Corpus analysis and results visualisation using self-organizing maps

Dr. H. Moisl and Dr. J. Beal
Centre for Research in Linguistics, University of Newcastle
{Hermann.Moisl, Joan.Beal}@ncl.ac.uk

This paper addresses the related issues of statistical analysis of text corpora and of intuitively-accessible representation of the results of such analysis, with reference to the Newcastle-Poitiers Electronic Corpus of Tyneside English (NPECTE) project. It proposes topographic mapping as a tool for analysis and visualization of the NPECTE corpus, and is in three main parts. The first gives a brief account of the NPECTE project, the second explains the nature of topographic mapping and the motivation for its use, and the third gives an example of how topographic mapping can be implemented using the Self-Organizing Map artificial neural network architecture.

1. The NPECTE project

The NPECTE project is based on two separate corpora of recorded speech:

(i) The earlier of the two corpora was gathered during the Tyneside Linguistic Survey (TLS) (Strang 1968, Pellowe 1972) in the late 1960s, and consists of 86 loosely-structured 30-minute interviews. The informants were drawn from a stratified random sample of Gateshead in North-East England, and were equally divided among various social class groupings of male and female speakers, with young, middle, and old-aged cohorts. Some transcription and analysis was done on this material at the time, but little of it was published, and work on it languished until 1995, when Joan Beal of DELLS secured funding from the Catherine Cookson Foundation to salvage the original reel-to-reel tapes to audio cassette format and to catalogue and archive the cassettes. This material is now housed in the Catherine Cookson Archive of Tyneside and Northumbrian Dialect in the Department of English Literary and Linguistic Studies (DELLS), University of Newcastle upon Tyne

(ii) The more recent corpus was collected in the Tyneside area in 1994 for an ESRC-funded project 'Phonological Variation and Change in Contemporary Spoken English' (PVC). This data is in the form of 18 DAT tapes, each of which averages 60 minutes in length. Dyads of friends or relatives were encouraged to converse freely with minimal interference from the fieldworker, and informants were again equally divided between various social class groupings of male and female speakers in young, middle, and old-age cohorts. This material is housed in the Department of Speech, University of Newcastle upon Tyne;

Recently, an AHRB grant was awarded under the Resource Enhancement Scheme to combine the TLS and PVC collections into a single corpus and to make it available to the research community in a variety of formats: digitised sound, phonetic transcription, standard orthographic transcription, and various levels of tagged text, all aligned.

2. Topographic mapping and its application to NPECTE

a) Topographic mapping

The aim of topographic mapping is to represent relationships among data items of arbitrary dimensionality n as relative distance in some m-dimensional space, where $m < n$. In practice, it is used in applications where there is a large number of high-dimensional data items, and the interrelationships of the dimensions are not obvious: the data items are typically represented as a set of length-n real-valued vectors $V = \{v_1, v_2 \ldots v_k\}$, and these vectors are mapped to points on a 2-dimensional surface such that the degree of similarity among the v_i is represented as relative distance among points on the surface.

b) Motivation and application to NPECTE

Corpus analysis is often concerned to discover regularities in the interrelationships of certain features of interest in the data – correlations of phonetic or graphemic features, for example, or of such things as social class, age, gender and geography with aspects of linguistic usage. Cluster analysis (Everitt 1993) has been widely and successfully used for this purpose (Manning and Schütze 1999), and topographic mapping is in fact a variety of cluster analysis. Its chief advantage over standard

cluster analysis techniques is the intuitive accessibility with which analytical results can be displayed: projection of a large, high-dimensional data set onto a two-dimensional surface gives an easily-interpretable spatial map of the data's structure.

With regard to the application of topographic mapping to the NCEPTE corpus in particular, the project's aim is not only to create an electronic resource, but to make that resource the basis of analytical research projects. We are therefore developing software tools to supplement those generally used in corpus analysis, and topographic mapping is the first of these.

3. Implementation of topographic mapping using the SOM architecture

There are several ways of implementing topographic mapping, that is, of forming two-dimensional projections of data distributions in high-dimensional spaces: principal component analysis (Jolliffe 1986, Everitt 1993), multidimensional scaling (Borg and Groenen 1997, Everitt 1993), and self-organizing maps (SOM) (Kohonen 1995). This paper adopts the last of these because SOMs have been successfully used in natural language corpus processing, and the relevant work provides a good basis for development of the applications required for the NPECTE. This section briefly describes the SOM architecture, then gives pointers to current applications of SOM in processing of textual corpora, and finally presents an example of how a SOM can be used in analysis of corpora like the NPECTE.

a) SOM

The self-organizing map, also known as the Kohonen net after its inventor, is a *k*-dimensional surface of processing units, where *k* is usually 2. Associated with each unit is a set of connections from an input buffer such that, for a buffer of length *n*, there are *n* connections per unit (for clarity, only sample connections are shown in Figure 1):

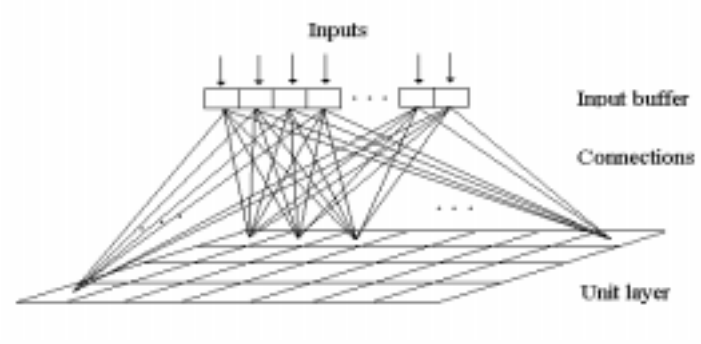

Figure 1: A self-organizing map.

Given a set V of input vectors of length *n*, such a net can, using the SOM training algorithm, learn to approximate the similarity relations among the $v_i \in V$ in *n*-dimensional space on the two-dimensional surface of processing units. After training is complete, each of the v_i is associated with a specific unit u_j in the sense that it activates u_j more strongly than any other; when the activations for all the v_i are plotted on the unit surface, the distances among activated units represent the similarity relations in the input vector space. Details of the network training algorithm can be found in most textbooks on artificial neural networks (for example Haykin 1999, Rojas 1996); the standard reference is Kohonen 1995.

b) SOM and corpus analysis

SOMs have found application in a wide range of disciplines (Kohonen 1995, chapter 7). In natural language processing (Kohonen 1995 pp 237-249, 301; Honkela 1997), the main application to date has been in the classification of texts in large document collections. In particular, Kohonen and his research group have developed WEBSOM (Kohonen *et al* 2000, Kaski *et al* 1998, Lagus *et al* 1999), a system that has successfully classified over one million web documents on the basis of their lexical content. WEBSOM underlies development of the NPECTE-specific analytical tool being described here.

c) Example

Assume the existence of a phonetically-transcribed spoken corpus, like NPECTE, consisting of a fairly large number of interviews, each labelled for region, age, gender, and social class. One is interested in the, say, region and age distribution of phonetic segments in two environments, that is, of segments that occur between two specific (phonetic prefix - phonetic suffix) pairs. To carry out the analysis, the transcribed corpus is scanned for the relevant prefix-segment-suffix sequences, labelling

each such sequence with the regional and age information associated with the interview from which it came. The aim is to show how a SOM can generate and display a topographic map of the structure of such a data set. To show this data with a known structure is required; for clarity of exposition, a small, artificially constructed data set D1 will be used.

The phonetic segment of interest comprises 5 variants, V1 - V5, distributed as shown in Figure 2:

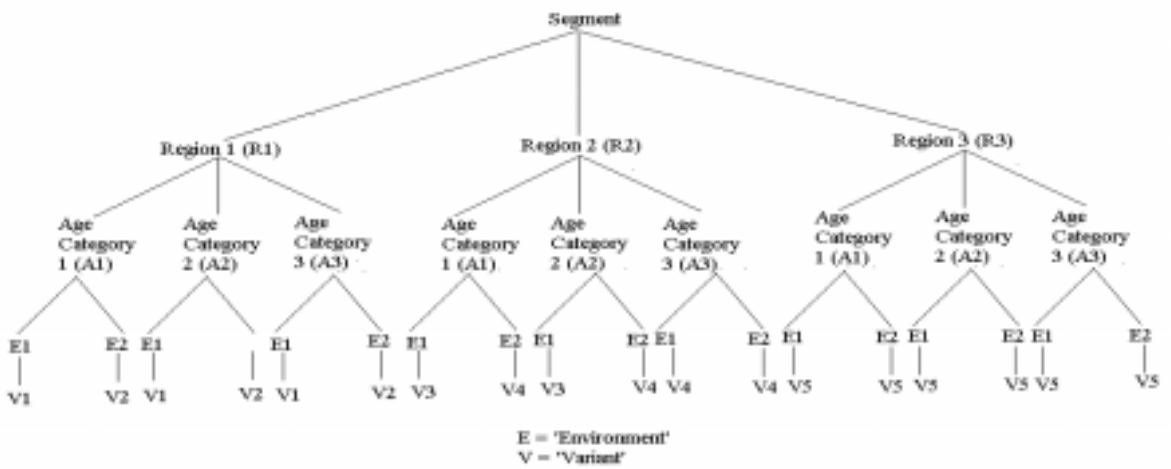

Figure 2: The structure of the example data set D1

- In Region 1 all age categories use variant 1 in environment 1, and variant 2 in environment 2
- In Region 2 age categories 1 and 2 use variant 3 in environment 1 and variant 4 in environment 2, but age category 3 uses variant 4 in both environments
- In region 3 all age categories use variant 5 in all environments

The first step is to encode the environmental prefixes and suffixes and the segment variants for processing by a SOM. This means some form of vector encoding. Again for clarity, a binary encoding is adopted, where 1 and 0 represent the presence and absence respectively of a phonetic feature; prefixes and suffixes are encoded as 3-bit and segment variants as 6-bit binary vectors:

E1 prefix: 010	V1: 100011
E1 suffix: 101	V2: 100111
E2 prefix: 011	V3: 010011
E2 suffix: 110	V4: 010111
	V5:001111

The encodings are arbitrary, and are not intended to be interpretable as specific phonetic features. The data set corresponding to the structure in Figure 2 is thus:

1. R1 A1 E1	0 1 0 1 0 0 0 1 1 1 0 1	10. R2 A2 E2	0 1 1 0 1 0 1 1 1 1 1 0
2. R1 A1 E2	0 1 1 1 0 0 1 1 1 1 1 0	11. R2 A3 E1	0 1 0 0 1 0 1 1 1 1 0 1
3. R1 A2 E1	0 1 0 1 0 0 0 1 1 1 0 1	12. R2 A3 E2	0 1 1 0 1 0 1 1 1 1 1 0
4, R1 A2 E2	0 1 1 1 0 0 1 1 1 1 1 0	13. R3 A1 E1	0 1 0 0 0 0 1 1 1 1 0 1
5. R1 A3 E1	0 1 0 1 0 0 0 1 1 1 0 1	14. R3 A1 E2	0 1 1 0 0 0 1 1 1 1 1 0
6. R1 A3 E2	0 1 1 1 0 0 1 1 1 1 1 0	15. R3 A2 E1	0 1 0 0 0 0 1 1 1 1 0 1
7. R2 A1 E1	0 1 0 0 1 0 0 1 1 1 0 1	16. R3 A2 E2	0 1 1 0 0 0 1 1 1 1 1 0
8. R2 A1 E2	0 1 1 0 1 0 1 1 1 1 1 0	17. R3 A3 E1	0 1 0 0 0 0 1 1 1 1 0 1
9. R2 A2 E1	0 1 0 0 1 0 0 1 1 1 0 1	18. R3 A3 E2	0 1 1 0 0 0 1 1 1 1 1 0

Hierarchical cluster analysis (squared Euclidean distance, average linkage) reveals the structure of this data (Figure 3):

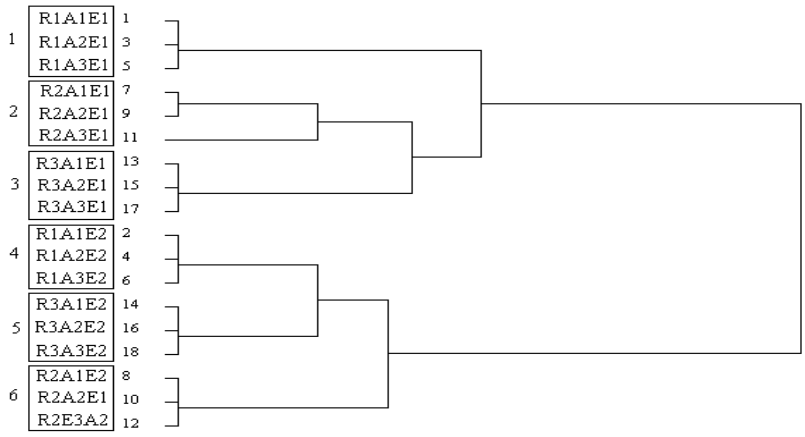

Figure 3: A hierarchical cluster analysis of D1

The two main clusters (1-3) and (4-6) correspond to environments E1 and E3, and within both of these there is subclustering first by region and then by age. This is the structure which the SOM is expected to discover from the data.

A SOM was trained on S1 with the following parameters:

Map axis: 9 (that is, a 9 x 9 unit layer)
Initial learning rate: 0.9
Learning rate decrement: 0.01
Learning rate decrement interval: 10 iterations
Initial neighbourhood: 9
Neighbourhood decrement interval: 40 iterations
Number of training iterations: 10000

After training the S1 vector set was presented to the net, with the following result (Figure 4):

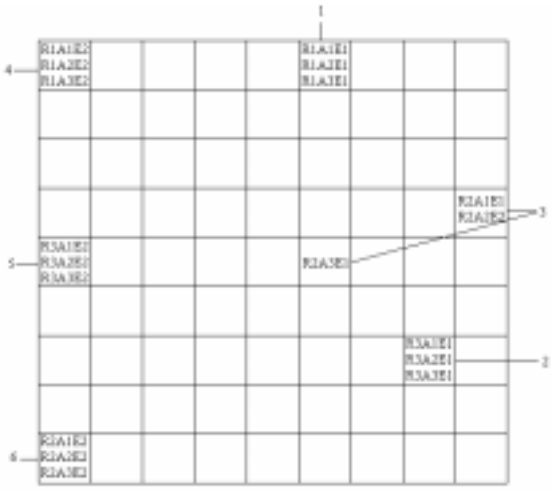

Figure 4: The SOM's analysis of D1

The main E1 and E2 clusters are clearly separated from one another on the left and right sides of the map. The groups in the E2 region are equidistant from one another, corresponding to the E2 subtree in Figure 2; the distance match with the hierarchical cluster tree, where the distances among 4, 5, and 6 are slightly asymmetrical, is not exact, but the map approximates this as closely as possible given its coarse granularity. The EI region also closely reflects the EI cluster subtree since, in both, 2 and 3 are closer to one another than they are to 1. In addition, the cluster tree shows that the structure of group 2 is unlike that of the other groups 1 and 3-6 in that R2A3E1 differs substantially from the other two members of the group; the map shows a corresponding distance relation.

It can, therefore, be said that the SOM gives a good 2-dimensional spatial representation of the vector similarity relations in the data set, which itself encodes regional, age, and phonetic environment variation in our hypothetical corpus.

Now, the hierarchical cluster tree is easily as clear about the structure of the data as the SOM. What, therefore, is the advantage of SOMs over established cluster analysis methods like hierarchical analysis in corpus work? The answer is that, as data sets grow larger and their structure more complex, existing hierarchical methods become increasingly difficult to interpret, whereas SOMs remain clear. Consider, for example, another artificial data set D2 of 1000 length-24 real-valued vectors. These were generated by a process which subdivided them into 5 main groups, numbered (0-200), (201-500), (501-800), (801-950), and (951-1000). Each was then given some subsidiary structure, and finally noise was injected into the whole set by randomly selecting two components of each vector and incrementing the values found there by a small random amount. Figures 5 and 6 show the results of hierarchical cluster analysis (Euclidean distance, single linkage) and SOM analysis respectively.

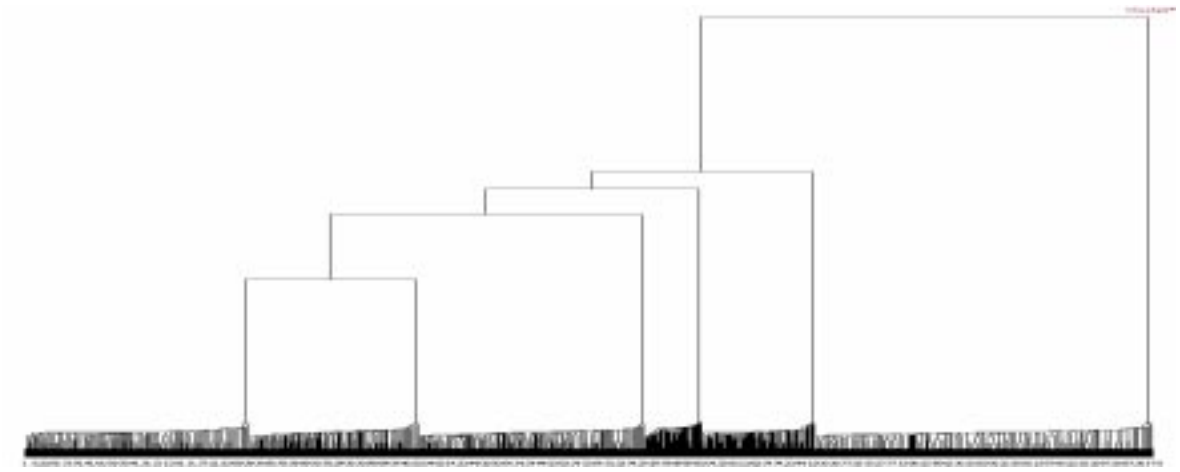

Figure 5: A hierarchical cluster analysis of D2

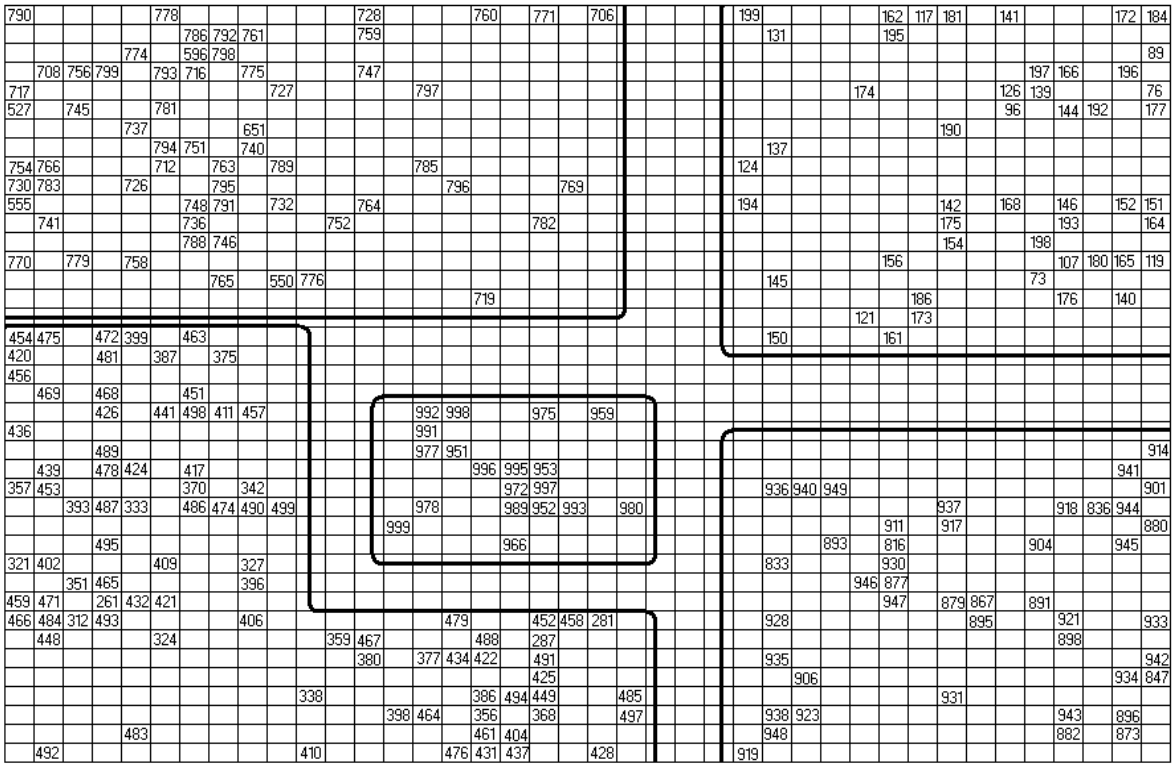

Figure 6: A SOM analysis of D2

Both are clear with respect to the main structure of the data, but the subsidiary structure within the main clusters is much clearer in Figure 6. Both hide some structural information. In Figure 5 it is almost entirely opaque and difficult if not impossible to comprehend. In Figure 6 not all 1000 vectors

are present on the map: many vectors are mapped to a single unit and, since only one vector label can be represented at any given location, only one vector location can be displayed. Thus, the upper right-hand corner of the SOM represents vectors 1-200, but only a minority of these is visible. The solution is both cases is to implement a graphical interface that permits interactive browsing of the structure display. For Figure 5 this would reveal ever more detailed subtrees, but, as data sets grow larger, such zooming-in soon make it difficult to see the selected subtree region in relation to the structure tree as a whole. For Figure 6 there are at least two possibilities. On the one hand, selection of a given unit could display a list of vectors associated with that unit in a way that allows the analyst to maintain a clear view its place in the overall structure map, as in Figure 6a. On the other, and more interestingly, one could use a hierarchical feature map (Merkl 2000) in which each unit of the main SOM has its own SOM associated with it, allowing the structure of the vectors mapped to the node of interest to be displayed, as in Figure 6b.

Conclusion

Topographic mapping is a nonhierarchical clustering technique that can project high-dimensional data sets onto low, usually two-dimensional surfaces such that the similarity relations of the data are represented as spatial distribution of points on the surface. In relation to text corpus analysis, its main

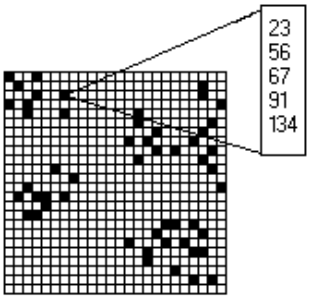

Figure 6a: Vector list at selected node

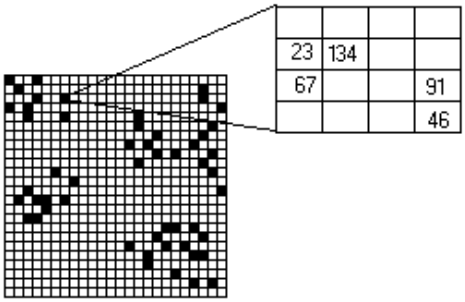

Figure 6b: SOM at selected node

advantage over standard hierarchical cluster analysis methods for the purpose is in the intuitive clarity of results visualization as a spatial structure map, and the scope of that visualization for interactive exploration of the map. The aim is to develop a SOM-based implementation of a topographic mapping tool for analysis of the NPECTE corpus.

References

Borg I, Groenen P 1997 *Modern Multidimensional Scaling - Theory and Applications*. Springer.

Everitt B 1993 *Cluster Analysis*, 3rd ed. E. Arnold.

Haykin S 1999 *Neural Networks. A Comprehensive Foundation.* Prentice Hall International.

Honkela T 1997 *Self-Organizing Maps in Natural Language Processing*. PhD thesis, Helsinki University of Technology, Espoo, Finland.

Jolliffe I 1986 *Principal Component Analysis*. Springer.

Kaski S, Honkela T, Lagus K, Kohonen T 1998 WEBSOM--self-organizing maps of document collections. *Neurocomputing* 21: 101-117.

Kohonen T 1995 *Self-Organizing Maps*, 2nd ed. Springer.

Kohonen T, Kaski S, Lagus K, Salojärvi J, Paatero V, Saarela A 2000 Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks* 11(3): 574-585.

Lagus K, Honkela T, Kaski S, Kohonen T 1999 WEBSOM for textual data mining. *Artificial Intelligence Review* 13(5/6): 345-364.

Manning C, Schütze H 1999 *Foundations of Statistical Natural Language Processing.* MIT Press.

Merkl, D Text data mining. In Dale R, Moisl H, Somers H (eds), *Handbook of Natural Language Processing*, Dekker: 889-903

Pellowe J *et al* 1972 A dynamic modeling of linguistic variation: the urban (Tyneside) linguistic survey. *Lingua* 30: 1-30.

Rojas R 1996 *Neural Networks. A Systematic Introduction.* Springer.

Strang B 1968 The Tyneside Linguistic Survey, *Zeitschrift für Mundartforschung*, Neue Folge 4: 788-94.