Zeitschrift: SPELL : Swiss papers in English language and literature
Herausgeber: Swiss Association of University Teachers of English
Band: 11 (1998)

Artikel: PDP : a recent shift in phonetic performance
Autor: Rudin, Ernst / Elmer, Willy
DOI: https://doi.org/10.5169/seals-99963

PDP: A Recent Shift in Phonetic Performance

Ernst Rudin and Willy Elmer

There is hardly any other branch of linguistics as thoroughly permeated by the notion of performance as dialectology: fieldwork is performance, from designing a questionnaire and finding suitable informants to successful interviews, and the stages of the description and organisation of dialect material are prototypical of what is frequently laborious performance. This relationship to the data never seems to leave the dialectologist and may be responsible for the fact that often in dialectology the description is the explanation – not an ideal state of affairs. We are not the first to suggest that dialect studies would profit from a reduction in the time-consuming manual assessment of the data, allowing dialectologists to spend more time and effort on interpretation. By making one of the core collections of English dialect data available for use on the computer, however, we are now in a position to add substance to this insight.

The use of computers in human and social sciences has been steadily increasing over the last decades. In addition to their use as general word-processing and research tools, they have been used for the compilation of dictionaries, concordances and bibliographies, and today we experience a boom in such manuals on CD ROM. In applied linguistics, the market of software for language teaching is rocketing, and linguistic investigation can rely on a variety of language corpora available on CD ROM. As far as English is concerned, there are three reference works of this type: the *Brown Corpus* (written American English), the *Lancaster-Oslo-Bergen Corpus* (written British English) and the *London-Lund Corpus* (spoken British English), which all have become important sources for statistical evaluations in the fields of morphology, syntax, lexicology and stylistics. The only field of linguistics that cannot take advantage of the material collected in these corpora is phonetics. Since all of them rely on the *lexeme* as the basic unity, they are of little use for the study of the *sound structure* of English, and up to now there is no comparable corpus based on phonetic material.

The *Survey of English Dialects: The Basic Material* (quoted in the following as *SED*) comprises the most substantial collection of phonetic data for the English language. A questionnaire consisting of 1326 questions administered to well over a thousand informants in 313 localities in England resulted in roughly 500,000 answers in detailed phonetic transcription, which were published in twelve volumes – three volumes each for the Northern Counties, the East Midland Counties, the West Midland Counties, and the Southern Counties. The basic structural unit of the *SED* is the *keyword,* which represents the Standard English response to each question asked. Each set of data is headed by a number and a keyword, followed by the question used to elicit it, the lexical variants given, and, finally, the actual responses in phonetic transcription, ordered by county and locality. The keyword *to WINNOW* in the West Midland Counties may serve as an example (our comments are marked by curly braces):

II.8.4 To WINNOW *{NUMBER AND KEYWORD}*

Q. What was their word for separating the grain from the husks? {QUESTION}

Rr. THRASHING, WIN, WINDOW~WINNOW, WITHER *{RESPONSES}*

7 Ch *{CHESHIRE}* 1 wɪðə [wɪðəɹɪnmɪʃe:n *withering-machine* (i.e. winnowing-m.)] 2 wɪnəɹɪn 3-5 wɪnə 6 wɪnɪ

8 Db *{DERBYSHIRE}* 1 wɪnɪ 2 wɪnə, °wɪndə[2] 3 wɪnə 4 wɪndə 5 wɪnə,°~[1] 6-7 wɪnᵊ

11 Sa *{SHROPSHIRE}* 1-2 wɪnə 3 wɪno· 4-8 wɪnə 9 wɪno· 10-11 wɪnə

12 St *{STAFFORDSHIRE}* 1[1] wɪnə 2 wɪnoɷ-ɪn 3 wɪnəɹ [+ V.] 4 wɪnəwɪŋg 5 wɪnəɹɪn [wɪnəɹɪnməʃi:n *winnowing-machine*] 6 wɪnəɹ [+ V.] 7 wɪnəɹɪn 8-9 wɪnə 10-11 wɪnoɷ-ɪn

(. . .)

Various linguistic atlases based on the *SED* have shown its relevance for dialectal geography: Kolb's *Phonological Atlas of the Northern Region* and *The Atlas of English Sounds*, Orton's *Linguistic Atlas of England* and Viereck's *Computer Developed Linguistic Atlas of England 1*, which concentrates on the lexical information contained in the *SED* but disregards its phonetic aspects.

Comprehensive analyses of the phonetic data contained in the *SED* are very difficult to accomplish by manual methods. To extract and statistically analyze phonetic data from among half a million entries is a very arduous,

not to say impossible, task. It is not surprising, therefore, that the collective maps in those Atlases that deal with phonetics tend to be based on a relatively small number of *SED* entries. So far, there is no statistical study that would take into account *all* the phonetic material. The sheer abundance of phonetic data interferes with any overall attempt at assessing it and leaves many of the hidden treasures of the *SED* still unrevealed.

The hidden treasures of the *SED* could be unearthed much more easily if the work were available in computerized form. In the field of dialectal geography this would make it possible to produce accumulative maps revealing general phenomena. The static representation of phonetic data could give way to the representation of phonetic *processes*. And in the area of general phonetics and phonology, accessing the *SED* by computer would for the first time allow the investigation of an entire series of context-dependent processes, such as rounding or unrounding, monophthongization or diphthongization, palatalization or velarization, in a statistically relevant way. Moreover, detailed quantitative analyses of the *SED* could back up qualitative analyses in the field of dialectology as well as in general linguistics. Finally, the immediate access to large-scale statistical information could not only help to confirm or to challenge the results of phonetic analyses so far, but might also lead to new insights and to the formulation of new questions.

Until recently, the phonetic material in the *SED* was not available in computerized form. Wolfgang Viereck, e.g., says in the introduction to his *SED*-based *Computer Developed Linguistic Atlas of England*, under the heading "Problems Encountered in Computerizing the Data:"

> The conventions show that phonetics is not dealt with. This means that a certain type of question cannot be asked, which will no doubt be deplored by some. The aims we had in mind naturally determined our computerization procedure. In view of the fact that a quantification of the data and a dictionary were envisaged, phonetic transcriptions had to be transformed into normal orthography in any case. (5)

In 1992 we started the *Phonetic Database Project (PDP)*, with the aim of making the *Basic Material (BM)* contained in the 12 volumes of the *SED* accessible in digitalized form, without sacrificing any of its phonetic information (Rudin and Elmer). The bulk of the actual scanning of the *BM* was done by Michael Gasser, while Ernst Rudin was responsible for encoding and structuring the data, for routines that checked their accuracy and for the search procedures. In the following, we will briefly delineate three aspects of our project:

1. Scanning and encoding phonetic script
2. Search procedures
3. Map-making

1. Scanning and Encoding Phonetic Script

Our workplace was equipped with a 486/25DX computer with 8 Megabytes of RAM, a 200 megabyte hard disk – an incredibly huge storage capacity five years ago – a *Syquest* removable disk, a *Panasonic* scanner with an optical resolution of up to 400 dpi, and the *proLector* software for optical character recognition (OCR). Until a few years ago, the main obstacle to electronic reading of phonetic transcription was that none of the OCR programs that fit into the average budget of a University department was able to perform such a task. *ProLector* has changed this situation. On the one hand, it is highly accurate and able to distinguish between the minute graphic differences that can occur in phonetic transcription – the difference, say, between the same phonetic sign accompanied by different diacritics. On the other hand, it is trainable and allows users to define any scanned sign by a string of up to four characters or numbers – to code phonetic characters, in other words.

When we started our project, none of the different versions of the ASCII-code – the basic and internationally standardized computer character sets – included the complete phonetic alphabet. Word processors that could display phonetic symbols on screen and in print used a system of *encoding* those symbols that are not within the ASCII set. Today, computing still relies to a great deal on ASCII. *Unicode,* a new international character set with a potential of 65,000 symbols, has only just started to replace the ASCII code for Windows and Internet applications. Since we wanted to keep the scanned data as open and as convertible as possible, we decided against using a code that would tie us down to a specific word-processing software. As a first requirement, our system of encoding had to fit into the 128 characters of the 7 bit ASCII set. Discounting the 32 control characters and the space character, we were left with 95 symbols, including the small and capital letters of the roman alphabet and numbers. We defined three additional requirements for our code. It should be:

a. as long as necessary and as short as possible;
b. as specific as necessary; tailored to our task, i.e. able to reproduce all the phonetic details contained in the *SED*, although not necessarily those of other collections of phonetic data;
c. readable to the human eye at least to some extent. In order to facilitate error searches and the editing of the scanned material, the code should not be entirely arbitrary but make certain categories apparent.

The IPA (International Phonetic Association) 1989 Kiel Convention Workgroup agreed on a coding system for phonetic symbols in which "each accepted symbol or diacritic should be assigned a unique numerical equivalent" (81-82). The IPA code represents each phonetic character by a three-digit number, the first digit of which indicates the category of the symbol. Thus, IPA code units starting with the numerals 1 or 2 represent consonants, those starting with 3 are vowels, while the numbers in the 400s and 500s are reserved for diacritics and suprasegmentals, respectively. This code seemed to fit our requirements; the fact that it covers not only the symbols contained in the *SED* but the entire IPA alphabet was no disadvantage, and we had every interest in using an internationally standardized code. By the same token, the IPA computer coding system imposes some restrictions on computer searches and does not fulfill all our requirements:

- It assigns a three digit number to "each accepted symbol *or* diacritic" (emphasis added). This means, that a given sound is not always represented by three digits. Depending on the number of diacritics that accompany the basic symbol, six, nine, or even twelve digits are needed for codification. Thus [a] is 304, [ä] is 304493, [a̤] is 304415497, and [a̤:] is 304415497503. Search procedures can only recognize the code-length of a given sound by going beyond the basic unit in order to detect whether the next unit is an additional diacritic or a new basic symbol.
- The three-digit IPA units cannot simply be stringed together. If they were, the resulting data string would hardly be readable, and therefore go against requirement 3 above. Moreover, search errors would invariably occur. The coded form of [ɒs], 313132, to give an example, includes the string 131, which in its turn is the code for [Eth]. In order to avoid such errors, the three-digit units would have to be separated by delimiters, e.g., @313@132, or the search program would have to keep track constantly of its exact position within the data string.

While none of these restrictions make it impossible to work with the IPA code, we nevertheless decided on a code that is more concise, more suitable to our requirements, and allows more elegant programming. We now represent each sound by a four-digit unit, regardless of whether it carries diacritics or not. The four digits are classified according to their function:

a. The first digit serves as a delimiter and indicates at the same time the basic category of the sound concerned. It has one of the four following values: % for vowels, & for superscript vowels, # for consonants, and $ for superscript consonants.
b. Every basic consonant or vowel symbol is represented by a two-digit cardinal number that occupies positions 2 and 3 in the unit. In the case of vowels, the first of these two digits indicates the vowel *type* – the codes for [a], [ɐ], [ɑ], [ɒ], and [ʌ], for example, all begin with 0. This allows searches not only for specific vowels, but also for vowel categories.
c. The last digit consists of a numeral or letter which represents a diacritic or a combination of diacritics. With symbols that do not carry any diacritics, the value of this digit is 0. In our encoding system, to use examples already cited, [a] is %000, [ä] is %001, [a̤] is %00F, and [ä̤] is %00F:.

2. Search Procedures

Since the basic unit for dialectologists interested in phonetic patterns is not the word but the sound, isolated or in combinations, we could not simply take advantage of an existing text-retrieval or hyper-text program. Had we been interested mainly in lexical searches, we could have used any of the commercially available retrieval programs, which offer a great variety of search options and speed up searches by indexing the data. For our purposes, however, an alphabetic index would only be useful for searches that are restricted to the very beginning of entries: with most of the searches we do, the sound or phonetic pattern that we are looking for can occur anywhere in an entry, not only at the beginning. We therefore decided to write a search routine of our own. We would like to demonstrate, using a basic search procedure, how this program, named *QScan* and written in Modula-2, works. We will not hunt for hidden treasures, but will illustrate a fairly well-known phenomenon (see Elmer and Rudin for a more detailed example that also involves a lexical search). The search string we propose is *#710* and corresponds to [ɹ], one of the realizations of the phoneme /r/ in English. Since the

four-digit code unit is framed by asterisks, the search will include all the occurrences of [ɪ]. Omitting the first or the second asterisk would limit the search to words that start with [ɪ] and that end on [ɪ], respectively. For reasons of time, we will not perform a search on the entire data, but will restrict ourselves to a 15% sample. Searching the complete set of data would take about a minute with an average Pentium Computer – a significant time gain in comparison to the weeks and months you would have to invest to perform such a search manually.

We start our search by entering *QSCAN *#710** at the DOS prompt. While it is running, the program keeps us informed about the number of keywords it has checked. Once the search is finished, the following message appears on the screen:

PHONETIC DATABASE PROJECT UNIVERSITY OF BASEL 1997
The English Seminar

Searchstring *#710* found 11109 times in 762 keywords

15.11.1997 RS.Q2 lists all the FINDINGS
12.05 ST.Q2 gives a STATISTICAL SURVEY
Type "QMAP" to create maps based on search results

QScan produces a list of all the items that correspond to the search criteria, indicating the county and the locality of each entry:

LIST OF FINDINGS. Original Filename: RS.Q2. For STATISTICS, see file ST.Q2

Keyword Pattern: * SEARCHSTRING: *#710*

N11-2.EXT; I. 1 .2 FARMSTEAD
La 3: #420%043$710#710:#200#460#030%130#040
La 4: #420%000$710:#710#200#460#030%220%110#040
La 8: #420%001:#710#200#460#030%220%110#040
La 9: #420%000$710:#710#200#460#030%130#040
La 11: #420%150$710:#710#200
Y 5: #420%000:#710#200
Y 6: #420%000#710#200
N11-3.EXT; I . 1 . 3 FARMYARD
La 8: #420%001$710:#710#200#820%001$710:#710#040
La 9: #420%000$710:#710#200#820%000$710:#710#040
(. . .)

While we consider our code in ASCII format to be an ideal basis for search procedures, it is not a very practical basis for dialectologists to work from. PDP therefore includes a routine that reconverts the search results into phonetic script so that you get them on screen and in print as they appear in the *SED*. For the time being, we have such reconversion algorithms for *Nota-bene Lingua* and for *Microsoft Word*. The result of this conversion looks like the following:

LIST OF FINDINGS. Original Filename: RS.Q2. For STATISTICS, see file ST.Q2

Keyword Pattern: * SEARCHSTRING: *□*

N11-2; I. 1 .2 FARMSTEAD
La 3: fæɹːmstɛd
La 4: faˈːɹmstɹəd
La 8: fäːɹmstɹəd
La 9: faˈːɹmstɛd
La 11: fɜˈːɹm
Y 5: faːɹm
Y 6: faɹm
N11-3; I . 1 . 3 FARMYARD
La 8: fäˈːmjäˈːd
La 9: faˈːmjaˈːd
(...)
N11-5; I.1.5 PIGSTY
Cu 1: pɪgkɹoː,
Du 4: pɪgkɹi:
Du 5: pɪgkɹˈi:
La 1: hɷŋəɹɷlˈ IM
La 5: pɪgkɹöu
La 6: pɪgkɹu:
La 6: [bɷlkɹuːˈ IM REP
La 6: kɔːfkɹuːˈ IM REP
La 7: pɪgkɹu: BC [usual]
Y 25: [pɪgəɹɪˈ IM
Y 27: pɪgəɹɪ BC ["polite"]
(...)

As you can see under the last main entry, *QScan* not only lists the relevant items, but also gives some additional information about them. The abbreviation *IM* stands for incidental "material," responses, i.e., that are not spontaneous reactions to a given question, but mentioned as a second possibility, produced after the fieldworker pressed for other responses, or that occur later during the interview. *REP* indicates the "repetition" of the search

string within the same locality, *BC* signals "brief comments" by the informant. In a second file, *QScan* gives a basic statistical survey of the search results, again indicating repetitions and incidental material:

STATISTICAL SURVEY. Original filename ST.Q2
RS.Q2 lists the FINDINGS

Keyword Pattern: * SEARCHSTRING: *#710*
Keywords scanned: 762

11109 FINDINGS OVERALL, of which:
1346 REPETITIONS within locality <R>;
886 occurrences in INCIDENTAL MATERIAL <I>, of which
420 REPETITIONS <(R)>
9297 BASIC MATERIAL findings without Incidental Material and Repetitions <B>

NUMBER OF FINDINGS WITHIN ONE LOCALITY Highest: 152
Lowest: 0
Average: 35

SCOREBOARD
North (75 loc.) : 188 keywords checked; 2877 findings.
East (87 loc.) : 192 keywords checked; 3924 findings.
West (76 loc.) : 190 keywords checked; 3130 findings.
South (75 loc.) : 192 keywords checked; 1178 findings.

NORTH: 2877 FINDINGS

Northumberland: 42

Nb	5:	1	1B		
Nb	6:	2	2B		
Nb	7:	14	12B	1R	1I
Nb	9:	25	20B		5I

Cumberland: 184

Cu	1:	13	12B	1I	
Cu	2:	35	31B	3R	2I (1)
Cu	3:	25	23B	1R	2I (1)
Cu	4:	35	30B	3R	3I (1)
Cu	5:	47	39B	6R	4I (2)
Cu	6:	29	27B	1R	1I

Durham: 158

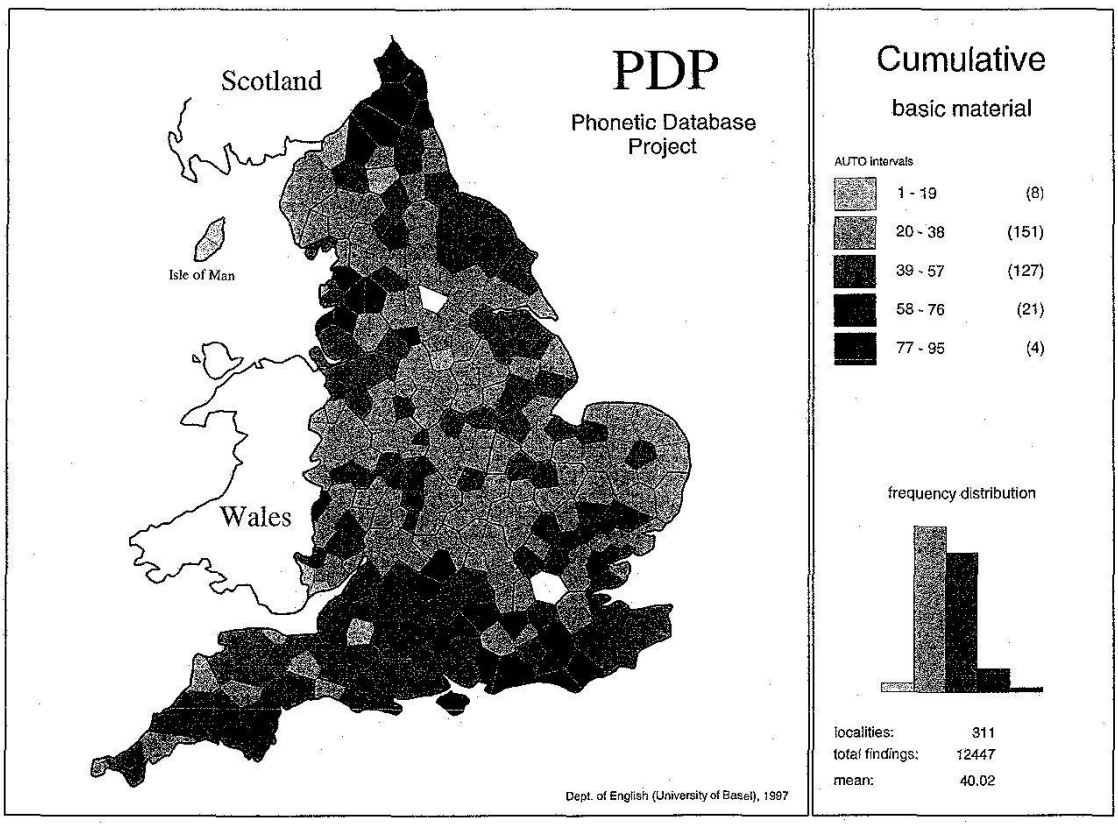
Du	3:	19	19B		
Du	4:	45	38B	5R	3I (1)

Du 5: 43 40B 3R
Du 6: 51 44B 7R 2I (2)
(. . .)

The regional statistics show a low representation of [ɪ] in the South (the East, with a comparable number of localities and keywords checked, has about three times as many findings as the South). The regional numbers for the North are slightly, but not significantly, lower than those for the East and West. More striking, as far as the North is concerned, is the absence of some localities in the ensuing list of localities: [ɪ] does not occur at all in Northumberland 1,2,3,4, and 8, and the same is true for Durham 1 and 2.

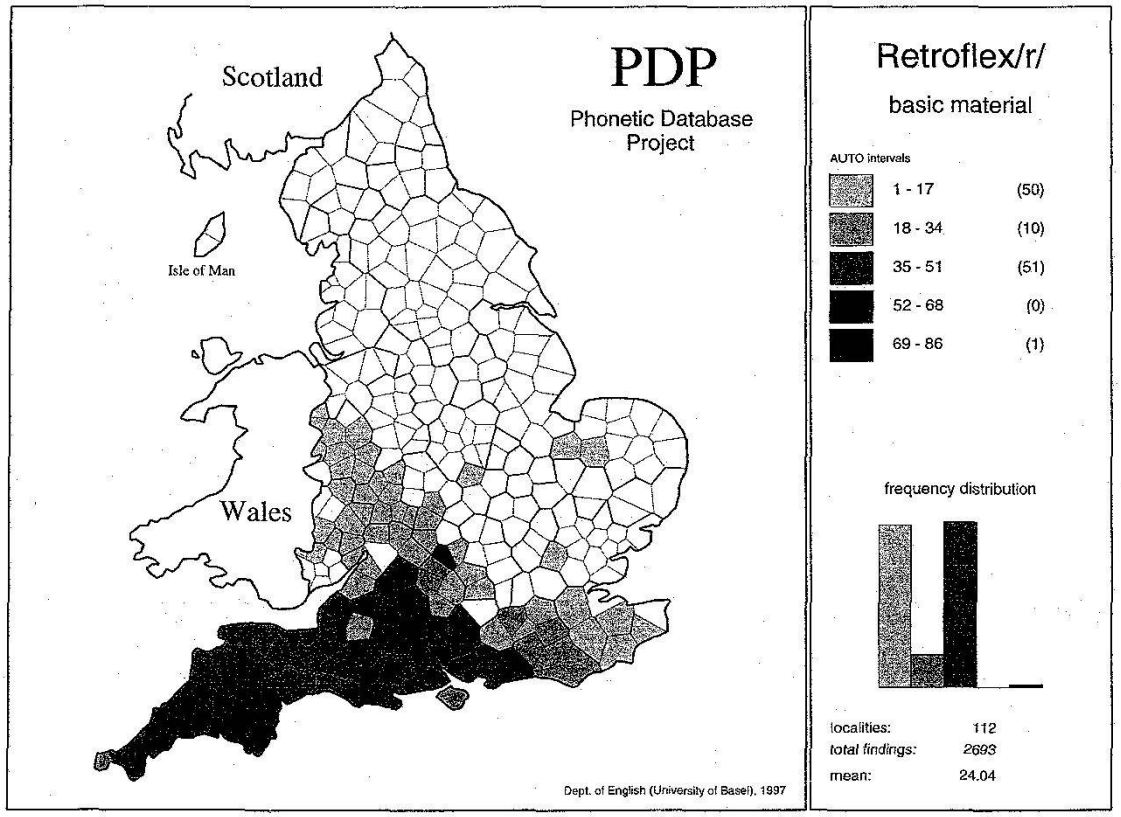
3. Map-making

QScan produces a third file, which feeds into a cartography program written by Guy Schiltz. Search results can thus be immediately visualized on screen or in print:

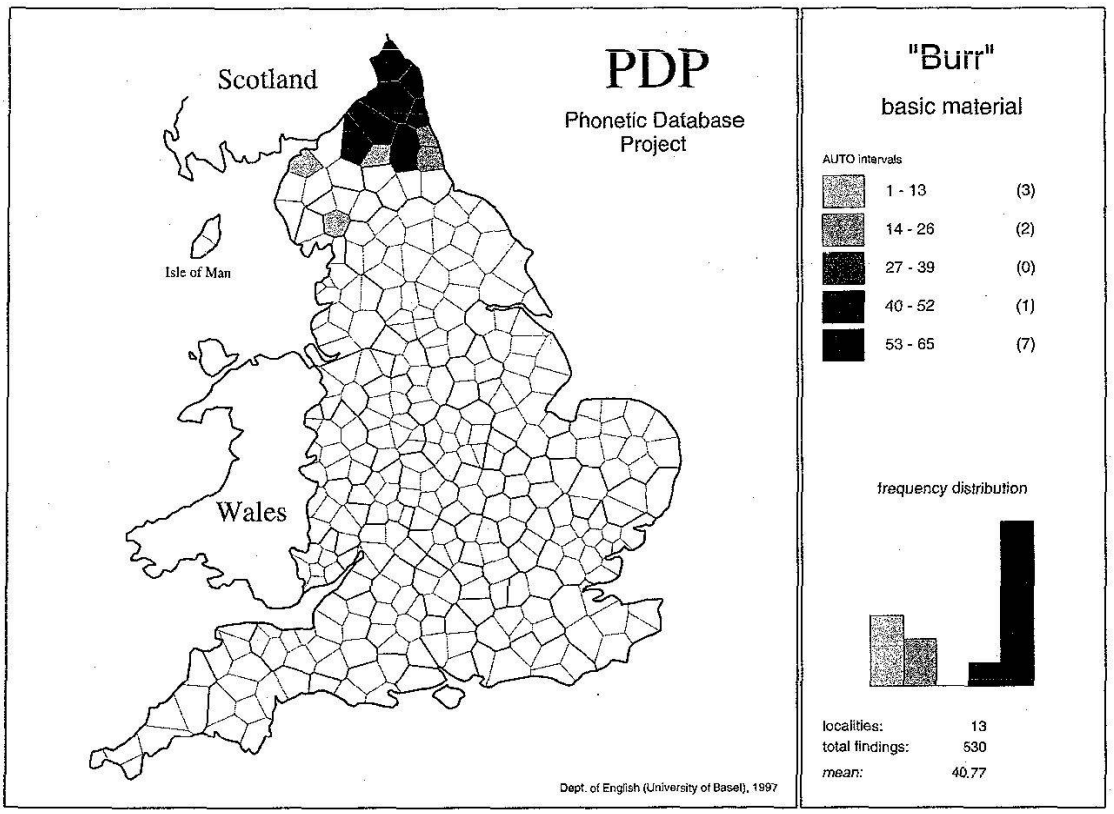


MAP 1

Map 1 illustrates the numbers of the statistics file graphically: in a large part of the South and in the most Northern localities, [ɪ] does not occur. We may know that the "burr" – [ʁ] – is one of the /r/ variants that occur in the far Northeast of England and that the same goes for retroflex /r/ – [ɽ] – in the Southwest. Let us therefore check the distribution of these two allophones. This time, we will not consult the list of the findings or the statistical survey, but go directly to the maps:

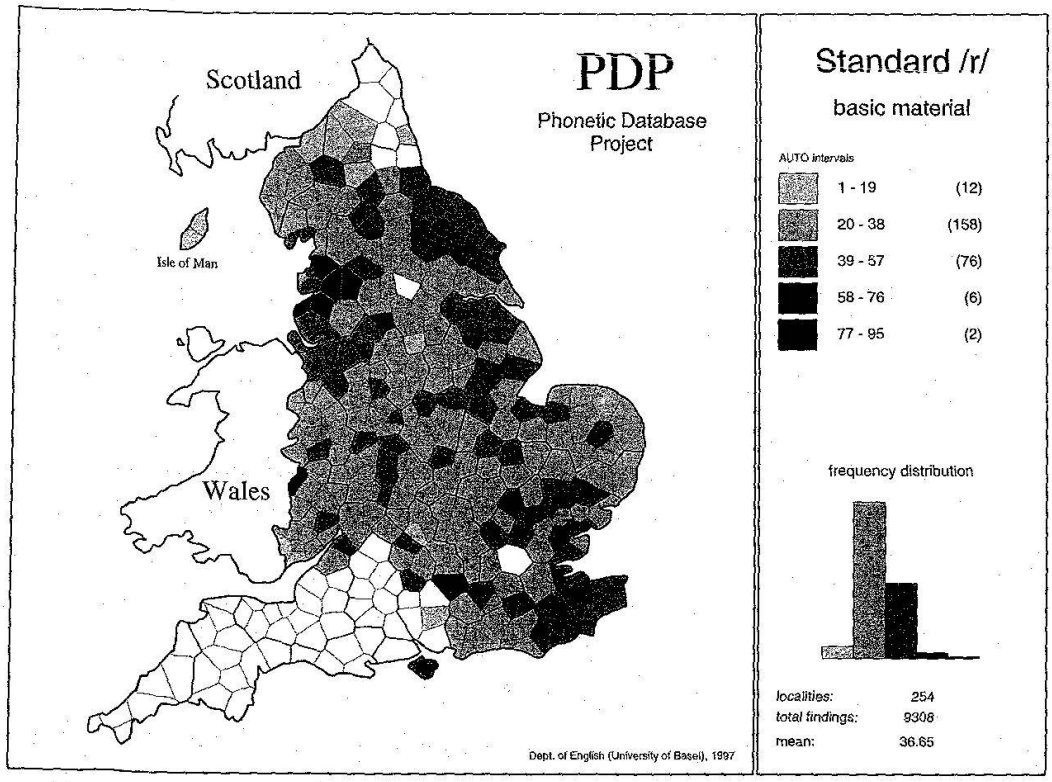


MAP 2



MAP 3

Maps 2 and 3 nicely complement Map 1. Since *QScan* allows up to three search strings, we can double-check this result by doing a simultaneous search for [ɪ], [ʁ], and [ɾ].



MAP 4

As was to be expected, this produces a fairly homogeneous picture. With more time at our disposal, we could now check further variants of /r/ – [r] or [ɾ], for example. Let us just add that the two blank spots on the map, corresponding to Leeds and Hackney, are not due to any particular pronunciation of /r/. The 15% sample that we used for our searches belongs entirely to the field of farming. While in Leeds questions on farming were not asked at all, in Hackney they were not "remunerative" (*SED*, 3.I 14).

PDP allows quick and accurate access to the 500,000 phonetic entries of the *SED* and thus brings along a shift in performance: dialectologists interested in the phonetic characteristics of the *Basic Material* no longer have to

go through a tedious process of searching the data manually for phonetic evidence, but can concentrate on interpreting the data.

Works Cited

Elmer, Willy and Ernst Rudin. "The *Survey of English Dialects* as an Electronic Database for Research in Areal and Variationist Linguistics." *Issues and Methods in Dialectology.* Ed. Alan R. Thomas. Bangor: University of Wales, 1997. 234-246.

"The IPA 1989 Kiel Convention Workgroup 9 Report: Computer Coding of IPA Symbols and Computer Representation of Individual Languages. Declaration." *Journal of the International Phonetic Association* 19.2 (1989): 81-82.

Kolb, Eduard. *The Phonological Atlas of the Northern Region.* Bern: Francke, 1966.

——— et al. *Atlas of English Sounds.* Bern: Francke, 1979.

Orton, Harold et al. *The Linguistic Atlas of England.* London: Croom Helm, 1978.

Rudin, Ernst and Willy Elmer. "¿Qué hace un ordenador con transcripción fonética detallada? La conversión del *Survey of English Dialects* en un banco de datos." *Jornadas internacionales de lingüística aplicada. Actas 2* (1993): 666-673. Instituto de Ciencias de la Educación. Granada: Universidad de Granada.

Survey of English Dialects: The Basic Material. 12 vols. Leeds: E.J. Arnold, 1962-1971.

Viereck, Wolfgang. *The Computer Developed Linguistic Atlas of England 1.* Tübingen: Niemeyer, 1991.