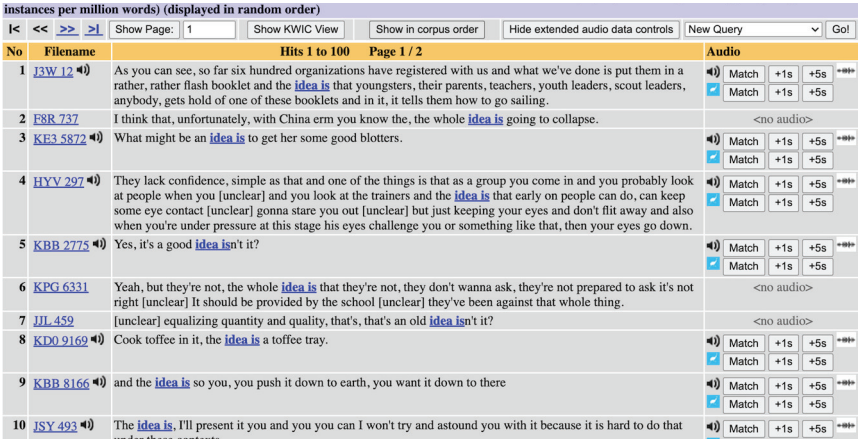
Given the phone-based definition of the ending of the first word, this type of study necessarily requires a corpus whose contents are (also) phonetically or phonemically transcribed. Once retrieved, all instances have to be manually checked and irrelevant contexts discarded; the remaining relevant items then have to be annotated for the presence or absence of intrusive *r*.[8]

4 The new audio features of BNCweb

4.1 General functionality

As a result of the integration of the Audio BNC, BNCweb now provides easy access to the audio for roughly half of the spoken component of the BNC. We will demonstrate the basic functionality of the new feature on the basis of a search for *idea is*. This sequence of words was chosen as one that typically invites the use of intrusive *r* and therefore seemed appropriate as a first look at

the phenomenon (e.g. to confirm its variability and to test the usability of the BNC data for answering our research questions). Figure 1 shows the first ten hits of the query result for *idea is*. The query was restricted to the spoken component and has 109 hits, corresponding to a frequency of about 10.5 pmw. The query result is displayed in random order.



instances per million words) (displayed in random order)

I< << >> >I Show Page: 1 Show KWIC View Show in corpus order Hide extended audio data controls New Query Go!

No	Filename	Hits 1 to 100 Page 1 / 2	Audio
1	J3W 12	As you can see, so far six hundred organizations have registered with us and what we've done is put them in a rather, rather flash booklet and the idea is that youngsters, their parents, teachers, youth leaders, scout leaders, anybody, gets hold of one of these booklets and in it, it tells them how to go sailing.	Match +1s +5s / Match +1s +5s
2	F8R 737	I think that, unfortunately, with China erm you know the, the whole idea is going to collapse.	<no audio>
3	KE3 5872	What might be an idea is to get her some good blotters.	Match +1s +5s / Match +1s +5s
4	HYV 297	They lack confidence, simple as that and one of the things is that as a group you come in and you probably look at people when you [unclear] and you look at the trainers and the idea is that early on people can do, can keep some eye contact [unclear] gonna stare you out [unclear] but just keeping your eyes and don't flit away and also when you're under pressure at this stage his eyes challenge you or something like that, then your eyes go down.	Match +1s +5s / Match +1s +5s
5	KBB 2775	Yes, it's a good idea isn't it?	Match +1s +5s / Match +1s +5s
6	KPG 6331	Yeah, but they're not, the whole idea is that they're not, they don't wanna ask, they're not prepared to ask it's not right [unclear] It should be provided by the school [unclear] they've been against that whole thing.	<no audio>
7	JJL 459	[unclear] equalizing quantity and quality, that's, that's an old idea isn't it?	<no audio>
8	KD0 9169	Cook toffee in it, the idea is a toffee tray.	Match +1s +5s / Match +1s +5s
9	KBB 8166	and the idea is so you, you push it down to earth, you want it down to there	Match +1s +5s / Match +1s +5s
10	JSY 493	The idea is, I'll present it you and you you can I won't try and astound you with it because it is hard to do that	Match +1s +5s

Figure 1: The first ten hits of a query result for idea is *in the spoken component of the BNC*

For seven of the ten instances, extended audio controls are shown to the right of the query hit.[9] By clicking the buttons in the upper row, users can listen to the audio of the query match and its immediate context (+/- 1 second, +/- 5 seconds). This will reveal that only some of the audio data is perfectly aligned; this is the case for query hits 8, 9 and 10. In some other cases (e.g. hit no. 3), the audio extract that is played covers slightly more than the words that are matched; in the case of hit 3, another speaker can be heard towards the end of the extract played. The fifth hit, however, displays slightly more serious issues since the extract played when clicking on the left-most button of the extended audio controls contains other words than the ones of the query term. A click on the button labelled '+ 1 second' will, however, also play the correct audio for *idea is*. In other words, the alignment is off by about a second. If users click on the query match, they are taken to a display of the larger context – see Figure 2. Here, too, the audio can be played, either by clicking on the audio controls in the top right corner or by clicking on individual words in the text.

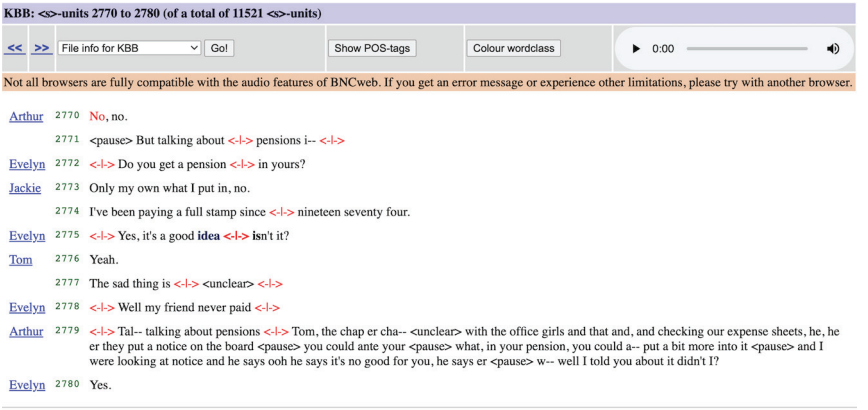


Figure 2: The larger context display of the fifth hit for the query idea is

The corpus extract in Figure 2 is a conversation between four people, and some overlap occurs (indicated by the symbol "<-|->" in the transcription shown in Figure 2). This kind of data naturally poses challenges to the automated forced alignment process, which cannot assign portions of the audio to several speakers at the same time. For users of BNCweb who wish to play only the correct extract of the audio for *idea is*, an additional range of options is available by clicking on the small picture of a wave form that can be found to the right of the query hit (i.e. the fourth element in the upper row of the extended audio controls). This feature allows users to select the exact length of the audio clip to be played, thus making it possible in cases of more severe misalignment to play audio outside the maximum window provided by the audio control buttons.

Figure 3 displays the wave form of the audio for the 5th hit of the query result after adding 2.5 seconds on both sides of the – slightly faulty – time-stamps for the query hit. In addition, a part of the wave form has been selected using 'click and drag'; this region corresponds to the exact position in the audio where the words *idea is* are uttered. Regions can be flexibly moved and shortened/lengthened until the exact offsets are located. Once a region has been created, clicking the 'play'-button will only play the selected portion of the audio.[10] In addition, the feature offers the functionality of changing the audio speed of the recording down to 50 per cent without changing its pitch. In this way, auditory analyses might detect features that would escape the analyst when the audio is played at full speed. A combination of the larger context display (Figure 2)

with the wave form feature is likely to ease the detection of the corresponding audio in misaligned contexts. We will return to an additional functionality offered by this feature in Section 4.4, which presents the option available to correct faulty audio alignment.

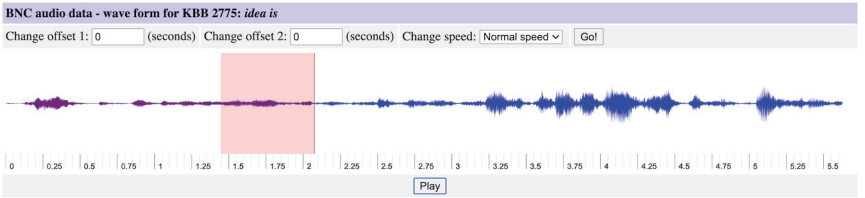


Figure 3: The fifth hit for the query idea is *with extra 2.5 seconds on both sides and a user-defined region to play the relevant portion of the audio*

Hits 1 and 4 (cf. Figure 1) of the query for *idea is*, finally, represent an issue that cannot be resolved by checking the larger context of the query hit. Unfortunately, a number of tapes among the original BNC recordings could not be successfully matched with their corresponding transcriptions. This situation may improve in the near future when resources become available to carry out more extensive work on the data.

The second row of the extended audio controls (cf. Figure 1) offers access to a user-customizable spectrogram provided through the EMU-webApp (Winkelmann and Raess 2014) hosted by the Institute of Phonetics and Speech Processing, LMU. This feature also offers access to phoneme-level segmentation and includes tiers for the phonemic transcription, the stress pattern and the number of syllables of each word; see Figure 4, which displays +/-1 seconds of the eighth query hit shown in Figure 1.

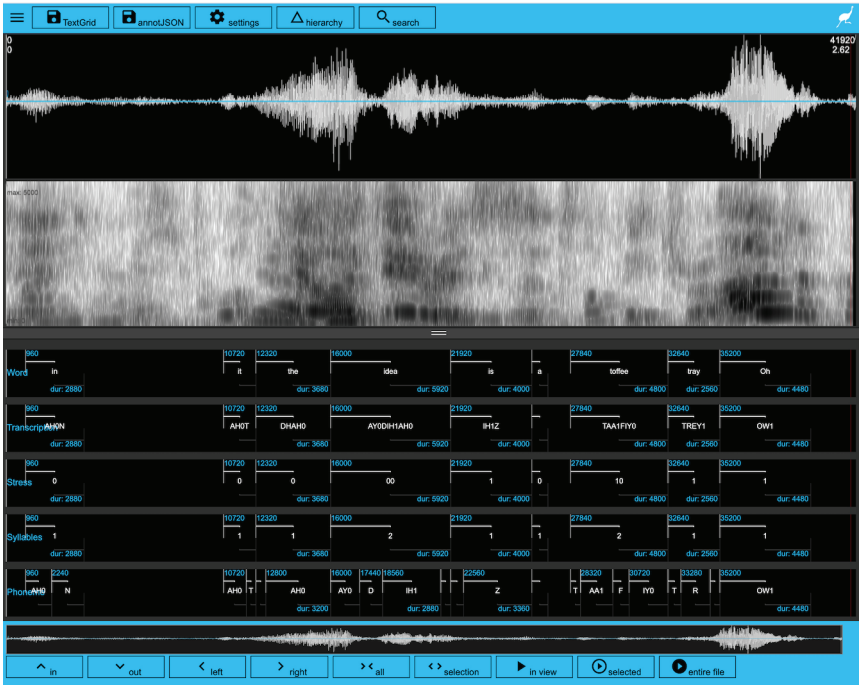


Figure 4: The eighth hit of the query result +/- 1 seconds in the EMU-webApp

Finally, users of BNCweb also have the option of downloading the audio clips of query matches (and their immediate context, i.e. +/- 1, +/- 2, and +/- 5 seconds) for further analysis in tools such as Praat, together with corresponding TextGrid files and/or various tab-delimited files with alignment information. This feature is available via the 'Download…' option in the drop-down menu in the top right corner of the query result page; the relevant section is found at the bottom of the 'Download concordance' page – see Figure 5:[11]

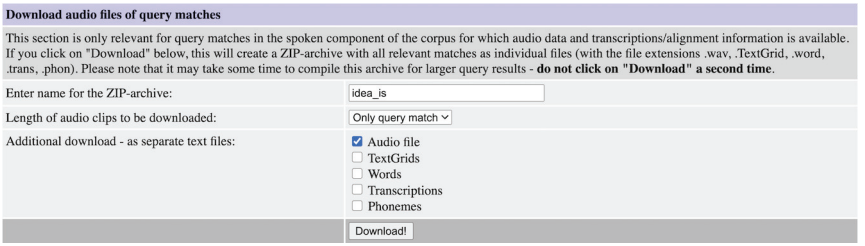


Figure 5: The audio download options available on the 'Download concordance' page

Even a cursory glance at the hits retrieved by the query for *idea is* provides ample confirmation of the feasibility of studying intrusive *r* with the help of the data contained in the BNC. Thus, the hits for which correctly aligned audio data is available corroborate the fact that the phenomenon is indeed variable. As a case in point, consider hits 8 and 9. The first of these two represents a clear case of an intrusive *r* while the latter does not, even though there is no pause between the two words. Furthermore, it is interesting to note that hiatus avoidance phenomena also occur across suprasegmental units of prosody. This is confirmed by hit 5, where an intrusive *r* is found between the anchor and the tag of a tag question even though the two are realised as two separate tone units. In what follows, we will introduce some additional functionality of the new audio feature of BNCweb that makes it possible to carry out an exhaustive analysis of the phenomenon at hand.

4.2 Additional index layers in BNCweb and sample searches

As a result of the integration of the Audio BNC into BNCweb, three new index layers are available in searches of the 5,418,016 words for which relevant information is available:

- phonemic transcriptions (`phon`)
- syllable counts of words (`n_syll`)
- stress patterns (`stress`)[12]

This functionality is only available for searches using the powerful Corpus Query Processor query syntax – referred to as 'CQP-syntax' (cf. Evert 2009) – which must be selected in the drop-down menu on the main query page in BNCweb to change the query mode. For users of BNCweb who are not familiar with CQP-syntax, it may be helpful to consult the 'Query history' feature, which

gives access to the internal CQP-query format of previous queries that were performed in Simple query mode.[13] For more extensive information about searching the BNC with CQP-syntax, readers are referred to Chapter 12 in Hoffmann et al. (2008).

The following sample searches exemplify the functionality of the added index layers:

(3) Phonemic transcription – see further Section 4.3 below:

```
[phon=".*AH0"]
```

→ finds all words ending in a schwa

(4) Syllable count:

```
[n_syll="8"]
```

→ finds all words with exactly 8 syllables

(5) Stress pattern:

```
[stress="1020"]
```

→ finds all words with four syllables that have primary stress on the first syllable and secondary stress on the third syllable

(6) Searches can combine various features:

```
[class="ADJ" & phon=".*(L|N)" & stress=
"\d{1,}10"]
[class="SUBST" & stress="1\d{1,}"]
```

→ finds adjectives whose last phoneme is a liquid, with at least three syllables, with primary stress on the penultimate syllable (followed by an unstressed syllable), immediately followed by a noun with at least two syllables that is stressed on the first syllable.[14]

At the time of writing, there is no straightforward way of restricting searches to only the portion of the spoken component for which audio recordings are available. The workaround is to include an element in a query that matches every single relevant word token for one of the three new index layers. This can for example be done with `n_syll="\d"`, which matches any token for which a syllable count is known. Thus, the following query will find all instances of *sure* in the 5,418,016 words for which audio recordings are available (%c means 'case insensitive'):

```
[word="sure"%c & n_syll="\d"]
```

4.3 ASCII coding of phone symbols used in the BNC transcriptions

As already suggested by sample query (3) above, the phonemic transcriptions provided in the Audio BNC do not employ the IPA but are based on a system of ASCII coding; the full set of codes is shown in Table 1, which has been reproduced in a slightly adapted format from the Audio BNC website (Coleman et al. 2012).[15] The forced aligner used pronunciation dictionaries as a reference, which were made specifically for the BNC and did contain some pronunciation variants, e.g. to reflect regional variation. In addition to the phone symbols, each vowel is coded for its stress value (0 – no stress; 1 – primary stress; 2 – secondary stress).

Table 1: ASCII coding of phone symbols used in the BNC transcriptions

Consonants		**Short vowels**	
IPA	**ASCII**	**IPA**	**ASCII**
p	P	ɪ	IH (Tensed unstressed [i] may be coded IY0)
b	B	ɛ	EH
t	T	a	AE
d	D	ə	AH0 (always unstressed)
ʧ	CH	ʌ	AH[12] (always stressed) For Midlands and Northern Varieties, UH[12] is used instead of AH[12]
ʤ	JH	ɒ	OH (often this is collapsed with AO, as we lack separate acoustic models)
k	K	ʊ	UH
g	G	**Long vowels and diphthongs**	
m	M	iː	IY
n	N	eɪ	EY
ŋ	NG	aɪ	AY
f	F	ɔɪ	OY
v	V	aʊ	AW
θ	TH	əʊ	OW
ð	DH	uː	UW
s	S	ɪə (ɪəˠ)	IH[12] AH0 (IH[12] ER0 in rhotic dialects)

ʃ	SH	ɛə (ɛɚ)	EH[12] AH0 (EH[12] ER0 in rhotic dialects)
ʒ	ZH	əː	ER[12] (even in nonrhotic varieties)
h	HH	ɑː	AA
r	R	ɔː	AO
l	L	ʊə	UH[12] AH0
w	W	1 – primary stress, 2 – secondary stress	
j	Y	0 – unstressed	

(Adapted from: http://www.phon.ox.ac.uk/files/docs/BNC_transcription_alphabet.html)

Table 2 displays a number of sample transcriptions to give readers an impression of both the ASCII coding and the kind of variation found in the data. The word *action*, for example, is found in two different transcription variants (AE1KSHN and AE1KSHAH0N), the first of which displays a syllabic consonant. In the case of the word *chicken*, however, the variants are CHIH1KAH0N and CHIH1KIH0N; there is no variant with a syllabic consonant.

Table 2: Sample transcriptions

Word	Transcription
theatre	THIH1AH0TAH0 THIY1AH0TER0
industrialization	IH0NDAH0STRIH0AH0LAY0ZEY1SHN
action	AE1KSHN AE1KSHAH0N
chicken	CHIH1KAH0N CHIH1KIH0N
joining	JHOY1NIH0NG
perpendicular	PER2PAH0NDIH1KYAH0LER0
woodworking	WUH1DWER2KIH0NG
idea	AY0DIH1AH0 AY0DIY1AH0
overreact	OW1VER0RIY0AE1KT
linguistics	LIH0NGGWIH1STIH0KS
close	KLOW1Z KLOW1S

4.4 Correcting the audio alignment

As shown in Section 4.1, some of the alignment information provided in the Audio BNC is faulty. For users who wish to have direct access to the exact audio extracts of their query results, BNCweb offers a way of correcting alignment information – on condition that the correct position can be found in the not-too-distant context of the query match. For the purpose of demonstration, we return to the query for *idea is* already used in Section 4.1. As noted before, the audio for the third hit covers slightly more than needed: the word *intuitive* is also heard on the audio clip.[16]

Correcting the audio alignment requires the use of the 'Categorize hits…' feature, which is available from the drop-down menu in the top right corner of the query result page. The main use of the feature is to manually annotate a query result according to user-defined categories (e.g. 'relevant', 'irrelevant', 'unclear'). Once the annotations are stored, the different categories can be saved as separate sets of data that can then be analysed with any of the post-processing features available in BNCweb; for further information on how to use the 'Categorize hits…' feature, please consult Chapter 9 in Hoffmann et al. (2008). In the context of the current task at hand, only one category needs to be added, and its label is irrelevant.

After creating a 'categorized query' of the 109 hits, the user is returned to the same concordance view of the query result – with an extra column that contains the choice of available categories. Using the fourth item in the upper row of the extended audio controls – i.e. clicking on the small picture of a wave form – it is now again possible to find and select the correct extract of the audio clip. For the current purpose, we extend the length of the audio by one second on both sides by entering '-1' in the left field and '1' in the right field. Once the page has reloaded, the exact position of *idea is* can be selected using 'click and drag'. This is shown in Figure 6, where the section from about 0.99 to 1.64 seconds of the clip has been selected. This time, however, an extra button with the label 'Update offsets' appears as soon as a region of the wave form is selected. Once this button is clicked, the alignment information is updated on the server and a confirmation message is shown – see Figure 7:

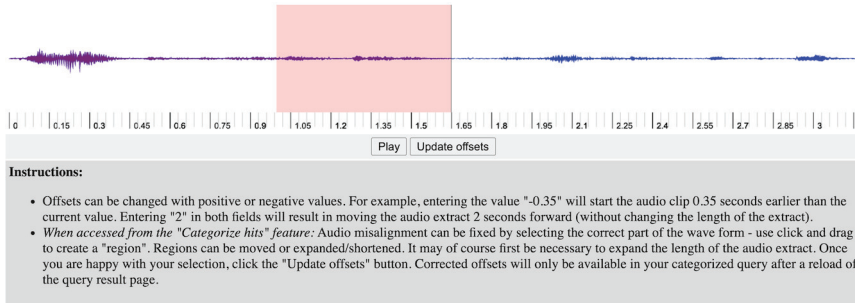


Figure 6: The correct region for the audio of the third hit for the query idea is

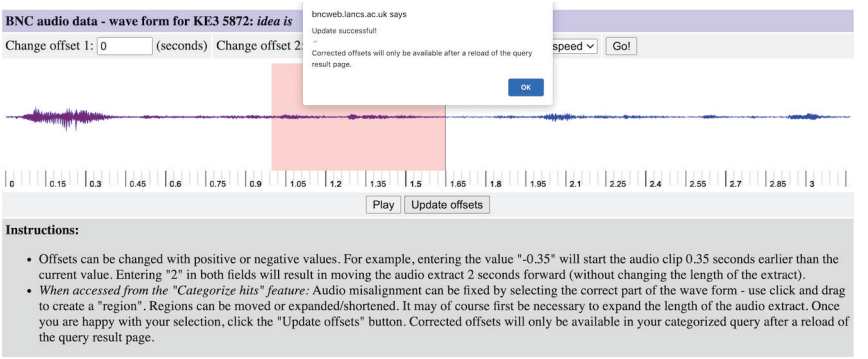


Figure 7: The correct region for the audio of the third hit for the query idea is

As indicated in this confirmation message, it is necessary to reload the query result page of your categorized query before the corrected alignment is available via the other buttons of the extended audio controls.

Once all alignments have been corrected, the corresponding audio of all hits in the query result can be downloaded to the user's hard-disk by clicking on the 'Download categorized query' button. As described towards the end of Section 4.1, the bottom part of the 'Download concordance' page is used to do this. It must be noted, however, that corrections to the audio clips only apply to the length of the audio itself. The alignment of individual word and phoneme segments is not changed. As a result, alignment of the downloaded TextGrid and audio files will be off in a tool like Praat, and will therefore require manual post-processing.

4.4 Problems in matching the Audio BNC with the BNC transcriptions

A number of problems emerged in the process of matching the TextGrid files provided by the Oxford Phonetics Lab with the transcriptions of the BNC. For example, the two differ in their tokenisation of the text. Thus, instances involving contracted negation (e.g. *haven't*, *don't*) are represented as a single token in the TextGrid files but they are separate tokens in the BNC. Furthermore, there were differences in the way dysfluencies and related phenomena of speech were treated. In many cases, for example, truncated words (e.g. false starts) were indicated as {OOV} in the TextGrid files, but they are actual (partial) word tokens in the BNC. Mismatches of this type made it difficult to link the two resources in a straightforward way. A number of heuristics had to be developed to counter these issues, and their implementation is certainly not 100 per cent successful.

More importantly, the data in the Audio BNC presented a number of issues that could not be resolved. For example, for some audio recordings, two TextGrid files with somewhat different phonemic transcriptions were found. In such instances, one of the two available files was randomly chosen. Finally, as already referred to in the discussion of the first ten hits of the query for *idea is*, there are a number of cases where the method of matching the audio recordings with their corresponding TextGrids failed to link up the correct files: although the file names suggest a match, the contents of the TextGrid file clearly correspond to a different recording. As a result of this, users of BNCweb will at times be presented with audio data that is clearly faulty. No complete assessment of the extent of these issues has so far been carried out, but it is clearly a relatively minor issue. Future releases of the audio features of BNCweb may improve this situation.

In any case, users must bear in mind that the alignment process was carried out fully automatically and on the basis of lexicon-based transcription models. Furthermore, although some phenomena of connected speech are represented (e.g. syllabic consonants), the actual realisation of words (including their stress patterns) may be very different from what is contained in the corpus. In other words, the features presented in this paper cannot be used without manually checking each and every instance against the audio data.

5 A pilot study on intrusive **r**

After having summarized the new functionality of BNCweb in Section 4, we would like to return to the topic of intrusive *r* and present the results of a pilot study that makes full use of the feature. For this purpose, the formal description

of potential hiatus avoidance contexts as presented in Section 3 have to be translated into a query in CQP syntax; this is shown in (7):

(7)
```
[phon = ".*(AO\d|AH0|AA\d)" & word=".*[^r]" %c &
word!=".*(a|e|i|o|u|t|y|\')re" &
word!="(to|the|a|du|erm|er|wa|ha|ah|aha|ahh|aah|aarg
h|ca|da|la|le|nah|sa|sha|tha|wha|into|yeah|onto)" %c
& phon !="NAH1MBAH0"] [phon = "(A|E|I|O|U).*" &
word!="(erm|er|s)"] within s
```

As a reminder, this query retrieves sequences of two words in which the first word ends in /ɑ:/, /ɔ:/, /ə/, /ɪə/, /ʊə/ or /aɪə/ – but does not end in <-r> or <-re> – and the second word starts with a vowel. In addition, a number of lexical exclusions (e.g. interjections, filled pauses) are defined in order to improve the precision of the query; a final constraint is that the sequence of words cannot span across a sentence boundary.[17] The query retrieved a total of 4,162 hits, of which a random subset of about 38 per cent were analysed for the present pilot study. The audio of about 95 per cent of all instances could be successfully located with the help of BNCweb; the 'Categorize hits…' feature was used to manually analyse and annotate the data.

5.1 Findings

Table 3 displays the proportion of intrusive *r* in the 1,254 potential contexts so far analysed.[18] The figure of just over a quarter of all instances tallies well with other published research on the phenomenon.

Table 3: The use of intrusive *r* in the spoken component of the BNC

Type	N	per cent
intrusive *r*	340	27.1 %
no intrusive *r*	914	72.9 %
Total	1254	100 %

Table 4 is concerned with the distribution of intrusive *r* over male and female speakers. It further distinguishes between the demographically sampled and the context-governed components. Interestingly, female speakers display a greater proportion of intrusive *r* use than male speakers, but this is much more pronounced – and also statistically significant – in the spontaneous conversations of the demographically sampled component. This is a relevant finding because there is disagreement in the literature about the preferred/dispreferred status of

intrusive *r*. Given the higher use by female speakers, an interpretation of intrusive *r* as a stigmatised form seems less likely.

Table 4: The use of intrusive r by male and female speakers in the demographically sampled and the context-governed components of the BNC

Type	intrusive *r*	no intrusive *r*	Total	% intrusive *r*
Spoken Demographic:				
male	50	170	220	22.7 %
female	105	231	336	31.3 %
Total	155	401	556	
Context-governed:				
male	103	308	411	25.0 %
female	22	57	79	27.8 %
Total	125	365	490	

The final finding to be reported concerns the kind of vowel found before an intrusive *r*. As Table 5 shows, there are considerable differences between the individual vowels. In other words, the data suggests that the occurrence of intrusive *r* is codetermined by the final vowel of the preceding word.

Table 5: The proportion of intrusive *r* after different vowels; infrequent vowels are excluded

Type	Example	intrusive *r*	no intrusive *r*	Total	% intrusive *r*
/ɔ:/	*dr<u>aw</u>*	28	239	267	9.7 %
/ɑ:/	*grandm<u>a</u>*	17	159	176	10.5 %
/ə/	*Americ<u>a</u>*	139	331	470	29.6 %
/ɪə/	*id<u>ea</u>*	153	181	334	45.8 %
Total		337	910	1247	27.0 %

Why should there be such a pronounced difference between the various vowels? One explanation could be the influence of high-frequency sequences such as *idea is*, which is by far the most frequent combination in the data, and which may therefore be more entrenched than combinations that are rarely found in spoken interaction. This may in turn have an influence on how it is produced.

However, another intriguing explanation can be found in the possible influence of an analogical process at work (cf. Soskuthy 2013 for a proposal along these lines). As Table 6 shows, there appears to be a strong correlation between the probability of intrusive *r* ("non-etymological *r*", see Section 3) in our data and the probability of encountering a word that is phonologically similar – i.e. it ends in the same vowel – but does not have a syllable-final <r> in its spelling ("etymological *r*"). In other words, the fact that the frequency of words ending in /ɪə/, without etymological *r* is not dramatically different from similar words with etymological *r* might support the use of intrusive *r*. For words ending in /ɔ:/, however, there is an almost 15-fold difference between the frequencies involved. Spelling can thus be seen to influence pronunciation.

Table 6: The proportion of intrusive *r* after different vowels and the proportion of words without and with etymological *r*

Final vowel	% intrusive *r*	N tokens without etymological *r* / N tokens with etymological *r*
/ɔ:/.	9.7	0.068
/ɑ:/	10.5	0.075[19]
/ə/	29.6	0.146
/ɪə/.	45.8	0.396

5.2 Evaluating the pilot study

As mentioned above, much more work is required to capture the full range of factors influencing the use of intrusive *r* in British English. However, the main point of this presentation was to show how the new feature of BNCweb can be fruitfully employed by the two research communities we wish to address. From the corpus linguist's perspective, this is yet another study on the (spoken component of the) BNC. It touches on the question of how sociolinguistic factors influence language use and hypothesises a connection between frequency and entrenchment that has an impact on how the phenomenon under study is employed. This is well within the remit of what a typical paper at an ICAME conference would be concerned with; it just happens to be on a phonological phenomenon.

For the 'speech-oriented linguist', too, there is little that wouldn't fit the agenda of their typical research modes. Admittedly, the data is to some extent more 'messy', i.e. less amenable to instrumental – and at times even auditory – analysis. However, this is, we believe, counterbalanced by the advantage of hav-

ing access to a significantly larger, carefully sampled dataset of authentic language use. As a case in point, consider Table 7, which gives an overview of previous quantitative/empirical studies on intrusive *r* in English. The table is an extended version of the table found in Pavlik (2016: 113). The middle column, labelled 'N instances', refers to the total number of potential instances of intrusive *r* in the data; given the variability of the phenomenon, the actual number is much smaller.

Table 7: Previous research on intrusive *r* based on speech data

Study	N instances	Type of data
Bauer (1984)	74	recordings of university students and staff
Foulkes (1997)	174	conversational data from sociolinguistic interviews
Hay and Sudbury (2005)	198	ONZE
Hannisdal (2006)	558	newsreaders
Mompeán and Mompeán-Guillamón (2009)	148	newsreaders
Mompeán and Gómez (2011)	399	newsreaders
Pavlík (2011)	300	newsreaders
Pavlík (2016)	613	newsreaders

As mentioned above, the query for potential instances of intrusive *r* in the BNC retrieved 4,162 hits, and based on the results of the pilot study, about 3,300 of these would likely offer an opportunity for intrusive *r* to occur. In other words, the data retrieved via the audio features of BNCweb is more than five times larger than the most extensive data set used in any of the studies shown in Table 7. More importantly, however, it is also much more varied: virtually all recent studies employ the output of newsreaders as their database.

5 Concluding remarks and outlook

As mentioned in the introduction, the present paper has three aims – two of which have so far been addressed. We have given a detailed description of the new audio features of BNCweb, which allow users to search a sizeable proportion of the spoken part of the corpus with the help of three additional layers of annotation (phonemic transcription, stress pattern and the number of syllables in a word). We have also shown how the partially faulty alignment information

provided in the data can be dealt with by users of BNCweb. Apart from offering ways to search for the correct audio in the immediate context of the query match, the tool also gives users the opportunity to correct the alignment more permanently. BNCweb also makes it possible to download audio data – and corresponding TextGrid files – to the user's computer.

The second of our aims was to show how this type of data can appeal to two strands of linguistics or research communities that we have claimed to be more distant than necessary, i.e. what we have termed 'corpus linguists' and 'speech-oriented linguists'. A pilot study on intrusive *r* was used to support our plea for further cooperation between the two groups.

The final aim naturally emerges from the previous two: assuming that we have successfully illustrated the potential of both the audio features of BNCweb and their utility for the two research communities, a call for further data and material that would benefit both groups can come as no surprise. Corpus linguists would more easily be able to work with spoken data, and speech-oriented linguists would gain access to more – and more varied – data. Compilers of new spoken corpora should therefore keep in mind the needs of researchers who are interested in phonological aspects. This first and foremost applies to the planning stage, when the necessary permission agreements are set up. Secondly, corpus compilers should factor in resources for aligning the audio with the transcriptions. Several projects are underway to improve existing forced alignment procedures and to make them compatible with additional varieties of English, and automatic speech recognition is advancing in large strides, placing within reach balanced and representative large-scale corpus resources with detailed annotation layers for speech analysis. Given the potential and attraction this type of data has, it may soon emerge that the two research communities we have juxtaposed in this paper are in fact not as distinct as we have made them out to be.

Notes

1. The current paper is based on one of the exceptions, Hoffmann et al. (2018), presented at Tampere in 2018 (ICAME 39). In the same spirit, the authors of this paper organised a pre-conference workshop on the topic of "English Corpus Phonetics and Phonology at ICAME" at ICAME 41 in Heidelberg.
2. In fact, versions of the corpus were soon in circulation that were stripped of the 'disruptive' annotation in order to make it more amenable to corpus-linguistic research. To the best of our knowledge, there is still no tool that

gives researchers full access to the prosodic information contained in the LLC.

3. The two major exceptions are ICE-Scotland and ICE-Nigeria, whose spoken components have been (partly) annotated phonemically – see Wunder et al. (2010); Gut and Fuchs (2017), Fuchs et al. (2019). With respect to suprasegmental features, a further notable exception is the pragmatically annotated version of ICE-Ireland, SPICE-Ireland (Kallen and Kirk 2011), which contains annotation of pitch movements for some of its texts.
4. The audio recordings for ICE-GB are sold separately and distributed on a set of 11 CD-ROMS. The corpus can be searched with ICE-CUP (Nelson et al. 2002), which also allows users to play the audio of particular text units in the corpus. For ICE-Nigeria, all speech data is freely distributed online.
5. The freely distributed 249.000-word Santa Barbara Corpus of Spoken American English (http://www.linguistics.ucsb.edu/research/santa-barbara-corpus) is an important exception. It is also available in CHAT-format (MacWhinney 2020a), which makes it possible to access both the audio data and the corresponding transcriptions with the help of the CLAN Program (MacWhinney 2020b). However, while the corpus texts are organised to represent prosodic units and contain a range of discourse-related annotations of (see Dubois et al. 1992), the transcriptions are again purely orthographic.
6. One reason for the lack of audio data is the fact that the recordings would have to be anonymised before the can be made public. For the time being, there is no funding available to carry out this process. (Andrew Hardie, p.c.)
7. It is worth mentioning, though, that the tool developed at the University of Canterbury to query ONZE, LaBB-CAT, is freely distributed as open-source project and can be used to query any spoken corpora available in a suitable format, including, for example, the Buckeye Corpus – see https://labbcat.canterbury.ac.nz/system/.
8. In the context of the present paper, we will not focus on the exact procedure involved in detecting intrusive *r*, which involved independent auditory analysis by two trained coders, supplemented by examination of spectral properties where possible, given the quality of the Audio BNC data. A comprehensive analysis of the phenomenon is currently underway and will be reported on in due course.
9. This functionality can be selected via a drop-down menu on the main query page or using a button on the query result page.
10. Regions can be deleted by way of a double-click; creating a second region will automatically delete the first one.

11. This download of audio files is currently limited to a maximum of 2,500 instances in order to avoid a browser time-out. If only text data (e.g. TextGrid files) is selected for download, this limitation does not apply. Depending on the size of a query result, preparing the ZIP-archive may take several minutes; users should avoid clicking on 'Download' a second time as this will only slow down the process and result in unnecessary data transfer between the servers involved.
12. Syllable counts and stress patterns for individual words are not provided as separate levels of annotation in the Audio BNC. The information was retrieved by way of a Perl script that analysed the phonemic transcriptions. Thus, each vowel contains a digit ('1' = primary stress, '2' = secondary stress and '0' = unstressed), and a regular expression was used to detect syllabic consonants. For each syllabic consonant, an unstressed syllable was added to the stress pattern of the word and the syllable count was increased by one.
13. The Simple query format is the default search mode; it gives access to all the most frequently-used features of CQP in a simplified format.
14. The regular expression `.*(L|N)` means 'zero or more instances of any character (indicated by the period sign), followed by either L or N', as in *school* or *pattern*. Curly brackets quantify the immediately preceding item; thus, `1\d{1,}` matches if the stress sequence begins with the number 1 and is followed by one or more digit characters (e.g. '10' as in *meeting* or '10200' as in *motorcyclist*).
15. We refer to the transcription as 'phonemic' in spite of the fact that they are in some respects more detailed than that. For example, some words show variation in the use of syllabic consonants; see Table 2. However, the transcriptions are clearly not optimised to represent connected speech phenomena (such as intrusive *r*).
16. Again, the misalignment can be explained by the nature of the audio data: A TV (or a radio) is playing in the background, but the words that can be heard do not form part of the BNC transcription.
17. The query also excludes the negation *no* transcribed as NAH1MBAH0 (*number*); this is a frequent error of the forced alignment process in the Audio BNC.
18. This figure excludes contexts where intrusive *r* was deemed impossible to occur, e.g. if a pause is found between the first and the second word. Pauses were defined as any period of absence of formant structure in the transition from the first to the second word, irrespective of length. Note that this procedure also eliminated intervocalic glottal stops.

19. The numbers for /ɑː/, involved some manual cleaning, e.g. the removal of interjections (e.g. *Yaaaaaa*) and mismatches between tokenisation and word type (e.g. *ca > can't*)

References

Baranowski, Maciej. 2013. Sociophonetics. In R. Bayley, R. Cameron and C. Lucas (eds.). *The Oxford handbook of sociolinguistics*, 403–424. Oxford: Oxford University Press.

Bauer, Laurie.1984. Linking /r/ in RP: Some facts. *Journal of the International Phonetic Association* 14: 14–79.

Boersma, Paul and David Weenink. Praat: doing phonetics by computer [Computer program]. Version 6.1.37, retrieved 01 January 2021 from http://www.praat.org/.

Burnard, Lou (ed.). 2007. *Reference guide for the British National Corpus (XML Edition)* Available at <http://www.natcorp.ox.ac.uk/docs/URG/> (last accessed 20.10.2020).

Coleman, John, Mark Liberman, Greg Kochanski, Lou Burnard and Jiahong Yuan. 2011. Mining a year of speech. In *Proceedings of New Tools and Methods for Very-Large-Scale Phonetics Research*, 16–19. Available at <http://www.phon.ox.ac.uk/jcoleman/MiningVLSP.pdf> (last accessed 20.10.2020).

Coleman, John, Ladan Baghai-Ravary, John Pybus and Sergio Grau. 2012. *Audio BNC: The audio edition of the Spoken British National Corpus*. Phonetics Laboratory, University of Oxford. Available at <http://www.phon.ox.ac.uk/AudioBNC> (last accessed 20.10.2020).

Dehdari, Jon, Tim Weale, Eric Fosler-Lussier and DJ Hovermale. 2009. *SpeechSearcher manual*. Available at <https://buckeyecorpus.osu.edu/SpeechSearcherManual.pdf > (last accessed 20.10.2020).

Du Bois, John W., Stefan Schuetze-Coburn, Danae Paolino and Susanna Cumming. 1992. *Discourse transcription*. Santa Barbara: The University of California.

Evert, Stefan. 2009. *The CQP query language tutorial.* Available at <http://cwb.sourceforge.net/temp/CQPTutorial.pdf> (last accessed 16.9.2019).

Foulkes, Paul 1997. English [r]-sandhi – a sociolinguistic perspective. Histoire Episteìmologie Langage 19: 73–96.

Fuchs, Robert, Bertus van Rooy and Ulrike Gut. 2019. Corpus-based research on English in Africa: A practical introduction. In A. U. Esimaje, U. Gut and B. E. Antia (eds.). *Corpus linguistics and African Englishes*, 37–69. Amsterdam: Benjamins.

Godfrey, John J., Edward C. Holliman and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. San Francisco, CA, USA, 517–520. doi: 10.1109/ICASSP.1992.225858.

Greenbaum, Sidney (ed.). 1996. *Comparing English worldwide: The International Corpus of English*. Oxford: Clarendon.

Gut, Ulrike and Robert Fuchs. 2017. Exploring speaker fluency with phonologically annotated ICE corpora. *World Englishes* 36(3): 387–403.

Hannisdal, Bente Rebecca. 2006. Variability and change in Received Pronunciation. A study of six phonological variables in the speech of television newsreaders. PhD dissertation, Bergen: University of Bergen.

Hardie, Andrew. 2012. CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409. doi: 10.1075/ijcl.17.3.04har.

Hay Jennifer and Andrea Sudbury. 2005. How rhoticity became /r/-sandhi? *Language* 81: 799–823.

Hoffmann, Sebastian, Sabine Arndt-Lappe and Valérie Keppenne. 2018. *Exploring the potential and the limitations of the Audio Edition of the 1994 Spoken BNC*. Paper presented at the 39th ICAME conference in Tampere, Finland, 30 May–3 June 2018.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee and Ylva Berglund Prytz. 2008. *Corpus linguistics with* BNCweb *– a practical guide*. Frankfurt am Main: Peter Lang.

Kallen, Jeffrey L. and John M. Kirk. 2012. *SPICE-Ireland: A user's guide. Documentation to accompany the SPICE-Ireland Corpus: Systems of pragmatic annotation in ICE-Ireland*. Queen's University Belfast, Trinity College Dublin, and Cloì Ollscoil na Banrìona.

Lehmann, Hans-Martin, Peter Schneider and Sebastian Hoffmann. 2000. BNCweb. In J. Kirk (ed.). *Corpora galore: Analysis and techniques in describing English*, 259–266. Amsterdam: Rodopi.

Love, Robbie, Abbi Hawtin and Andrew Hardie. 2017. *The British National Corpus 2014: User manual and reference guide (version 1.0)*. Available at <http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf> (last accessed 20.8.2020).

McWhinney, Brian. 2020a. *Tools for analyzing talk. Part 1: The CHAT transcription format.* Available at <https://talkbank.org/manuals/CHAT.pdf> (last accessed 20.12.2020).

McWhinney, Brian. 2020b. *Tools for analyzing talk. Part 2: The CLAN program.* Available at <https://doi.org/10.21415/T5G10R> (last accessed 20.12.2020).

Mompeán, Jose A. and Pilar Mompeán-Guillamón. 2009. /r/-liaison in English: An empirical study. *Cognitive Linguistics* 20(4): 733–776.

Mompeán, Jose A. and F. Alberto Gómez. 2011. Hiatus resolution strategies in non-rhotic English: The case of /r/-liaison. In E. S. Lee and E. Zee (eds.). *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS),* 1414–1417.

Nelson, Gerald, Sean Wallis and Bas Aarts. 2002. *Exploring natural language: Working with the British component of the International Corpus of English.* Amsterdam: John Benjamins.

Pavlík, Radoslav. 2016. A usage-based account of /r/-liaison in Standard British English. *Journal of Phonetics* 54: 109–122.

Pavlík, Radoslav. 2011. A quantitative analysis of British linking and intrusive /r/ in newsreading style. *Studies in Foreign Language Education* 3: 119–134.

Pitt, Mark A., Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume and Eric Fosler-Lussier. 2007. *Buckeye Corpus of Conversational Speech (2nd release)* [www.buckeyecorpus.osu.edu]. Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Sinclair, John McH. 1996. *EAGLES. Preliminary recommendations on corpus typology.* Available at <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html> (last accessed 18.8.2020).

Soskuthy, Marton. 2013. Analogy in the emergence of intrusive-*r* in English. *English Language and Linguistics* 17(1): 55–84.

Steen Francis, Cristóbal Pagán Cánovas, Anders Hougaard, Jungseock Joo, Inés Olza, Anna Pleshakova, Soumya Ray, Peter Uhrig, Javier Valenzuela, Jacek Woźny and Mark Turner. 2018. Toward an infrastructure for data-driven multimodal communication research. *Linguistics Vanguard: A Multimodal Journal for the Language Sciences* 4(1). doi:10.1515/lingvan-2017-0041.

Svartvik, Jan and Randolph Quirk (eds.) 1980. *A corpus of English conversation* (Lund Studies in English 56). Lund: Liber/Gleerups.

Trudgill, Peter. 1974. *The social differentiation of English in Norwich.* Cambridge: Cambridge University Press.

Uhrig, Peter. 2018. NewsScape and the Distributed Little Red Hen Lab – a digital infrastructure for the large-scale analysis of TV broadcasts. In A-J. Zwierlein, J. Petzold, K. Boehm and M. Decker (eds.). *Anglistentag 2018 in Regensburg: Proceedings. Proceedings of the Conference of the German Association of University Teachers of English.* Trier: Wissenschaftlicher Verlag Trier, 99–114.

Winkelmann, Raphael and Georg Raess. 2014. Introducing a web application for labeling, visualizing speech and correcting derived speech signals. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis (eds.). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA). Available at https://pdfs.semanticscholar.org/2c77/a396c4cd007e8f6380efac4bb2c2c525f2fb.pdf> (last accessed 19.8.2020).

Wunder, Eva-Maria, Holger Voormann and Ulrike Gut. 2010. The ICE Nigeria corpus project: Creating an open, rich and accurate corpus. *ICAME Journal* 34: 78–88.