The Computer Developed Linguistic Atlas of England, volumes 1 (1991) and 2 (1997): Dialectological, computational and interpretative aspects

Wolfgang Viereck
University of Bamberg, Germany

1 Description of the project and personal reminiscences

This project – the first computer developed linguistic atlas of England – has been a rather long time in the making. Its database is that provided by the *Survey of English Dialects* (SED), a nationwide survey conducted in England in the 1950s and 1960s and published in narrow phonetic transcriptions in four regional volumes, each in three parts (cf Orton et al 1962–1971; for a description of the survey see Viereck 1988). Our intentions had been overly ambitious in the beginning as we wanted to include phonetics, too. At that time the scanner had not yet been invented with which even today the error rate is quite high when narrow phonetic transcriptions and diacritics are involved. In those days it would have been necessary to devise a key for coding even minute phonetic differences. While this would surely have been possible, the task of using the phonetic key correctly and of putting the enormous amount of information on endless sheets of paper to be then punched on cards was too formidable to be attempted. The computer stone age is pretty recent; progress in computational dialectology has been unbelievably fast.

As a quantification of the data and a dictionary had originally also been envisaged, phonetic transcriptions had to be transformed into normal orthography, too[1]. This transformation of pronunciations into spellings proved very difficult indeed[2]. In addition, even without phonetics, coding is a rather monotonous business and one has to reckon with the human factor. In the course of our work I was often reminded of Roger Shuy's remark that he had been given a handsome sum of money for coding syntactic data. One should think that this is not too difficult a task, but when the work had been completed, Shuy had to throw everything

away as the results proved too unreliable! In addition to the monotonous coding procedure (cf Viereck 1991: 3–5 for the key according to which the data were coded) there were problems of a different kind. Orton had different co–editors working on the volumes of the Northern Counties, the West Midland Counties, the East Midland Counties and East Anglia and the Southern Counties and they often solved identical problems differently. The inconsistencies within these volumes and the discrepancies among them are many. How can one expect students instructed to code the material to come up with a uniform picture under such circumstances? But even a (Munich-trained) dialectologist did a bad job and almost ruined the first volume of our atlas as she did not decode identical transcriptions consistently. I noticed this late, but not so late that the errors could not be remedied, but it meant a not inconsiderable delay in publishing the volume and also a financial loss. While this problem occurred at Bamberg, Marburg, too, with whose 'Forschungsinstitut für deutsche Sprache – Deutscher Sprachatlas' we cooperated, was not immune to unpleasant surprises. After all the maps of volume 2 had been corrected and filmed, we received the locality map and great was my surprise to note that five localities had 'drifted' even into adjacent counties! I never thought that something like this could happen. It apparently occurred, I was informed, when the locality coordinates were transferred from the mainframe computer where volume 1 had been produced onto the PC. All the finished maps had to be thrown away and had to be done again. The financial loss was considerable but the loss of time was even worse. It meant that the dialectometrical part could not be finished in the estimated time. Although the Marburg collaborator, financed by the German Research Council, promised to finish it, he never did but simply disappeared. One year later he wrote me, demanding a testimonial – which he had already prepared for me just to sign. On top of that, a Marburg colleague who had not done a stroke for volume 2 requested to be put on its title page for the 'simple' reason that he is the boss of a real collaborator with 50 per cent of the latter's working time. Such are the joys of real teamwork! In view of such problems it is sometimes surprising that such long-lasting projects are finished at all in these financially difficult times.

Whereas the SED data published in CLAE 1 and 2 have been checked and rechecked several times, this is not the case with the other SED data. Thus our computer tape with the complete SED data can only be distributed among researchers after careful checking.

2 Dialectological aspects

CLAE 1 and 2 demonstrate the potentials of the computer in linguistic cartography very well. Their most important special features, which cannot be found in previous linguistic atlases based on the same data and produced in England (see Orton and Wright 1975; Orton, Sanderson and Widdowson 1978), are the following:

– The use of symbols. They allow the indication of linguistically important transition zones, whereas isoglosses suggest boundaries where in reality there are none. Also, isoglosses can be drawn, with equal justification, through different territory.
– The maps are documentary in character. They show what notions occur in the whole of England and document precisely in the legends what words, forms etc are put together and, if subsumed, what is subsumed under one and the same symbol and what notions remained unmapped and where they, too, were recorded.
– Of special importance on the maps is the integration of informants' and fieldworkers' remarks into the symbolisation. The system even allows information of more than one category to be indicated in any one symbol.

The map L 20 *Anvil* of CLAE 1 illustrates these points well. As the legend shows, we distinguish between seven categories of status indications. With the exception of 'preferred', their interpretation is straightforward. The reasons for preferring one word over another may vary. An informant's judgement may be based on the standard norm or on the dialectal norm. On the map L 20, the informants quite consistently preferred the dialect word *stiddy*, although, with equal consistency, the informants were quite aware that this was the older, less commonly used word, which was repeatedly elicited only under pressure or when suggested. Such judgements are indicative of linguistic change; they are a part of dynamic dialectology. Beyond any doubt, the Scandinavian loan *stiddy* will be ousted completely by Standard English *anvil*, and this will happen more quickly north of the Humber than in Lincolnshire, which seems in many ways to be a more stable relic area. By providing status indications, also varying degrees of competition between words can be distinguished, ie those that will be ousted earlier than others.

The map in question also provides insights into the basic principles of symbolisation. The greater the number of identical symbols, the larger the area covered by these symbols, the simpler are the strictly geometrical symbols. The rarer the answer, the more 'complex' or 'unusual' is the

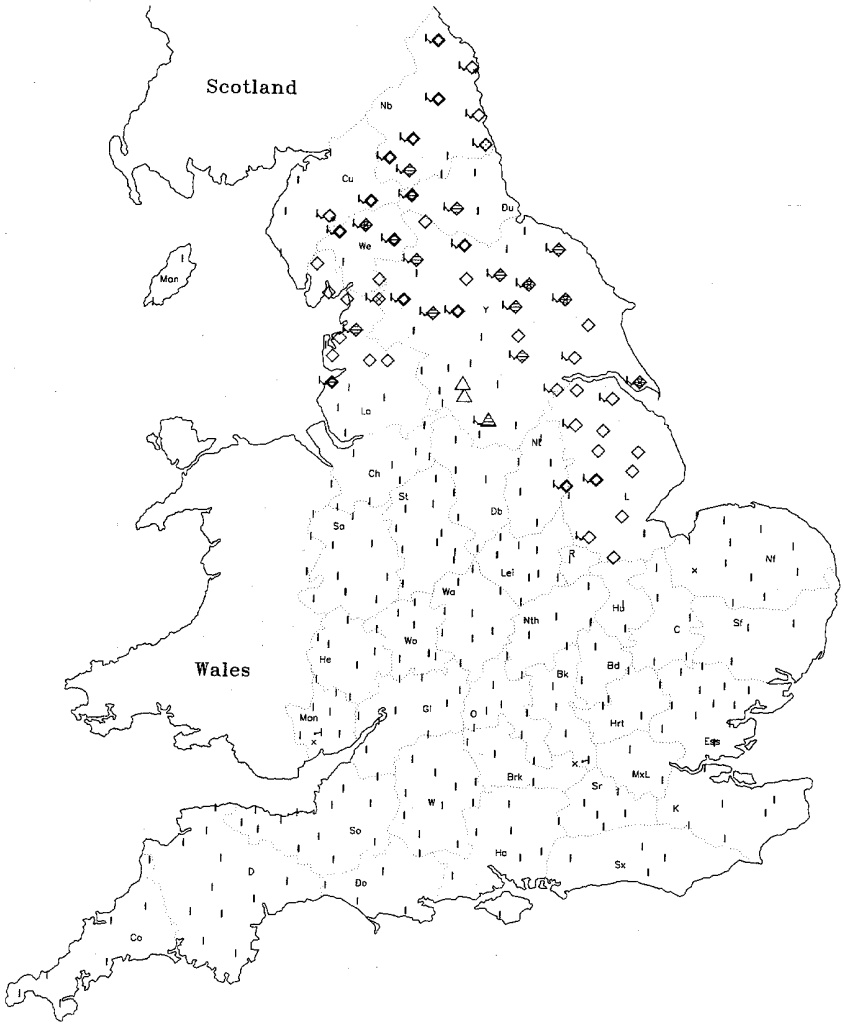

Figure 1: L 20: VIII.4.10 Anvil

By the way, we haven't mentioned the blacksmith, but what does he hammer things on?

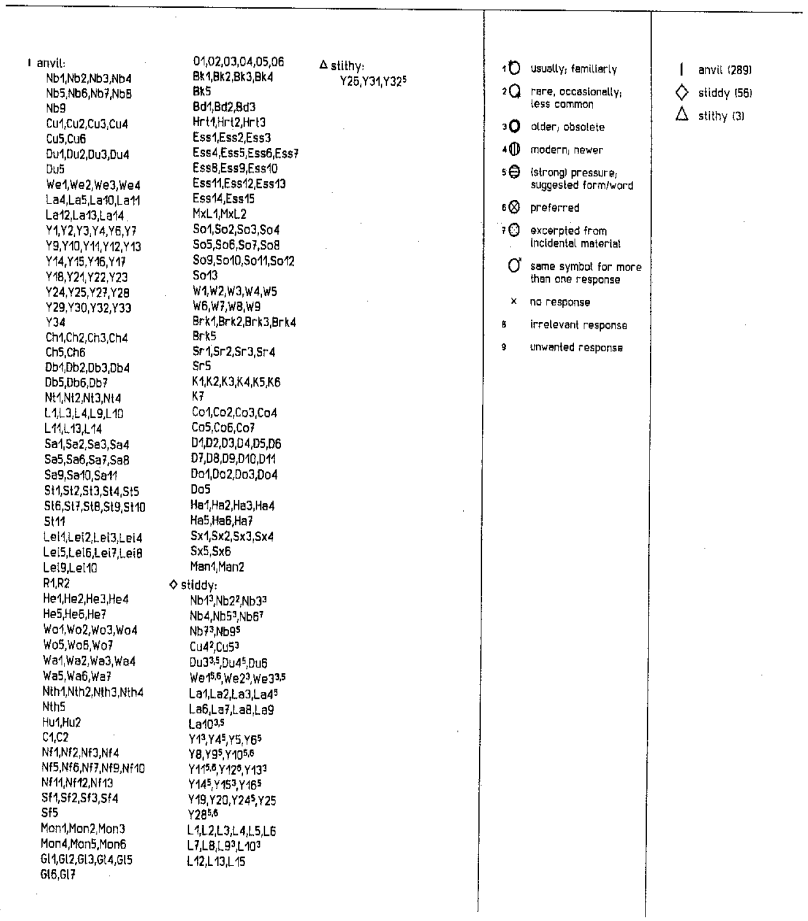

Figure 2: L 20: VIII.4.10 Anvil

symbol. Generally, the symbols of non-standard forms are larger and thus catch the eye more directly.

CLAE 1 contains 169 item-centered maps of the type reproduced here. Of these, 75 are lexical, 56 morphological and 38 syntactic in nature. CLAE 2, which is due for publication in the second half of 1997, adds 152 maps of the same type, of which 75 are lexical, 65 morphological and 12 syntactic in nature. From these numbers follows that, in both volumes, special attention was given to morphosyntax and I can honestly say that every rewarding item in this area was mapped. As the contribution CLAE 1 and 2 make to dialectal grammar is not only to be gathered from the two Tables of Contents, which can only give an incomplete picture, we have also added a comprehensive subject index covering all maps in both volumes, as grammatical evidence may be hidden in lexical maps, too.

3 Computational aspects

As regards the computational aspects of the project, I summarise Harald Händler's most important points that he puts forward in his introduction to CLAE 2. Since the publication of CLAE 1, substantial changes have occurred within the computer scene. The crucial development concerning linguistic geography is the fact that the graphical language *PostScript* has asserted itself worldwide as a standard language. Using this language it is possible to produce all kinds of graphics – fonts, drawings, diagrams, scanned images and so on – in a standardised manner. Without changing anything of the *PostScript* code the user can benefit from the highest quality output any given device is capable of achieving. This makes it possible to prepare, for instance, the linguistic map that is finally printed by using the desktop laser printer. In order to make use of this decisive progress, the whole program system that we used to produce the first volume of CLAE was replaced by completely new software, adapted to current requirements. In the process of this, the somewhat antiquated programming language FORTRAN was replaced by the more modern C++, thus enabling an easy application of the programs on nearly all computers.

It was not an easy decision to part from such an extensive program system capable of producing a linguistic atlas. However, the new possibilities were of such promise, and the amount of work to adjust the old software seemed acceptably so high that the risk was taken. To give an example of the problems involved in adapting the old software: In

producing CLAE 1 we used plotters; the graphical output programs controlled the drawing pen that drew characters, symbols, lines etc onto the foil in a traditional manner. With *PostScript* the programs have to formulate the whole linguistic map in the form of a text, a code written in a special language that is understood and worked out by the output devices. Anyone who is familiar with the traps of such a complex programming language like *PostScript* has an idea of the difficulties resulting from the ambitious project of building up an extensive program system from the very beginning.

Nevertheless, the advantages of *PostScript* were decisive. A seemingly trivial graphical element, such as the width of a stroke, demonstrates this impressively. A plotter with its drawing pen is hardly able to vary line width; wider lines are produced by drawing the standard line repeatedly. Using *PostScript*, this element can be set at will and can be changed constantly. The special features of a printed character, for instance, are often serifs: the stroke width has to be reduced continuously. For the plotter this is an unsolvable problem. *PostScript*, however, is equipped with hundreds of character fonts – phonetic, Latin, Cyrillic, Hebrew etc – all of them to be scaled optionally, with or without serifs. They are also easily accessible and reasonably priced.

The conceptual arrangement of *PostScript* partly exercises an influence on the programming work itself. The problem of multiple responses demonstrates this fact impressively. If there is a locality on the map where several symbols have to be drawn, they are arranged around the point, all of them marked with a tiny line from the symbol edge to the centre of the locality. In order to solve this problem using an ink pen plotter, the programmer has a great deal of work to do: he must know the mathematical function of the symbol shape, he has to intersect this function with the line from the locality point to the centre of the symbol and induce the ink pen to draw a stroke from the intersection point to the locality centre – not an easy task when operating with complicated symbols such as the ‘raspberry’. Using *PostScript*, this is much easier. The line (ie the code that will produce the line) is drawn from the symbol centre to the locality point. Then the symbol is drawn and filled with white colour. The disturbing stroke inside the symbol is still there, but it is no longer visible, since one of the basic principles of *PostScript* is *opaqueness*: if several graphical elements are to be drawn, ie coded, at a point, only the last one is visible. Every element is opaque, no matter whether it is black, grey or white. Already these few features of *PostScript* – the free choice of stroke width, filling of a closed line,

colour option and the principle of opaqueness – offer a whole range of new dimensions regarding the production of computer graphics.

In spite of all the substantial improvements with regard to the software, there were only minor changes concerning the dialectological part of the work. The conception of the *rough linguistic map* described in the introduction to CLAE 1 has again proved successful.

The main difference to the production of volume 1 is the avoidance of output devices of the previous generation. The entire production of maps during the early phases of the work was carried out by means of a laser printer. The tiny strokes, which are possible with *PostScript*, made even complicated maps scaled down to DIN-A4 format clearly legible. The speed of this way of production is impressive: linguistic maps that formerly required hours are now completed in minutes. The final production revealed the whole range of the *PostScript* conception: only slightly enlarged, the working maps were produced on a laser image setter. The films produced in this way were used directly for printing, thus giving rise to several advantages: high quality, high speed and low costs, as film exposures are much cheaper than the special foils and ink pens necessary for map production in a traditional way.

4 Interpretative aspects

Scholars associated with the SED have been reluctant to make the necessary generalisations and to define dialect areas. In 1983, Raven McDavid remarked:

> The English interpretive works are ... disappointing. The 1975 *Word Geography of England* [by Orton – Wright] is not a word geography; for it nowhere summarizes, in statement or maps, the characteristic vocabulary of any region in England. One may similarly judge the 1978 *Linguistic Atlas of England* [by Orton – Sanderson – Widdowson] and Eduard Kolb's somewhat redundant 1979 *Atlas of English Sounds*. In none of these works is the cartography comparable to Kurath's; their charts treat too many variants with too many symbols. The delineation of English dialect areas from Orton's *Basic Materials* (SED) is yet to come, from someone who will look at patterns rather than items (49).

Wakelin has interpreted some SED data, but almost exclusively in the area of phonetics/phonology (cf Wakelin 1977, 1983). He remarks: 'In

the phonology, I have attempted to define dialect areas, but this is not possible for the morphological and lexical data' (Wakelin 1984: 70). This assessment is clearly wrong, as shown by a number of insightful publications. Until 1986 no nationwide picture existed. In that year, I presented such an overview based on SED lexical data (Viereck 1986a) to be followed in the same year by using SED phonetic and grammatical evidence (Viereck 1986b). Their methodology was traditional. Then followed some dialectometric analyses with CLAE 1 lexical and morphosyntactic data (Viereck 1992, 1995a, 1995b, 1996a, 1996b). Although these analyses corroborated to a great extent the findings of my traditional analyses of the structure of English dialects, they also yielded insights that clearly went beyond an analysis on traditional lines.

I cannot survey here the various quantitative methods available[3]. They fall into several categories, namely arithmetic methods among which the isogloss method, the identity method and the gravity centre method figure prominently and multivariate or multidimensional analyses. In contrast to the former approaches, it is the discrepancy between the geographical and the linguistic map that is of importance with the latter procedures. A number of these methods are used in treating CLAE 1 and CLAE 2 data.

For CLAE 2, the following seven dialectometric contributions were especially written. Sheila Embleton of York University, Canada, together with Eric Wheeler, Ontario, Canada, report on the methods and results of a study using multidimensional scaling on the CLAE data. Chitsuko Fukushima of Niigata Women's College, Japan, investigates morphological standardization of English English based on the data of both CLAE volumes. The Romance scholar Hans Goebl of Salzburg University, Austria, is especially concerned with dendrographic classifications of the CLAE data. In a second contribution, he and Guillaume Schiltz deal with quantitatively important boundaries as well as with what they call dialect integration. Harald Händler of the Philipps University of Marburg, Germany, and I describe the findings obtained with a specially developed gravity centre method. Fumio Inoue of the University of Foreign Studies, Tokyo, Japan, and Chitsuko Fukushima apply 'Hayashi's quantificational theory type 3' to the CLAE data, a multivariate analysis successfully used with Japanese dialects. In the final paper Alan R. Thomas of the University of Wales in Bangor subjects the lexical data of CLAE 1 to 'a two stage analysis, comprising a "lateral" clustering procedure followed by hierarchical cluster analysis of the lateral clusters identified'.

Preference is here given to analysing the data actually elicited rather

than making conjectures on 'the probability that a target linguistic feature might have been elicited [at a location] at the time of the survey' (Kretzschmar 1996a: 28), as do Linn and Regal (1993) as well as Kretzschmar (1996a, 1996b) with their (density) estimation and probability mapping.

Notes

1 The plan of producing a dictionary has since been abandoned although even after the publication of Upton, Parry and Widdowson (1994) there would still be plenty of room for a dialect dictionary based on the data of the SED (cf my review Viereck 1997). A pronouncing dictionary of dialectal English is also a desideratum, now much more feasible, but difficult to accomplish even with modern technology.

2 On the problems involved cf Viereck (1988, 269–271, 277). Quite a few dialectologically important distinctions were either overlooked or made wrongly in Orton and Wright (1975) and Orton, Sanderson and Widdowson (1978). On these, see Viereck in the introductions to both CLAE 1 and 2.

3 On a number of these, with examples also from our project, see Inoue (1996a, 1996b).

References

Inoue, Fumio. 1996a. Computational dialectology (1), *Area and Culture Studies* [Tokyo] 52: 67–102.

Inoue, Fumio. 1996b. Computational dialectology (2), *Area and Culture Studies* [Tokyo] 53: 115–134.

Kolb, Eduard, Glauser, Beat, Elmer, Willy and Renate Stamm. 1979. *Atlas of English sounds*. Bern: Francke.

Kretzschmar, Jr., William A. 1996a. Foundations of American English. In *Focus on the USA*, E. W. Schneider (ed), 25–50. Amsterdam: Benjamins.

Kretzschmar, Jr., William A. 1996b. Quantitative areal analysis of dialect features. *Language Variation and Change* 8: 13–39.

Linn, Michael D. and Ronald R. Regal. 1993. Missing data and computer mapping. In W. Viereck (ed) *Proceedings of the International Congress of Dialectologists. Bamberg 29.7.–4.8.1990*. Vol. 1, 253–267. Zeitschrift für Dialektologie und Linguistik. Beihefte, 74. Stuttgart: Steiner.

McDavid, Raven I., Jr. 1983. Retrospect. *Journal of English Linguistics* 16: 47–54.

Orton, Harold, et al. 1962–1971. *Survey of English dialects* (*B*): *The basic material*. Vols. 1–4. Leeds: E.J. Arnold [SED].

Orton, Harold and Nathalia Wright. 1975. *A word geography of England*. London: Seminar Press.

Orton, Harold, Sanderson, Stewart and John Widdowson. 1978. *The linguistic atlas of England*. London: Croom Helm.

Upton, Clive, David Parry and John Widdowson. 1994. *Survey of English dialects: The dictionary and grammar*. London and New York: Routledge. Reviewed by Wolfgang Viereck in *Zeitschrift für Dialektologie und Linguistik* 64 (1997).

Viereck, Wolfgang. 1986a. Dialectal speech areas in England: Orton's lexical evidence. In D. Kastovsky and A. Szwedek (eds) *Linguistics across historical and geographical boundaries. In honour of Jacek Fisiak on the occasion of his fiftieth birthday. Vol. 1: Linguistic theory and historical linguistics*, 725–740. Berlin: Mouton de Gruyter.

Viereck, Wolfgang. 1986b. Dialectal speech areas in England: Orton's phonetic and grammatical evidence. *In memory of Raven I. McDavid, Jr., Journal of English Linguistics* 19: 240–257.

Viereck, Wolfgang. 1988. The data of the 'Survey of English Dialects' computerized – problems and applications. In M. Kytö, O. Ihalainen and M. Rissanen (eds) C*orpus linguistics, hard and soft*, 267–278. Amsterdam: Rodopi.

Viereck, Wolfgang. 1991. In collaboration with Heinrich Ramisch. *The computer developed linguistic atlas of England 1*. Computational production: Harald Händler et al. Tübingen: Niemeyer [CLAE 1].

Viereck, Wolfgang. 1991. Dialectological aspects. In Viereck and Ramisch 1991: 3–9.

Viereck, Wolfgang. 1992. The computer developed linguistic atlas of England (CLAE): A preview on volume 2. *Nihon Hôgen Kenkyû-kai Genko-shû* [Dialectological Circle of Japan] 55, 55–114 (in Japanese and in English).

Viereck, Wolfgang. 1995a. Analyzing English dialect data: A quantificational approach. In K. Sornig et al (eds) *Linguistics with a human face. Festschrift für Norman Denison zum 70. Geburtstag,* Grazer Linguistische Monographien 10: 413–427.

Viereck, Wolfgang. 1995b. Méthodes en dialectométrie. *Idéologies dans le Monde Anglo-Saxon* 8: 51–73.

Viereck, Wolfgang. 1996a. English dialectology and dialectometry. In J. Klemola, M. Kytö and M. Rissanen (eds) *Speech past and present. Studies in English dialectology in memory of Ossi Ihalainen,* 447–469. University of Bamberg Studies in English Linguistics, 38. Frankfurt: Lang.

Viereck, Wolfgang. 1996b. Quantitative geolinguistics in England. In M. Reinhammar et al (eds) *Mål i sikte. Studier i dialektologi tillägnade Lennart Elmevik,* 419–432. Svenska landsmål och svenskt folkliv 1995.

Viereck, Wolfgang and Heinrich Ramisch. 1997. *The computer developed linguistic atlas of England 2*. Computational production: Harald Händler and Christian Marx. With dialectometrical contributions by Sheila Embleton, Chitsuko Fukushima, Hans Goebl, Harald Händler, Fumio Inoue, Guillaume Schiltz, Alan R. Thomas, Wolfgang Viereck and Eric Wheeler. Tübingen: Niemeyer [CLAE 2].

Wakelin, Martyn F. 1977. *English dialects. An introduction.* London: Athlone Press.

Wakelin, Martyn F. 1983. The stability of English dialect boundaries. *English World-Wide* 4: 1–15.

Wakelin, Martyn F. 1984. Rural dialects in England. In P. Trudgill (ed) *Language in the British Isles,* 70–93. *Cambridge: Cambridge University Press.*