Recent Advances in Salzburg Dialectometry

Hans Goebl

Department of Romance Philology, University of Salzburg, Austria

Abstract

This paper documents the many taxometric and cartographic achievements of the Salzburg school of dialectometry. The paper discusses the following topics: (1) problems of measurement of linguistic atlas data (with particular consideration of Romance linguistic atlases), (2) establishment of the data matrix, (3) choice of the similarity index (Relative and Weighted Identity Value), (4) generation of the respective similarity and distance matrices, (5) their subsequent cartographic exploitation, which encompasses the following cartographic tools: similarity maps, parameter maps, dendrograms (and their spatial projection), and correlation maps. The ultimate purpose of these highly sophisticated cartographic techniques (choropleth and isopleth maps) is to increase our knowledge of the complex mechanisms of the dialectal management of space by man. From a methodological point of view our paper deals with problems related to (Romance) dialectology and linguistic geography, historical linguistics, numerical classification, statistics and statistical cartography. The examples are drawn from the French linguistic atlas ALF (*Atlas linguistique de la France*) published by Jules Gilliéron and Edmond Edmont (Paris: Champion, 1902–1910, 10 volumes) more than one hundred years ago. The taxometric calculations and their respective visualizations are realized by a powerful computer program called 'Visual DialectoMetry' (VDM), created by Edgar Haimerl (Blaustein, Germany) between 1997 and 2000 in Salzburg, which is freely available for research purposes.

Correspondence:
Hans Goebl,
Fachbereich für Romanistik,
Akademiestrasse 24,
A-5020 Salzburg, Austria.
E-mail:
Hans.Goebl@sbg.ac.at

1 Preliminary remarks

Since 1973, when the French geolinguist Jean Séguy (1914–1973) created a new approach to dialectology, including even the term *dialectométrie* (DM), the range of activities which can be regarded as 'dialectometrical' have largely diversified in both methodical and regional respects. Thus, it is legitimate to speak today of different dialectometrical schools.[1] One of these research centres is in Salzburg.

1.1 Brief characterization of the Salzburg school of dialectometry (S-DM)

In the early seventies, the Salzburg version of dialectometry developed geolinguistic research with the exploration of Romance linguistic atlases (in particular the *Atlas linguistique de la France* (ALF) and the *Atlante italo-svizzero* (AIS). S-DM assumes that in linguistic atlases such as the ALF and the AIS there are not only superficial organizing geolinguistic structures, but also 'deeper structures' which can only be uncovered by the quantitative

synthesis of many single atlas maps. This quantitative synthesis is realized with a statistical method called 'numerical classification, taxometry, data analysis, data mining', etc. as exemplified in the classical handbooks of Sneath/Sokal (1973), Chandon/Pinson (1981) and Bock (1974). The resulting measurement values are directly cast in maps, i.e. appropriately visualized. S-DM is therefore characterized by three basic methodological principles:

- empirically: by the investigation of Romance linguistic atlases,
- numerically: by the use of appropriate procedures of numerical classification,
- heuristically: by the generation of complex maps (choropleth maps, honeycomb maps, beam maps,[2] trees, etc.) which lend visual form to quantitative evidence.

These objectives of the S-DM also continue the tradition of classical Romance geolinguistics, which has always worked with maps and appropriate visualizations.

1.2 The central question in S-DM research

The Salzburg research team defines the main aim of DM in the exploration of the 'basilectal management of space by the HOMO LOQUENS'. This basic theoretical position assumes that geographical 'space' is a texture of complex relations in which man evolves and settles by means of speech. Thus, human speech is considered as being *variable* and space as *invariable*. Insofar as this theoretical point is also relevant to many other forms of human manifestations of life in space and can further be applied to all sciences dealing with these relations (such as history, geography, economy, population genetics, ethnography, anthropology, etc.), there result *eo ipso* many opportunities for interdisciplinary cooperation with S-DM.

1.3 Methodical principles

As shown in Figure 1, S-DM makes use of a chain of methods, including the following elements or stages:

(1) establishing of the data matrix

(2) establishing of the similarity and distance matrix

(3) different interpretations[3] of the similarity and/or distance matrix.

Ad 1: The data matrix (with N inquiry points, localities or sites, and p working maps) is established by measurements ('taxations') of the original (qualitative) atlas data (here: ALF): the phonetic, morphosyntactic and lexical raw data are fed into the data matrix in the form of discrete nominal (qualitative) units (called 'taxates'). In fact, this is a procedure which has already been in use in Romance linguistics for several decades.

Ad 2: The similarity matrix is created by measuring the (aggregated and thus quantitative) similarity between (pairs taken from) the N vectors of the data matrix. This kind of measurement should not be confounded with implicational scales. For a better understanding of the general procedure of the similarity measurement, let us take a look at the left half (Data matrix A) of Figure 2. Assuming that we wish to measure the overall similarity between the atlas points 2 and 3, we have to take into account the quantitative relation between the following pairs of linguistic attributes: row 1: a/a, row 2: *e/f*, row 3: h/h, row 4: j/j. Hence we see that there exist pairwise identities (also called 'co-identities', COI) in rows 1, 3 and 4 while the row 2 constitutes a *co-difference* (*COD*). Summing up the number of the COI (here: 3) and dividing it by the sum of the COI and the COD (here: 4) we obtain 'Relative Identity Value' (RIV), in this instance, the $RIV_{2,3}$ score (here 75%). The score of 75% can be found in the scheme of the Similarity matrix s(a) A, located to the right of the Data matrix A.

It is possible – as demonstrated in the above mentioned handbooks of numerical classification (such as Sneath/Sokal 1973) – to use different similarity measures or indexes in creating the similarity matrix. S-DM has indeed attempted to re-use and represent the traditional concepts and ideas of geolinguistic similarity with taxometrical methods, as S-DM views itself as being in the service of traditional geolinguistics and as its supplement or 'sympathetic successor'.
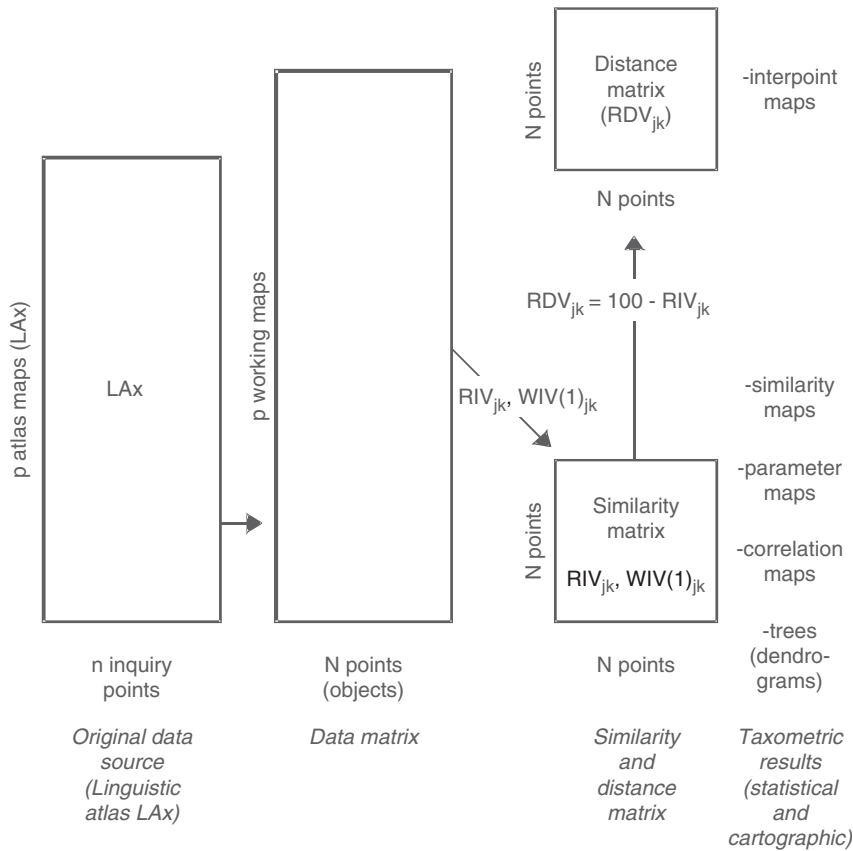
Fig. 1 Flow chart of the dialectometrical methods used by the Salzburg school of dialectometry.

The distance values (*dist*) in the distance matrix are calculated from the similarity values (*sim*), with the following formula: *dist + sim = 100*. Usually, the distance matrix is used solely to generate so-called 'honey comb maps',[4] which show the quantitative combination of isoglosses.

Ad 3: The S-DM tries to interpret the measurement values stored in the similarity and distance matrices in such a way that traditional geolinguistics can work with them as well. Therefore, it is of utmost importance that the respective dialectometrical maps give clear answers to the research questions which had already been fixed and discussed before. However, such answers ought to be more elaborate, more comprehensive and more precise than the previous ones and suggest in addition some new insights, too.

A complete version of the most important DM methods developed in Salzburg is given in my habilitation thesis *Dialektometrische Studien* (Goebl, 1984, 3 volumes). Since 2000, most of the methods and procedures described there have been converted into a very handsome computer program called 'Visual DialectoMetry' (VDM), which was created by our senior research assistant Edgar Haimerl. The aim of VDM is double:

- numerical: rapid execution of many statistical computations,
- graphical: the optimal visualization of the numerical results. Usually, the visualization employs different colours. As the present contribution shows black and white graphics only, the reader unfortunately cannot see the whole range of VDM's visual capabilities.
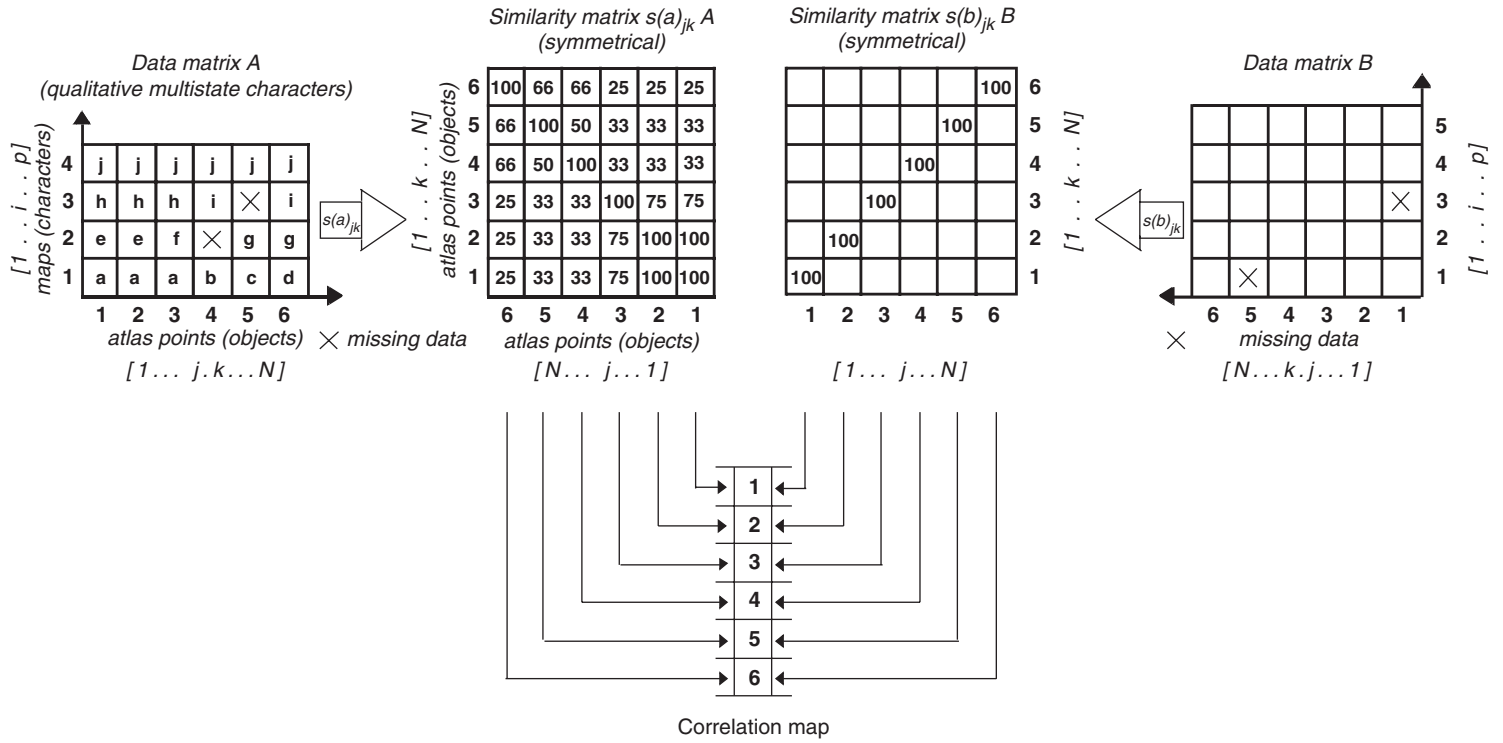
Fig. 2 Scheme of the calculation of the interdialectal similarities and of the correlation maps. Note the symmetry of the two similarity matrices (A and B)!

1.4 Some statistical-mathematical remarks

The basic epistemological purpose of the S-DM is explorative, which means that the empirically existing deeper structures (viz. regularities, or even 'laws') of the atlas data have to be identified and recognized, in order to know more about the nature of 'language and speech in space'.[5] For this reason procedures of descriptive statistics are preferable in the first instance.

At this point, the following dilemma must be mentioned: the quantitative maps of S-DM represent (at times very complex) patterns and structures which are optically perceived by the geolinguist holistically, and are mentally processed initially in this holistic way. Although it is possible to discuss or test the underlying quantitative measurements by means of inferential statistics, it is not (yet) possible to do so with the holistic patterns and structures (insofar as their visual form is concerned). The geolinguist is only able to appreciate them *de visu*, according to the more or less 'fruitful' suggestions they make. But he cannot holistically test the scientific sense of the patterns with traditional statistical methods. This dilemma has not been appreciated sufficiently, perhaps not at all, to-date.

1.5 Previous reception of S-DM

Apart from Salzburg or the other places where I have worked in the past,[6] S-DM has met with a only limited interest; see the respective bibliography in Goebl, 1993: 277–278, and compare also the publications of my scholar Roland Bauer (2003 and 2004). Whereas in Germany and different Romance speaking countries like France (including the French speaking part of Canada), Spain (including Catalonia,[7] Galicia, and the Basque Provinces[8]) and Portugal[9] the interest in DM was and is still quite great, it is completely absent in Italy (cf. Grassi 2001). Actually, in Italy there seems to be no intellectual need for the synthetic interpretation of linguistic atlases, as well as for these specific research objectives. Nerbonne and Kretzschmar (2003) discuss it prominently in their introduction to a special issue of the journal *Computers and the Humanities* on dialectometry.

Furthermore, I have the impression that traditional linguistic geography is still more interested in the setting up of linguistic atlases, followed by a atomistic interpretation or reading of the data, than in putting the question whether there are *prima vista* hidden global structures or anything similar in the huge amount of data of the linguistic atlases. I personally also think that many geolinguists have still not grasped the empirical importance and the theoretical range of the communicative challenge of geographical space in human verbal behavior.

1.6 Desiderata and perspectives for the future

All scientists working with DM should try to reunite and contrast different DM-approaches and thereby to increase their own experiences in DM. This can mainly be realized by analyzing a given linguistic atlas (e.g. the French atlas ALF, the Italian atlas AIS or the Dutch atlas RND) with the means of different DM-methods, and by then comparing the respective results. Our VDM program is prepared for such analyses. It is even able to process externally generated data and similarity matrices and to visualize the corresponding deeper structures using all methods implemented in VDM. This opportunity for cooperation is of course not limited to geolinguistics; it is open to all geographically oriented sciences.

1.7 Technical remarks concerning the 16 maps of this paper

As it is intended that the 16 maps of the appendix be compared by the reader, they are reproduced in groups of four. It is indeed regrettable that the colouring of the maps could not be realized in the present contribution. Notice that coloured DM-Maps are published in the cartographic appendices of some of my earlier DM-works (cf. 1993, 2000, 2002, 2003, 2004, and 2005).

2 From raw data to similarity maps

This contribution will only deal with the results of the dialectometrization of the French atlas ALF.

For the purposes of this project (realized in the years 1997–2000) 626 out of the 1421 original ALF-maps were analysed and 'mined' for their phonetic, lexical and morphosyntactic information. We added three artificial sites, or 'false dialects' to the original 638 inquiry points of ALF (corresponding to the standard forms of French, Italian, and Catalan). Subsequently, the grid of the 641 sites was triangulated and then reset in a polygon structure in accordance with the principles of Delaunay-Voronoi-geometry. See Figure 3.

The data matrix derived from the 626 original ALF-maps contains 641 sites ($=N$) and 1687 ($=p$) 'working maps'. We can derive 1687 *working* maps from 626 *original* ALF maps because many ALF maps, which are based on only a single lexical type (e.g. ALF 13 *aigle* [< lat. ÁQUILA], 14 *aiguille* [< lat. *ACÚCLA[10]], etc.), result nonetheless in 2, 3 or even more different phonetic taxations, i.e. working maps.

An example for a (very simple) lexical taxation is given in Map 1 (derived from ALF 6 *acheter*): it relies on only three lexical taxates (*acheter* < lat. ACCAPTARE, *croumpa* < lat. COMPARARE, *ana croumpa* < lat. AMBITARE + COMPRARE) which obviously appear in many different phonetic realizations in the map. There are, however, also lexical working maps in our data matrix which have up to 91 taxates. The skewed distribution of the ALF's taxates' frequencies recalls the well-known power laws for linguistic distributions discovered by the Harvard linguist Georges Kingsley Zipf.[11]

The data matrix contains nominal (or categorical) data. For the subsequent measurement of the similarities between (the pairs of) the 641 vectors of the data matrix the 'Relative Identity Value' (RIV_{jk})[12] and the 'Weighted Identity Value (with the weight 1)' [$WIV (1)_{jk}$] are usually applied. Actually, the RIV_{jk} is the standard procedure for computing in S-DM. It corresponds very closely to common ideas of linguistic similarity shared by many (geo)linguists.

After setting up the (square) similarity matrix (here: with the dimensions 641 by 641), we have to interpret it step by step. One of the first steps is the establishment of similarity maps. Each of these similarity maps relies on one of the N ($N = 641$) similarity distributions of the similarity matrix and consists of 641 measurement values, where one[13] of them, the (reflexive) similarity of the reference site j to itself, is equal to 100% ($RIV_{jj} = 100$). The remaining 640 measurement values below 100% are subsequently visualized with the competent support of statistical cartography. See Map 2, which shows the position of a Gascon dialect (corresponding to ALF site 682, Tartas, *Département Landes*) within the whole investigation grid. In this instance, the 640 measurement values that are to be visualized oscillate between 37,62[14] (%) and 86,25[15] with as arithmetic mean 54,09[16] (see the numerical legend at the lower end of the of Map 2).

The algorithm of visualization MINMWMAX[17] provides for 6 intervals: three of them (intervals 1–3) below the arithmetic mean, and three above the arithmetic mean (intervals 4–6). The spatial distribution of the hatchings (intervals 1–3) and of the shadings (intervals 4–6) generates a suggestive and well-structured choropleth profile, which is of the utmost interest and significance for the geolinguist and for the dialectometrician.[18] Three results can be derived:

(a) the position of the dialect of Tartas within the Occitanian area (*Domaine d'Oc*), and its relations to other similar sites (or 'related' dialects),

(b) the position of Gascoigne and the whole *Domaine d'Oc* in the ALF grid,

(c) the position of Northern French (*Domaine d'Oïl*) in regard to Southern French, respectively Occitanian (*Domaine d'Oc*), in general.

One can further note a progressive, though not regular drop in RI values with increasing geographical distance from the reference point j. This constant drop is visible in all similarity maps and points to something like 'spatial laws'.

Moreover, each single similarity map can be interpreted analogously:

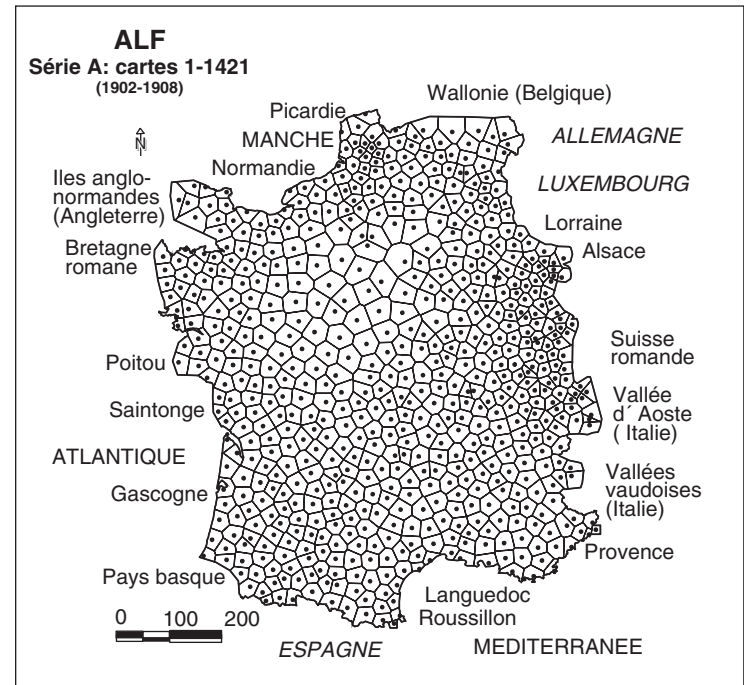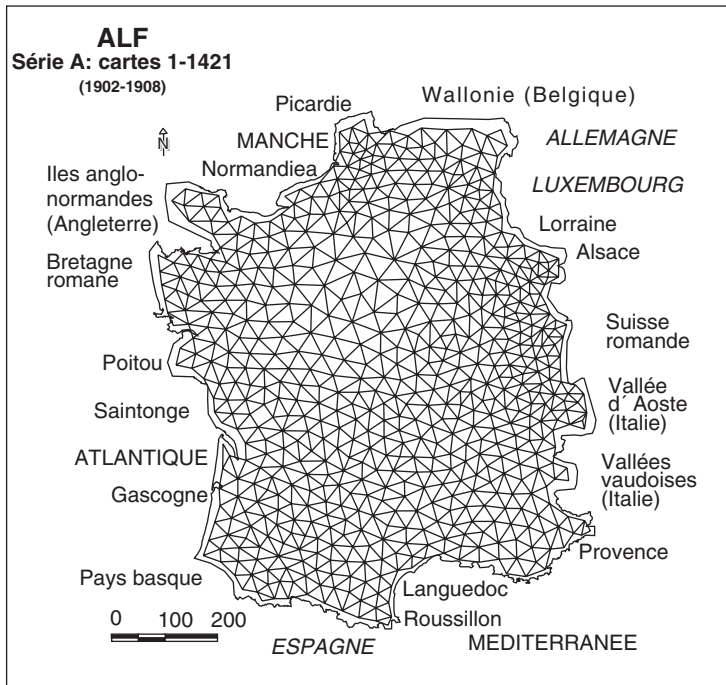(a) by analogy with human societies: the polygons in the intervals 6 and 5 indicate the position of

Fig. 3 Delaunay-triangulation (left) and Voronoi-tesselation (right) of the grid of the ''Atlas linguistique de la France'' (ALF).

the actor's (=the reference site) 682 (Tartas) 'best friends', whereas the intervals 1 and 2 locate his 'greatest enemies'.

(b) by analogy with general diffusion processes: the stratification of the choropleth profile in Map 2 indicates the actor's (=reference site) 682 (Tartas) 'efforts' in the whole dissemination process, and shows where and to what extent they succeed within the whole grid.

2.1 Similarity maps and different linguistic categories

See the Maps 3 and 4.

In Romance linguistics researchers have long inquired whether quite similar or clearly different patterns are found when a large number of working maps belonging to different linguistic categories (such as phonetics, vocabulary, etc.) are synthesized. The syntheses of various isogloss provide evidence that relatively similar patterns appear. These comparisons can also be realized by means of DM. See Maps 3 and 4, which rely on two different subcorpora of our total ALF grid: Map 3 for phonetics (1117 working maps), and Map 4 for vocabulary (471 working maps).

It is fairly easy to recognize that the geographical regularity of the choropleth profile of Map 3 (phonetics) is greater that the one of Map 4 (vocabulary). This difference, however, does not prevent the sites in close vicinity to the reference site from being quite similar to each other: the spatial distributions of the phonetic and lexical 'best friends' are similar. Although it seems that the lexical management of space in the dialectal Gallo-Romance area basically follows the same principles as the phonetic management of that same space, the deeper linguistic regularities which are involved are apparently not completely identical.

2.2 Similarity maps and different similarity measurements

See Maps 5 and 6.

Obviously, the nature and quality of the selected similarity index determines the graphic structure of the similarity map. It is further obvious that the dialectometrician's choice of the similarity index has to be guided previously by definite (geo)linguistic ideas or 'theories'. In the field of Indo-European studies, as well as in German and Romance linguistics, language features which rarely occur are given a particular prominence in language typology and classification: less common linguistic features are either quite old or are an indication of borrowing and therefore need to be treated with special attention.

Taking into account these considerations I defined in 1983[19] the 'Weighted Identity Value (with the weight x)' [WIV(x)$_{jk}$], which has also been implemented in the VDM program. The corresponding similarity maps display a far better structured choropleth profile in the neighbourhood of the reference site (intervals 5 and 6) and in the range of the smallest similarity values (i.e. at the 'antipodes' of the inquiry site). The resulting iconic profiles are less smoothly structured than the profiles generated with the 'Relative Identity Value' (RIV). This may be very useful for certain classifications. One may also wish to introduce the option of weighting in the formula of the WIV(x)$_{jk}$. Theoretically, the factor x could vary between 1 and ∞. As it increases, the measurement values of WIV(x)$_{jk}$ and the values of the RIV$_{jk}$ converge.

A comparative look at the Maps 5 and 6 illustrates very clearly this last remark. Both maps show what is to be considered as a 'typical' choropleth profile for the centre of Normandy. The choropleth profile of Map 6 shows far better than that of Map 5 the iconic pattern in the proximity of the inquiry site (on account of the smaller number of polygons in the interval 6). The same can be noted concerning the position of the 'antipodes' of the inquiry site (on account of the much larger number of polygons in the interval 1). Because of these two enhancing effects, the choropleth profile of Map 6 shows a more uneven or 'hilly' structure, with a sharper differentiation than the profile of Map 5. This is a very profitable result for a certain number of geotypological purposes.

Nerbonne and Kleiweg (2006) argue for the use of WIV(x) (which they refer to as 'Goebl weighting') based on a quantitative heuristic which aims to reflect the geographic coherence exposed by dialectometric analysis.

3 Beyond the similarity matrix I: the parameter maps

See Maps 7 and 8.

Work with the similarity maps reveals that the inclusion of the different 'characteristic parameters' of the respective similarity distributions (such as maximum, minimum, arithmetic mean, standard deviation, etc.) is also very useful for geolinguistic objectives. Actually, many of these characteristic parameters highlight phenomena that are well-known in geolinguistics. Thus, the famous 'dialect kernels' can be analyzed by attending to the maxima of similarity distributions, and the phenomenon of 'linguistic compromise or exchange' by examining the 'skewness'.

3.1 In search of 'dialect kernels': the synopsis of the maxima of similarity distributions

See Map 7.

From a technical point of view, the whole procedure is simple. For each of the 641 sites, we calculate the respective similarity distributions using a given similarity index (e. g. RIV_{jk}). Ignoring the 641 reflexive similarity values ($RIVjj$), which are naturally always at 100%, we collect the maximum values (below 100%) of these 641 similarity distributions and then visualize them in the usual ways. The resulting profile resembles a 'landscape' with a harmonious structure of 'mountains and valleys', in which the individual 'mountains' (intervals 6 and 5) are distinctly separated by the 'valleys' in between (intervals 1 and 2). In Map 7, one clearly recognizes a 'plateau' in the North which is surrounded by isolated 'peaks' in the West and the North (Picardy), as well as a less compact 'plateau' in the area of Eastern Gascoigne, the Languedoc and – separated by some 'valleys' – the Provence.

In between there are transitional zones everywhere, which look like valleys (see Map 7).

The geolinguistic significance of this map is obvious: the agglomeration of extremely close dialects is always regionally limited, less in the North, more in the South. In between – i.e. in the transitional area between *Domaine d'Oc* and *Domaine d'Oïl* – the agglomerative effects decrease markedly. Naturally, this 'mountains and valleys' effect is more easily perceptible when different colours are used, or even with a coloured stereographic visualization.[20]

3.2 For a better understanding of 'linguistic compromise and exchange': the synopsis of the skewness values of similarity distributions

See Map 8.

The term and concept 'linguistic compromise or exchange' (called in German *Sprachausgleich*, in French *compromis linguistique*) was originally introduced in German linguistics,[21] in order adequately to describe the manifold linguistic contacts between the great dialect areas of the Middle Ages and the present (such as Rhine-Franconian [*Rhein-Fränkisch*], Bavarian [*Bairisch*], Alemannic [*Alemannisch*], Thuringian [*Thüringisch*], etc.). In substance, 'linguistic compromise' is defined as the accumulation of linguistic exchange and (ad)mixture phenomena of all kinds that can exist over a shorter and/or longer distance. The result of the exchange is a linguistic intermixing which is strong to different degrees. Obviously, it is less important in those dialectal zones which participate little or not at all in the dynamics of interplay and exchange.

This phenomenon can be detected statistically by the measurement of the symmetry of a given similarity distribution. Similarity distributions skewing to the left (negative skew, where most of the similarity values are concentrated below the arithmetic mean) tend to indicate isolated dialects, while similarity distributions skewing to the right (positive skew, where most of the similarity values are concentrated above the arithmetic mean) indicate expanding or well-integrated dialects.

This varied and puzzling picture is visualized by the changing shape of the histograms of the Maps 2, 5, 6, and 11.

In geolinguistic respects, Map 8 has a very clear shape which can easily be interpreted from a diachronic point of view. In the North, the polygons are arranged in a circular pattern (intervals 1 and 2), which has its centre in the Ile-de-France. On the Eastern border of the map, a smaller half circle (generated by polygons in interval 1) surrounds from the North, West and South, the Francoprovençal. In the South of the map four 'bulwarks' (marked in dark) can be detected, located in Gascoigne, the Roussillon, the Languedoc and the Provence. The transitions between these areas are loosely structured. The areas assigned to the intervals 1 and 2 correspond to zones of strong linguistic exchange and intermixing, whereas the areas assigned to the intervals 5 and 6 only weakly participate in the process of linguistic contact and exchange within the whole Gallo-Romance area. Thus, the Northern periphery of the ALF grid reveals much more open to linguistic contacts, whereas the South plays a rather conservative and defensive role in this process.

The whole profile of the map (which would show better in different colours or as a coloured stereographic visualization[22]) reveals the bipartite linguistic history of the Galloromania: in the North a dynamic propagation of the linguistic type of the Ile-de-France on the one side, in the South a narrow competition of different major regiolects (above all: the dialects of Gascoigne, Languedoc and Provence) on the other side. The Francoprovençal, representing the old Latin of Lugdunum/Lyon, and pressed (and thus lingui compromised) by the Domaine d'Oïl in the North and the Domaine d'Oc in the South, was forced to retreat gradually to the areas of Savoy, Aosta Valley, and French-speaking Switzerland.

A number of further dialectometrical analyses have also proven the value of the synopsis of the skewness values for uncovering the multiple facets of language compromise, diachronically and synchronically. Skewness actually possesses a high diagnostic value, as it shows language structures which are much more deeply anchored in the data masses of a given linguistic atlas than the structures of the similarity maps.

4 Beyond the similarity matrix II: the dendrographic DM

See Maps 9 (tree) and 10 (spatial projection of the tree)

S-DM assumes that a given similarity matrix can be interpreted taxometrically in multiple ways. Nevertheless, the selection of the methods should always be connected to already existing problems of traditional geolinguistics, as this is the only way to continue, diversify and thereby enrich older qualitative issues by the support of quantitative methods.

For more than a hundred years, one of the questions of historical and areal linguistics concerns the genealogical classification of languages, which may be approached, in the field of geolinguistics, by the dendrographic analysis of dialects in a given area. The very ancient scheme of the family (or phylogenetic) tree is indeed of great classificatory relevance, in diachronic as well as in synchronic respects. With the help of taxometrical methods it is quite easy nowadays to generate a great number of phylogenetic trees on the basis of a given similarity matrix; these trees can directly (and immediately) be projected spatially. It must be emphasized that there are many tree generating algorithms, and that the dialectometrician or geolinguist must therefore choose carefully the appropriate algorithm. It would be absolutely wrong to presume that there is only one, or only one 'right' tree generating algorithm or classification tree.

For the correct understanding and interpretation of dendrographic DM, a good knowledge of the statistical processes involved in the tree generation is of utmost importance. The specific algorithm which we used in this instance belongs to the class of 'hierarchic-agglomerative procedures'. By a successive fusion or agglomeration of (respectively) two most similar elements of the

similarity matrix – beginning with the 641 'leaves' of the tree – a binary hierarchy of classes or clusters (or 'dendremes') is generated. In this process, the (bigger) dendremes located near the root of the tree have a greater inner (numerical) heterogeneity than the (smaller) dendremes near the leaves.

The tree in Map 9 was generated with the algorithm suggested by Joe Ward in 1963. The left border of the figure shows the 641 'leaves' of the tree (corresponding to the 641 sites of the ALF-grid), the right border the root (or the trunk) of the tree. The (always binary) structure of the tree shows a clearly visible hierarchical structure; inside, seven (A–G) big clusters (dendremes) may be isolated. The projection of these particular seven dendremes was realized for the purpose of demonstration only and allows us to see the major dialectal subdivisions of the Galloromania. With the means of VDM it is possible to draw not only the tree, but simultaneously a map visualizing the spatial correspondences (=choremes) of these seven dendremes: see Map 10.

The results show:

(a) the perfect spatial coherence of the seven choremes,
(b) their high geolinguistic plausibility, as the actual linguistic landscapes are well-known in traditional linguistic geography (e. g. choreme C: Picardy and the Wallonia, D: Francoprovençal, E: Gascoigne, F: Languedoc and Roussillon, etc.).

In principle, tree and map can always be interpreted synchronically and diachronically. The synchronic interpretation concentrates on the determination of dialectal landscapes of different size and on their reciprocal dependence, i.e. similarity. A diachronic interpretation simulates, as in a theoretical 'game', the progressive fragmentation of a given linguistic area, beginning at the first bifurcation after the root. These views are also shared by lexicostatistics. Nevertheless, they depend on the basic assumption that ca. 1900 years ago Galloromania represented a linguistically homogeneous area which diversified progressively over the time. Map 9 (tree) shows that the first

bifurcation separates the Domaine d'Oc (node 2: sum of the choremes/dendremes E–G) from the Francoprovençal and the Domaine d'Oïl (node 1: sum of the choremes/dendremes A–D). In the same way, the next fragmentation (nodes 3 and 4) separates the Francoprovençal (node 4: choreme/dendreme D) from the rest of the Domaine d'Oïl (choremes/dendremes A–C). This analysis can also be continued in the 'depth' of the tree according to the geolinguist's respective classification goals.

The VDM program provides a direct computation of different tree graphs,[23] as well as the colouring (either automatic or manual) of the diverse parts of the tree (=dendremes) and the corresponding spatialized equivalents (=choremes).

Thus, dendrographic dialectometry gives also access to structures which are very deeply hidden in the atlas data. Nevertheless, their correct interpretation and heuristic fruitfulness needs a good understanding of the statistical-mathematical bases.

5 Beyond the similarity matrix III: the correlative DM

The main objective of the implementation (realized in 2004) of correlative DM into the VDM-program was the cartographic visualization of the highly fluctuating correlation of spatially prominent (linguistic and extra-linguistic) variables (such as vocabulary, phonetics, vocalism, consonantism, geographical proximity, etc.).

This problem can be solved taxometrically by the correlation by pairs of the N vectors of two similarity matrices which must naturally have exactly the same dimensions (i.e. $N \times N$): see Figure 2. As a result, the correlation values are calculated and subsequently mapped in the usual way. The correlation values are computed with the well-known Bravais-Pearson correlation coefficient [r(BP)]. This coefficient ranges between -1 and $+1$ and is used 'in normal instances' for the measuring of the linear connection between two (standardized) numerical variables.

5.1 On correlation between 'language and space'

See Maps 13–15.

In 2004, we first mapped the spatial correlation between linguistic similarity and geographical proximity. Séguy (1971) had investigated this earlier from a 'linear' point of view drawing a straight line across a geolinguistic grid as had *inter alia* Heeringa and Nerbonne (2001). As in VDM the *x*- and *y*-coordinates of the 641 sites of the ALF grid were already stored, it was easy to calculate with the famous Pythagoras-formula the Euclidean distances between all the points. Subsequently, the distance values (*dist*) were converted with the formula $dist + prox = 100$ into proximities (*prox*). We used the geographic distances directly, without the (sublinear) transformations favored by Heeringa and Nerbonne (2001).

The visualization of these proximity values is carried out according to the usual standards of VDM: see Maps 11 and 12, which in formal respects are normal similarity maps and therefore in principle closely related to the Maps 2–6.

Map 11 (left) shows the *basilectal* management of the ALF grid from the perspective of the ALF-inquiry point 1, whereas Map 12 shows the *Euclidean* management of the same grid and from the same perspective. The two iconic profiles are optically quite different; the two (underlying) similarity distributions are correlated at $r = +0.433$.

This isolated correlation value gives yet no (geo)linguistic evidence; it is only convincing when the remaining 640 *r*(BP)-values are first calculated and then visualized: see Maps 13 and 14. In these maps we compare the correlation of linguistic and geographic proximity separately for each of the sites in our study.

The basilectal component of Map 13 relies upon our total ALF-corpus (with 1687 working maps belonging to all linguistic categories), whereas the basilectal component of Map 14, which is very similar, relies only on the phonetically relevant part (1117 working maps) of this total corpus.

Besides the fact that both maps have an almost 'beautiful' iconic profile (which never could have been expected before!), they open genuinely new perspectives in geolinguistics. Again, the history of Galloromania plays its part in the explanation of this very regular choropleth profile: as a matter of fact, the 'game' began in the South (ca. 121 BC) with the establishment of the Provincia Narbonensis, and in the North with Julius Cesar's famous military campaign (58–51 BC). *Ab initio*, Latin, which slowly expanded in Gallia, had two linguistic focuses: Narbo/Narbonne in the South and Lugdunum/Lyon in the North. After the fall of the (West) Roman Empire (476 AD) and the political ascent of the Franks after their Christianization (496 AD), the central area of the *Ile-de-France* (with Lutetia/Paris as new capital) took over the linguistic leading role from Lugdunum/Lyon, first in the North, and then in the whole area of France. And precisely this very old antagonism – with its conflicts and manifold contacts – between a constantly expanding North and a defensive South – is reflected in Maps 13 and 14.

The darker zones of Maps 13 and 14 (intervals 6 and 5) indicate those areas where the diffusion of the linguistic affinities still corresponded to the chances or 'expectancies' of Euclidean space. This 'primitive harmony' between language and space gradually 'phased out' with the expansion of the linguistic type of the *Langue d'Oïl* to the South of the Loire: see the polygons in the intervals 1 and 2. This imbalance occurred not only on the Southern border of the old kernel zone of the *Langue d'Oïl*, but also on the Northern border of the old nuclear zone of the *Domaine d'Oc*: actually, the constant capillary interaction of two antagonistic linguistic systems[24] suspended progressively the relations that had 'naturally' arisen between language and space.

The relevance of this 'dislocation theory' can be convincingly illustrated by the position of a Northern French linguistic island near Bordeaux (=ALF-point 635, Andraut, Département Gironde); it corresponds to one isolated polygon on the Maps 13 and 14. It figures in the interval 1 and is clearly in visible contrast with its

environment, where there are only polygons in the intervals 4 and 5.

This specific linguistic island (traditionally called *la Petite Gavacherie*) was founded in the second half of the 15th century in North-eastern Gascoigne by peasants that had migrated from the Saintonge. We know that in the instance of the foundation of a linguistic island, as it relies basically on the factor of migration, older relations between language and space which are considered to be 'natural' are radically suspended and expire. Moreover, on both maps the minimum of the two r(BP)-measurements is exactly at the ALF-point 635: Map 13: −0,15, Map 14: −0,25.

It has already been demonstrated in Map 4 that vocabulary (represented by 471 working maps of our total ALF-corpus) behaves quite similarly to phonetics. Now, Map 15 – which is quite similar to Maps 13 and 14 – shows that vocabulary also obeys the same space-language imperatives as phonetics, and to a large extent.

5.2 On the correlation between two linguistic categories

See Map 16.

The correlation of the (RIV) similarity maps for phonetics (1117 working maps) on the one hand, with vocabulary maps (471 working maps) on the other, shows an amazing and moreover aesthetically pleasing result: see Map 16. In the two landscapes of the intervals 6 and 5, phonetics and vocabulary are 'in tune' (in a primitive harmony) between linguistic and Euclidean proximity: a larger zone is located in the North and a smaller one in the South (first of all in the Languedoc). In between a sweeping transition area (intervals 1 and 2) extends from the Atlantic to the Alps, which has a somehow different shape than in the Maps 13–15. However, this transitional zone shows again the contacts and conflicts between the Langue d'Oïl and the Langue d'Oc.

As phonetics and vocabulary can be considered as two integral components of one and the same *parole* and *langue*, Map 16 can also be interpreted in the following way: if one mentally draws a straight line between Picardy in the North and the Roussillon in the South and tests the correlation values along this line, one first crosses areas in which phonetics and vocabulary are almost 'in phase' (or 'in tune'), then one comes into zones where these categories are gradually growing out of phase with one another, arriving finally in areas where the alternative phasing is re-established. These empirical results of variables moving in and out of phase within the same *parole* and/or *langue* should be of great interest and importance for the theoreticians of variationist grammar.

It has to be mentioned that by correlating other linguistic subcategories of our ALF corpus (such as vocalism with consonantism, stressed vocalism with unstressed vocalism, etc.) the respective iconic profiles are very similar to those in Map 16.

Obviously, it was not possible to describe in detail the whole range of possible applications open to correlative dialectometry; in this respect I only refer to the very interesting correlations between (geo)linguistic, (geo)genetic and patronymic data.[25]

6 Final Remarks

The Salzburg version of DM represents a heuristic instrument of explorative data analysis of universal applicability, on account of its manifold instruments of analysis developed on the job (ranging from the similarity maps to correlative DM). Its external use outside Salzburg and beyond Romance linguistics has been enabled since 2000 thanks to the availability of the software package VDM created by Edgar Haimerl.[26] The basic prerequisite for successful data exploration by means of DM and VDM are linguistic atlases of large size and high empirical quality. Data analyses can of course also be based upon diachronic data. Thus, data collections relying on medieval linguistic material (such as A. Dees' scripta atlas from 1980) can be interpreted successfully

concerning their deep structures with the assistance of VDM.[27]

Acknowledgements

References and abbreviations

AIS: Jaberg, K. and Jud, J. (eds), *Sprach- und Sachatlas Italiens und der Südschweiz*. Zofingen: Ringier, 1928–1940, 8 vol. (reprint: Nendeln: Kraus, 1971).

ALF: Gilliéron, J. and Edmont, E. (eds), *Atlas linguistique de la France*. Paris: Champion 1902–1910, vol. 10 (reprint: Bologna: Forni 1968).

Aurrekoetxea, G. (1992). Naffaroako Euskara: azterketa dialektometrikoa. Uztaro. 5: 59–109.

Bauer, R. (2003). *Dialektometrische Analyse des Sprachatlasses des Dolomitenladinischen and angrenzender Dialekte (ALD-I)*. Salzburg: Institut für Romanistik, vol. 2.

Bauer, R. (2004). Dialekte – Dialektmerkmale – dialektale Spannungen. Von 'Cliquen', 'Störenfrieden' und 'Sündenböcken' im Netz des dolomitenladinischen Sprachatlasses ALD-I. *Ladinia*, **28**: 201–242.

Besch, W. (1967). *Sprachlandschaften und Sprachausgleich im 15. Jahrhundert. Studien zur Erforschung der spätmittelhochdeutschen Schreibdialekte und zur Entstehung der neuhochdeutschen Schriftsprache*. München: Francke.

Bock, H.-H. (1974). *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Cluster-Analyse)*. Göttingen: Vandenhoeck & Ruprecht.

Chandon, J.-L. and Pinson, S. (1981). *Analyse typologique. Théories et applications*. Paris, New York, Barcelona, Milan: Masson.

Clua Julve, E. (1999). Distància lingüística i classificació de varietats dialectals. *Caplletra Revista internacional de filologia*, **26**: 11–26.

Dees, A. (1980). *Atlas des formes et des constructions des chartes françaises du 13e siècle*. Tübingen: Niemeyer.

Dickinson, G. C. (1973). *Statistical Mapping and the Presentation of Statistics*. London: Arnold.

DM: dialectometry.

Goebl, H. (1983a). 'Stammbaum' und 'Welle'. Vergleichende Betrachtungen aus numerisch-taxonomischer Sicht. *Zeitschrift für Sprachwissenschaft*, **2**(1): 3–44.

Goebl, H. (1983b). Parquet polygonal et treillis triangulaire. Les deux versants de la dialectométrie interponctuelle. *Revue de linguistique romane*, **47**: 353–412.

Goebl, H. (1984). *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen: Niemeyer, vol. 3.

Goebl, H. (1987). Points chauds de l'analyse dialectométrique: pondération et visualisation. *Revue de linguistique romane*, **51**: 63–118.

Goebl, H. (1993). Dialectometry. A Short Overview of the Principles and Practice of Quantitative Classification of Linguistic Atlas Data. In Köhler, R. and Rieger (eds), *Contributions to Quantitative Linguistics*. Dordrecht, Boston, London: Kluwer, pp. 277–315.

Goebl, H. (2000). La dialectométrisation de l'ALF: présentation des premiers résultats. *Linguistica*, **40**: 209–36.

Goebl, H. (2002). Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane*, **66**: 5–63.

Goebl, H. (2003). Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur. *Estudis Romànics*, **25**: 59–120.

Goebl, H. (2004). Sprache, Sprecher und Raum. Eine kurze Darstellung der Dialektometrie. Das Fallbeispiel Frankreich. *Mitteilungen der Österreichischen Geographischen Gesellschaft*, **146**: 247–286.

Goebl, H. (2005). Dialekte und Familiennamen in Frankreich. Ein interdisziplinärer Vergleich mit den Mitteln der Dialektometrie. In Hauska, G. (ed.), *Gene, Sprachen und ihre Evolution. Wie verwandt sind die Menschen – wie verwandt sind ihre Sprachen?* Regensburg: Regensburger Universitätsverlag, pp. 68–99.

Goebl, H. and Schiltz, G. (2001). Der 'Atlas des formes et des constructions des chartes françaises du 13e siècle' von Anthonij Dees (1980) – dialektometrisch betrachtet. In Gärtner, K., Holtus, G., Rapp, A., and Völker, H. (eds), *Skripta, Schreiblandschaften und Standardisierungstendenzen. Urkundensprachen im Grenzbereich von Germania und Romania im 13. und 14. Jahrhundert*. Beiträge zum Kolloquium vom 16. bis 18. September 1998 in Trier. Trier: Kliomedia, pp. 169–221.

Grassi, C. (2001). La geografia linguistica. In Holtus G., Metzeltin M., Schmitt, Chr. (eds), *Lexikon der romanistischen Linguistik*, Tübingen Niemeyer, vol. **I,1**: pp. 207–235.

Haimerl, E. (2006). Data Base Design and Technical Solution for the Management, Calculation, and Visualization of Dialect Mass Data. *Literary and Linguistic Computing*, **21**(4): 437–444.

Heeringa, W. (2004). *Measuring Dialect Pronunciation Differences using Levensthein Distance*. Groningen: Groningen Dissertations in Linguistics 46.

Heeringa, W. and Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, **13**: 375–400.

Heeringa, W., Nerbonne, J. and Kleiweg, P. (2002). Validating dialect comparison methods. In Gaul, W. and Ritter, G. (eds), *Classification, Automation, and New Media*. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation (University of Passau, 15-17/3/2002). Berlin, Heidelberg: Springer, pp. 445–452.

Hoppenbrouwers, C. and G. (2001). *De indeling van de Nederlandse streektalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM* [=featurefrequentiemethode]. Assen: van Gorcum.

Kretzschmar, W. A. (2003). Mapping Southern English. *American Speech*, **78**(2): 130–149.

Kretzschmar, W. A., Schneider, E. W., and Johnson, E. (eds) (1989). Computer methods in dialectology. *Journal of English Linguistics*, **22**(1): 1–170 (special issue).

LAMSAS: McDavid, R. I.Jr., & O'Cain, R., (eds) (1980). *Linguistic Atlas of the Middle and South Atlantic States*. Fasc. 1, 2. Chicago: University of Chicago Press.

Manni, F., Heeringa, W., and Nerbonne, J. (2006). To what extent are surnames words? Comparing geographic patterns of surname and dialect variation in the Netherlands. *Literary and Linguistic Computing*, **21**(4): 507–527.

Nerbonne, J. and Heeringa, W. (2001). Computational comparison and classification of dialects. *Dialectologia et Geolinguistica*, **9**: 69–83.

Nerbonne, J. and Kleiweg, P. (2006). Toward a dialectological yardstick. Accepted to appear in *Quantitative Linguistics*.

Nerbonne, J. and Kretzschmar, W. A. (eds) (2003). Computational Methods in Dialectometry. *Computers and the Humanities*, **37**(3): 245–372 (special issue).

Palsky, G. (1996). *Des chiffres et des cartes. Naissance et développement de la cartographie quantitative française au XIXe siècle*. Paris: CTHS.

RND: Blancquaert, E. and Pee, W. (eds): *Reeks Nederlandsche dialect atlassen*. Antwerpen: De Sikkel, 1926–1982, vol. 17.

Saramago, J. (1986). Différenciacion lexicale (un essai de dialectométrie, appliquée aux matériaux de l'A.L.E.). *Géolinguistique*, **2**: 1–31.

Scapoly, C., Goebl, H., Sobota, S., Mamolini, E., Rodriguez-Larralde, A., and Barrai, I. (2005). Surnames and dialects in France. Population structure and cultural evolution. *Journal of Theoretical Biology*, **237**: 75–86.

S-DM: the Salzburg school of dialectometry.

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane*, **35**: 335–357.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane*, **37**: 1–24.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. San Francisco: Freeman.

VDM: Visual DialectoMetry (dialectometrical software, created by E. Haimerl).

Viaplana, J. (1999). *Entre la dialectologia i la lingüística. La distància lingüística entre les varietats del català nord-occidental*. Barcelona: Publicacions de l'Abadia de Montserrat.

Ward, J. H.,Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**: 236–244.

Notes

1 We refer to the important centres of dialectometry in the Netherlands and the United States: cf. for instance the works of John Nerbonne (2001, with W. Heeringa, and 2003, with W. A. Kretzschmar), Wilbert Heeringa (2002, with J. Nerbonne and P. Kleiweg, and 2004) and the brothers Cor and Geer Hoppenbrouwers (2001) on the one hand, and those of William A. Kretzschmar (1989, with E. W. Schneider and E. Johnson, and 2003) on the other hand, realized with the data of the East American linguistic atlas LAMSAS.

2 For reasons of space the map types *honeycomb map* and *beam map* cannot be analysed here: for a detailed discussion, see Goebl, 1984, I:183–196, and 1983b: passim.

3 Note that the different targets or objectives have to be defined previously by geolinguists or linguists.

4 See also note 2.

5 It is our hope that this research interest should create a strong link between areal, historical and general linguistics through its insistence on an empirical basis in linguistic behaviour.

6 Between 1973 and 1982 I worked mainly at Regensburg University (Bavaria, Germany).

7 Cf. Clue Julve, 1999: passim, and Viaplana, 1999.

8 Cf. Aurrekoetxea, 1992: passim.

9 Cf. Saramago, 1986: passim.

10 The ALF-map 14 *aiguille* (<lat. *ACÚCLA) has been analyzed five times and thus resulted in five working maps: 1: referring to the phonetic results of the pretonic A-, 2: referring to the phonetic results of the intervocalic -C-, 3: referring to the phonetic results of the stressed Ú, 4: referring to the phonetic results of the intervocalic nexus CL, 5: referring to the phonetic results of the final -A.

11 See the respective curves in Goebl, 2002 (13–15) and 2003 (68, 70–71).

12 Note that each atlas site is compared to each of the remaining $N-1$ sites. The symbol j is used for the reference site, and the symbol k for the atlas site that j is compared to.

13 I.e. the reference site with the symbol j.

14 This value corresponds to the lower threshold of the interval 1: see the numerical legend of Map 2.

15 This value corresponds to the upper threshold of the interval 6: see the numerical legend of Map 2.

16 This value corresponds to the upper threshold of the interval 3: see the numerical legend of Map 2.

17 The acronym MINMWMAX refers to three German statistical concepts: MIN (=*Min*imum), MW (=*M*ittel*w*ert, arithmetic mean), MAX (=*Max*imum). The use of well defined statistical algorithms for cartographic aims has a long tradition which goes back to the 19th century: see Dickinson, 1973: passim, and Palsky, 1996: passim.

18 The use of 6 (coloured) intervals is recommended by cartographers and geographers for psychological and optical reasons. Nevertheless our dialectometric software VDM allows the current (and rapid) generation of ten different types of mappings (based on 2 up to 20 intervals) using three different algorithms of visualization and a completely free colour spectrum.

19 See Goebl, 1983a: passim; cf. also Goebl, 1984, I: 83–86 and 1987: passim.

20 See the respective (coloured) visualizations in Goebl, 2003: 110–111.

21 Cf. the seminal book of W. Besch *Sprachlandschaften und Sprachausgleich* of 1967.

22 See the respective coloured figures in Goebl, 2002: 54–59 and 2003: 112–155.

23 Implemented hierarchic-agglomerative procedures: besides Ward also Complete Linkage, Single Linkage,

Simple Average Linkage, Average Linkage-UPGMA, and the Centroid Method.

24 In this 'game' Northern French (Langue or Domaine d'Oïl) took over the active part and expanded, whereas Occitanian (Langue or Domaine d'Oc) had the passive role and retreated.

25 Cf. Goebl, 2005 and Scapoli *et al.*, 2005. See also Manni *et al.*, *Literary and Linguistic Computing*, **21**(4): 507–527.
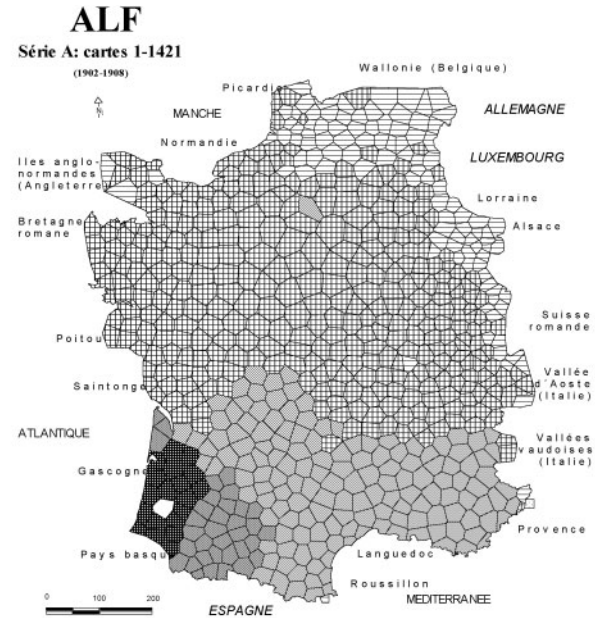
26 See Haimerl's paper, *Literary and Linguistic Computing*, **21**(4): 437–444.

27 Cf. Goebl/Schiltz, 2001: passim.

ALF
Série A: cartes 1-1421
(1902-1908)
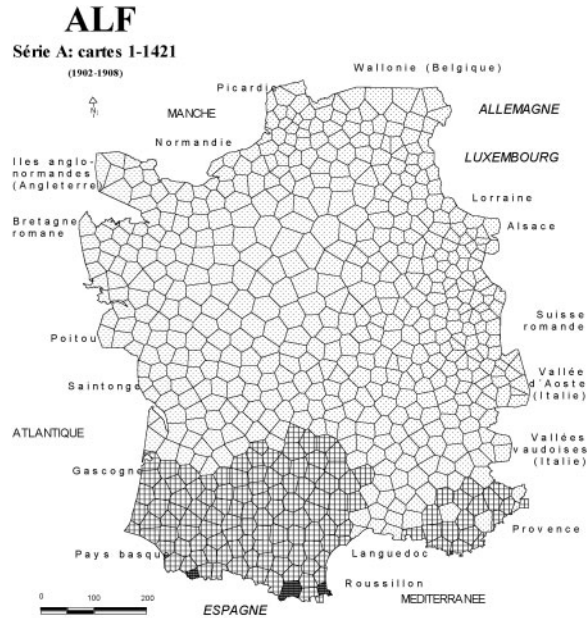
|		
1	acheter	(514)
2	croumpa	(124)
3	ana croumpa	(3)
		$\Sigma = 641$

Map 1: Sample of a lexical 'working map' showing the spatial distribution of the Gallo-Romance designations of 'to buy' (following ALF 6 *acheter*).

ALF
Série A: cartes 1-1421
(1902-1908)

Visualization
MINMWMAX 6-tuple

1	37.62	- 43.11	(1)
2		- 48.60	(102)
3		- 54.09	(336)
4		- 64.81	(148)
5		- 75.53	(34)
6		- 86.25	(19)
			$\Sigma = 640$

Similarity distribution
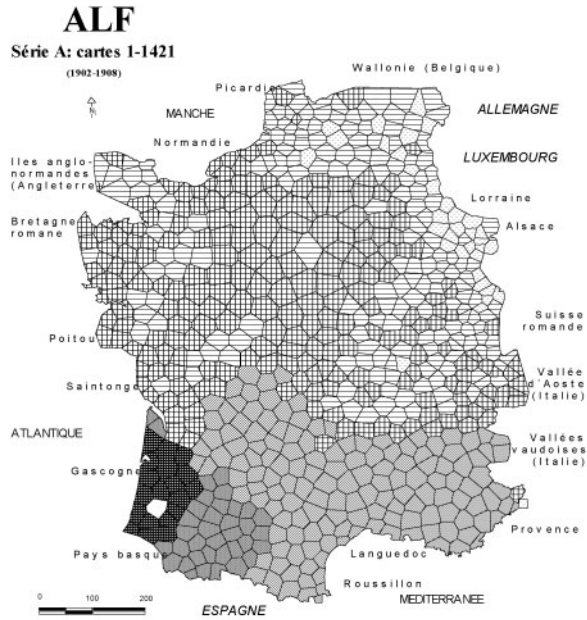MINMWMAX 12-tuple

$RIV_{682,k}(TOT)$
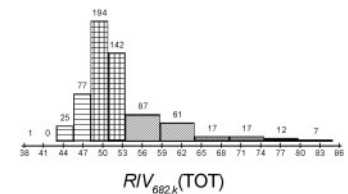
Map 2: A typical Gascon similarity profile: similarity map to the ALF-point 682 (Tartas, Département Landes).
Similarity index: $RIV_{682,k}$
Corpus: 1687 working maps (total corpus)
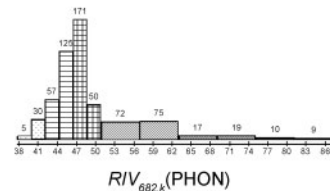Algorithm of visualization: MINMWMAX (6-tuple).

ALF
Série A: cartes 1–1421
(1902–1908)

Wallonie (Belgique)

Picardie

MANCHE

ALLEMAGNE

LUXEMBOURG

Normandie

Iles anglo-
normandes
(Angleterre)

Lorraine

Bretagne
romane

Alsace

Suisse
romande

Poitou

Vallée
d'Aoste
(Italie)

Saintonge

ATLANTIQUE

Vallées
vaudoises
(Italie)

Gascogne

Provence

Pays basque

Languedoc

Roussillon

MEDITERRANEE

ESPAGNE

0   100   200

Visualization
MINMWMAX 6-tuple

1	37.65	– 42.08	(35)
2		– 46.50	(182)
3		– 50.92	(221)
4		– 63.21	(147)
5		– 75.50	(36)
6		– 87.79	(19)

$\Sigma = 640$

Similarity distribution
MINMWMAX 12-tuple

171
120
57   50
30   72   75
5   17   19   10   9
38 41 44 47 50 53 56 59 62 65 68 71 74 77 80 83 86

$RIV_{682,k}(PHON)$

Map 3: A typical Gascon similarity profile: similarity map to the ALF-
point 682 (Tartas, Département Landes).
Similarity index: $RIV_{682,k}$
Corpus: 1117 working maps (phonetics)
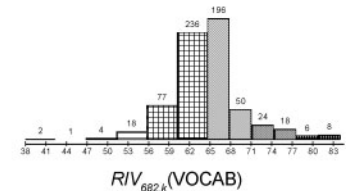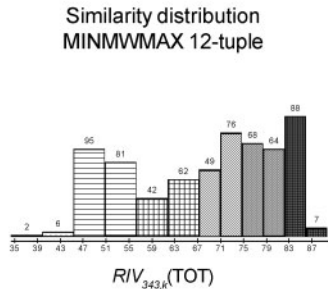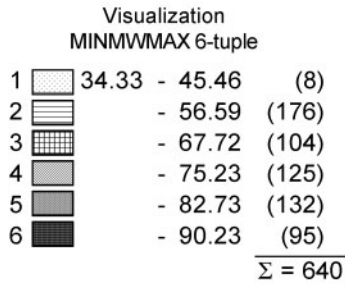Algorithm of visualization: MINMWMAX (6-tuple).

ALF
Série A: cartes 1–1421
(1902–1908)

Wallonie (Belgique)

Picardie

MANCHE

ALLEMAGNE

LUXEMBOURG

Normandie

Iles anglo-
normandes
(Angleterre)

Lorraine

Bretagne
romane

Alsace

Suisse
romande

Poitou

Vallée
d'Aoste
(Italie)

Saintonge

ATLANTIQUE

Vallées
vaudoises
(Italie)

Gascogne

Provence

Pays basque

Languedoc

Roussillon

MEDITERRANEE

ESPAGNE

0   100   200

Visualization
MINMWMAX 6-tuple

1	37.92	– 46.84	(3)
2		– 55.76	(22)
3		– 64.69	(313)
4		– 71.28	(246)
5		– 77.86	(42)
6		– 84.45	(14)

$\Sigma = 640$

Similarity distribution
MINMWMAX 12-tuple

196
236
77
50
2 1   4   18   24 18
38 41 44 47 50 53 56 59 62 65 68 71 74 77 80 83

$RIV_{682,k}(VOCAB)$

Map 4: A typical Gascon similarity profile: similarity map to the ALF-
point 682 (Tartas, Département Landes).
Similarity index: $RIV_{682,k}$
Corpus: 471 working maps (vocabulary)
Algorithm of visualization: MINMWMAX (6-tuple).
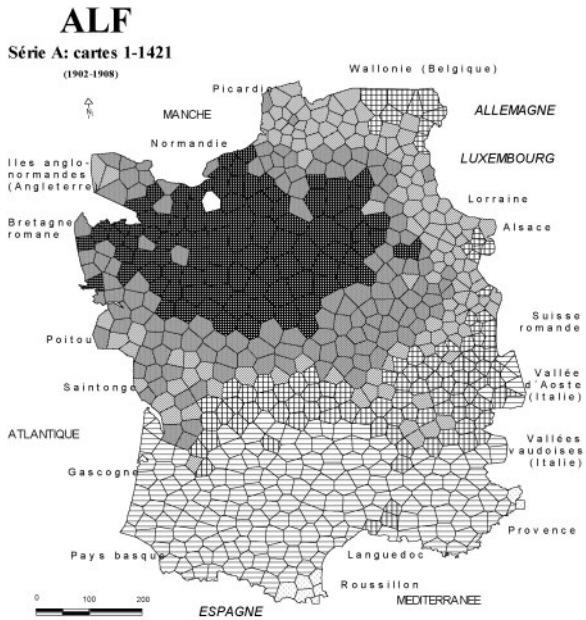
Map 5: A typical similarity profile of the northern Domaine d'Oïl: (normal) similarity map to the ALF-point 343 (La Chapelle-Yvon, Département Calvados).
Similarity index: $RIV_{343,k}$
Corpus: 1687 working maps (total corpus)
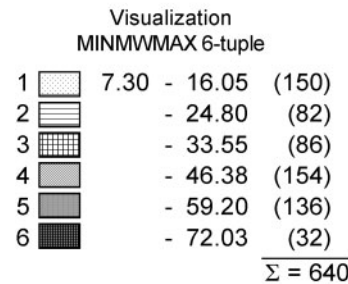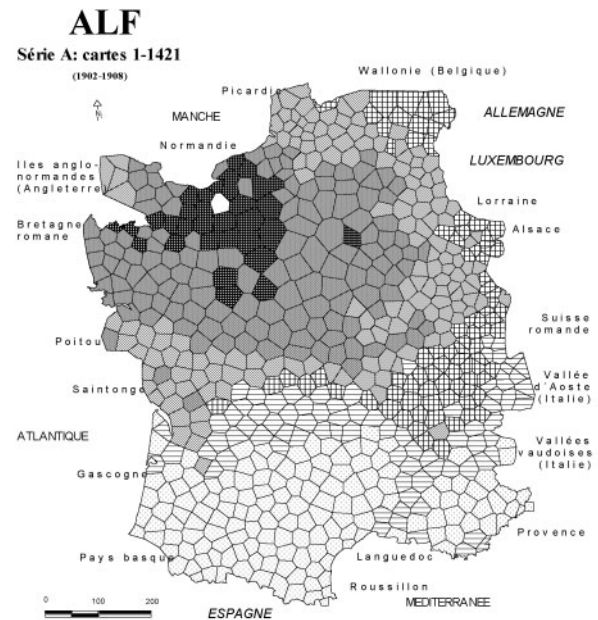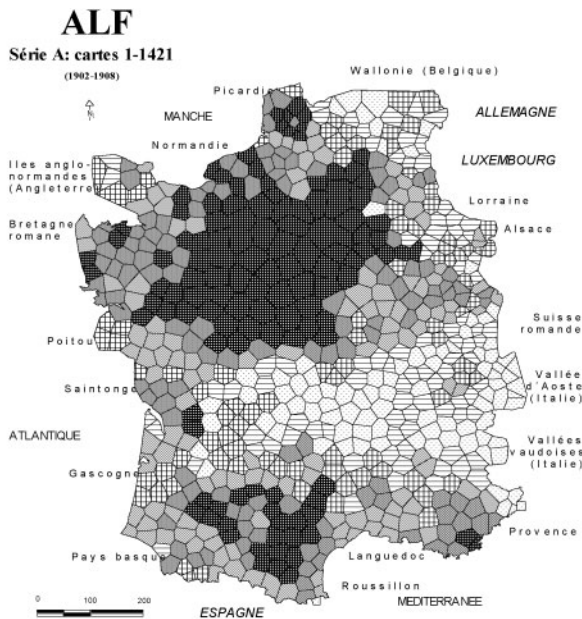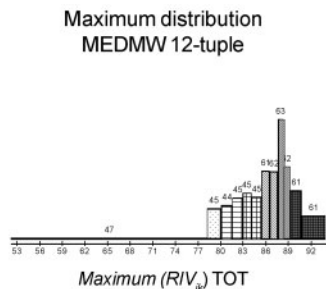Algorithm of visualization: MINMWMAX (6-tuple).
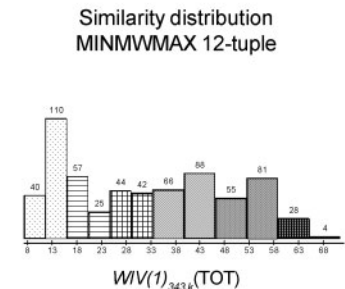
Map 6: A typical similarity profile of the northern Domaine d'Oïl: (weighted) similarity map to the ALF-point 343 (La Chapelle-(Yvon, Département Calvados).
Similarity Index: $WIV(1)_{343,k}$
Corpus: 1687 working maps (total corpus)
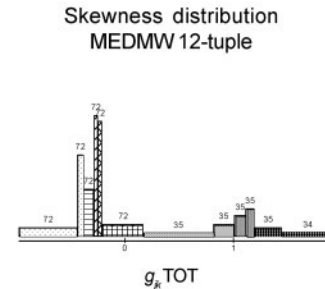Algorithm of visualization: MINMWMAX (6-tuple).

ALF
Série A: cartes 1-1421
(1902-1908)

ALF
Série A: cartes 1-1421
(1902-1908)

Visualization
MEDMW 6-tuple

1	52.17	-	80.12	(92)
2		-	83.12	(89)
3		-	85.71	(90)
4		-	88.02	(123)
5		-	89.77	(125)
6		-	94.60	(122)

$\Sigma = 641$

Maximum distribution
MEDMW 12-tuple

$Maximum\ (RIV_{jk})$ TOT

Map 7: Choropleth map of the synopsis of the maximal values of 641
similarity distributions.
Similarity index: RIV_{jk}
Corpus: 1687 working maps (total corpus)
Algorithm of visualization: MEDMW (6-tuple).

Visualization
MEDMW 6-tuple

1	-0.97	-	-0.37	(144)
2		-	-0.23	(144)
3		-	0.20	(144)
4		-	1.03	(70)
5		-	1.23	(70)
6		-	1.96	(69)

$\Sigma = 641$
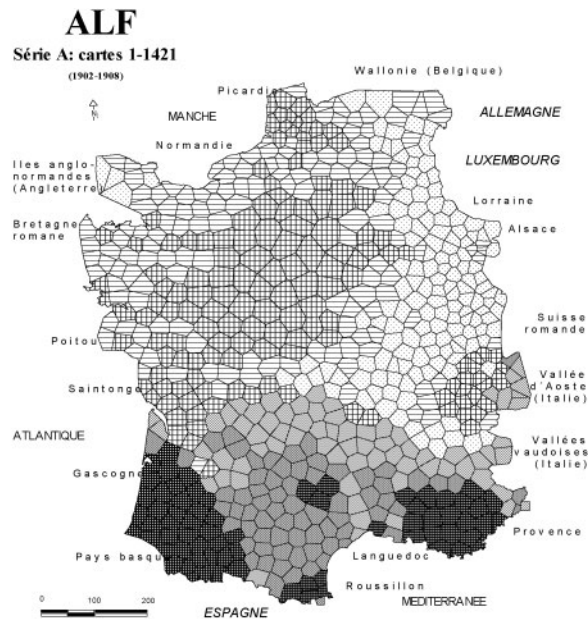
Skewness distribution
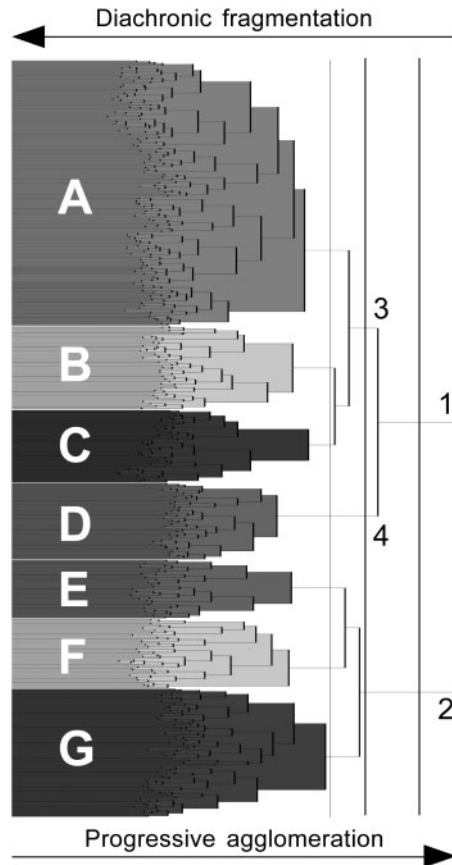MEDMW 12-tuple

g_{jk} TOT

Map 8: Choropleth map of the synopsis of the skewness values of 641
similarity distributions.
Similarity Index: RIV_{jk}
Corpus: 1687 working maps (total corpus)
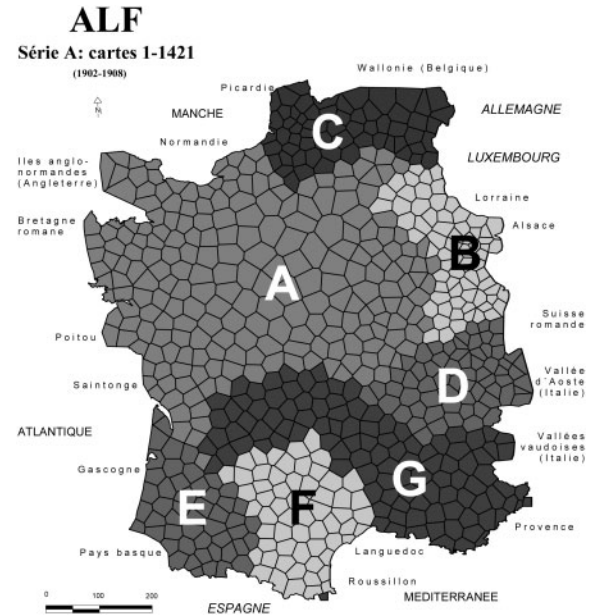Algorithm of visualization: MEDMW (6-tuple).

Map 9: Dendrographic classification of 641 dialectological objects
(ALF-points).
Similarity index: RIV_{jk}
Dendrographic algorithm: hierarchical grouping method of Ward
Number of dendremes (A-G): 7
1, 2: the two first ramifications (Domaine d 'Oïl versus
Domaine d'Oc)
3, 4: two subsequent subgroupings of the Domaine d 'Oïl.

Map 10: Spatial conversion of the tree of the map 9.
Number of choremes (A-G) [corresponding to the dendremes
of the tree of the map 9]: 7.

ALF
Série A: cartes 1-1421
(1902-1908)

Wallonie (Belgique)
Picardie
MANCHE
ALLEMAGNE
Normandie
LUXEMBOURG
Iles anglo-
normandes
(Angleterre)
Lorraine
Bretagne
romane
Alsace
Poitou
Suisse
romande
Saintonge
Vallée
d'Aoste
(Italie)
ATLANTIQUE
Vallées
vaudoises
(Italie)
Gascogne
Provence
Pays basque
Languedoc
Roussillon
MEDITERRANEE
ESPAGNE

0   100   200

Visualization
MINMWMAX 6-tuple

Similarity distribution
MINMWMAX 12-tuple

1	34.61 - 45.76	(8)
2	- 56.92	(169)
3	- 68.07	(115)
4	- 75.46	(115)
5	- 82.84	(141)
6	- 90.22	(92)

Σ = 640

$RIV_{1,k}$(TOT)

Map 11: A typical similarity profile of the southern Domaine d'Oïl: similarity map to the ALF-point 1 (Marcigny, Département Nièvre).
Similarity index: $RIV_{1,k}$
Corpus: 1687 working maps (total corpus)
Algorithm of visualization: MINMWMAX (6-tuple).

ALF
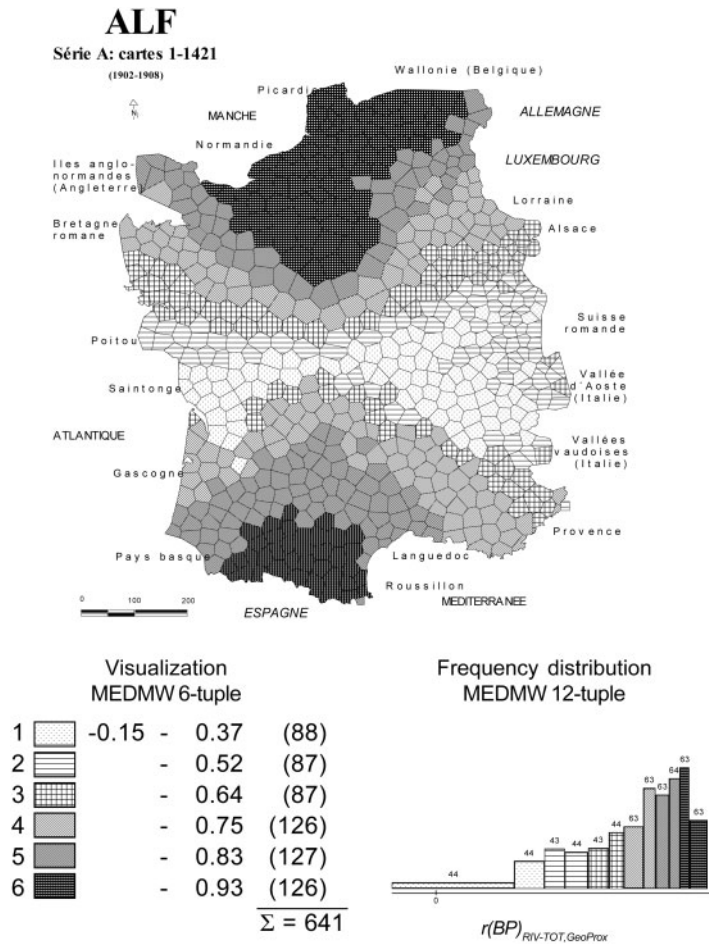Série A: cartes 1-1421
(1902-1908)

Wallonie (Belgique)
Picardie
MANCHE
ALLEMAGNE
Normandie
LUXEMBOURG
Iles anglo-
normandes
(Angleterre)
Lorraine
Bretagne
romane
Alsace
Poitou
Suisse
romande
Saintonge
Vallée
d'Aoste
(Italie)
ATLANTIQUE
Vallées
vaudoises
(Italie)
Gascogne
Provence
Pays basque
Languedoc
Roussillon
MEDITERRANEE
ESPAGNE

0   100   200

Visualization
MINMWMAX 6-tuple

Similarity distribution
MINMWMAX 12-tuple

| 1 | 47.85 - 55.15 | (58) |
2	- 62.46	(110)
3	- 69.76	(175)
4	- 78.80	(164)
5	- 87.84	(91)
6	- 96.88	(42)

Σ = 640

$GeoProx_{1,k}$(TOT)

Map 12: A proximity profile of the southern Domaine d'Oïl: proximity map to the ALF-point 1 (Marcigny, Département Nièvre).
Proximity index: Euclidean proximity
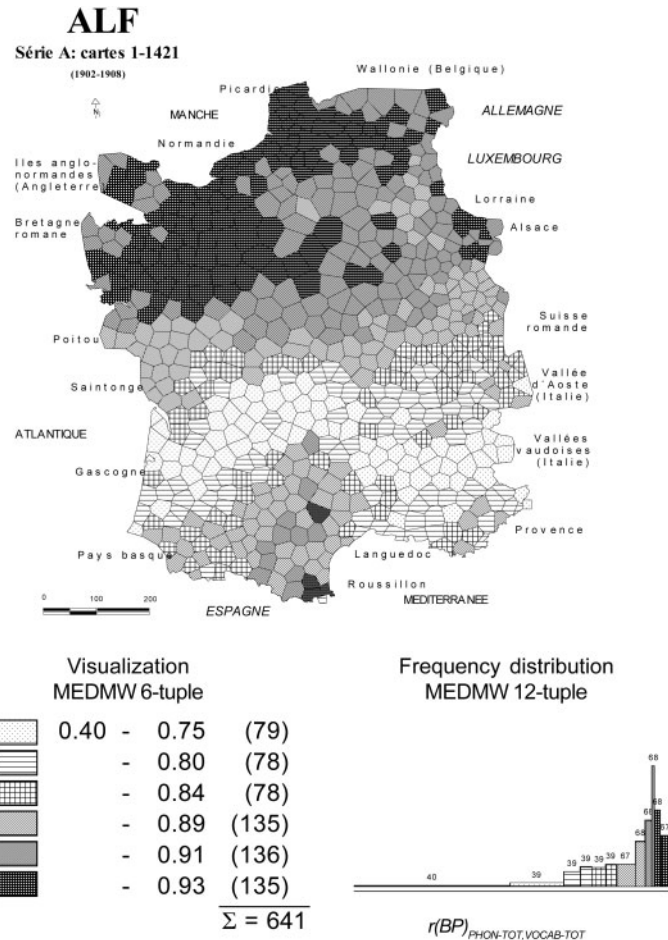Algorithm of visualization: MINMWMAX (6-tuple).

ALF
Série A: cartes 1-1421
(1902-1908)

Visualization
MEDMW 6-tuple

1		-0.15 - 0.37	(88)
2		- 0.52	(87)
3		- 0.64	(87)
4		- 0.75	(126)
5		- 0.83	(127)
6		- 0.93	(126)
			$\Sigma = 641$

Frequency distribution
MEDMW 12-tuple

$r(BP)_{RIV-TOT, GeoProx}$

ALF
Série A: cartes 1-1421
(1902-1908)

Visualization
MEDMW 6-tuple

1		-0.25 - 0.34	(88)
2		- 0.48	(88)
3		- 0.61	(88)
4		- 0.73	(125)
5		- 0.82	(126)
6		- 0.91	(126)
			$\Sigma = 641$

Frequency distribution
MEDMW 12-tuple

$r(BP)_{PHON-TOT, GeoProx}$

Map 13: Choropleth map of the correlation values [according to r(BP)] between 641 similarity values (according to RIV_{jk}) and 641 proximity values (according to the Euclidean proximity). Corpus (of the similarity measurement): 1687 working maps (total corpus) Algorithm of visualization: MEDMW (6-tuple).
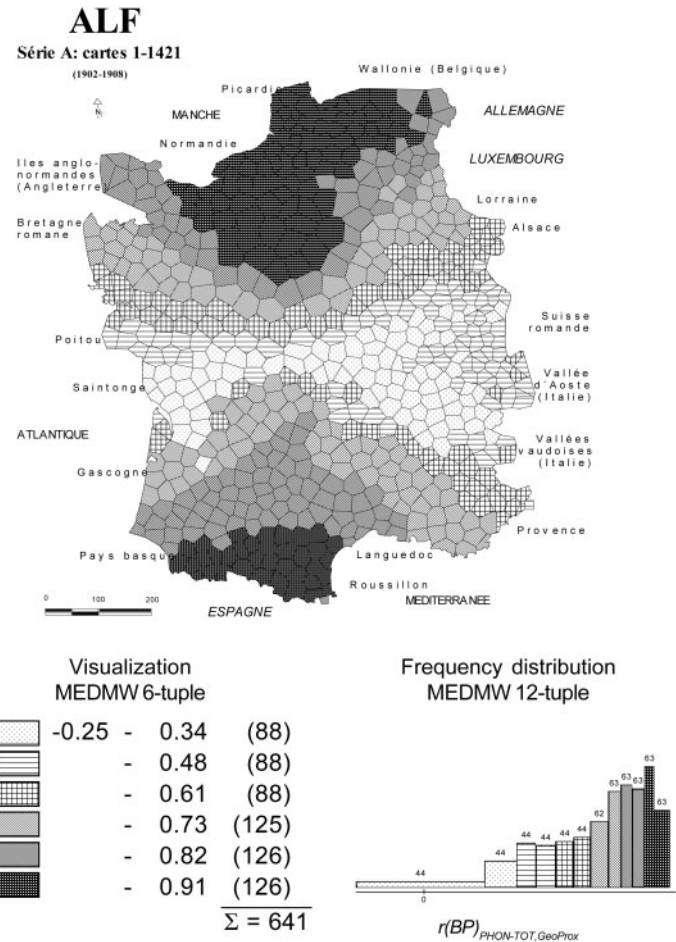
Map 14: Choropleth map of the correlation values [according to r(BP)] between 641 similarity values (according to RIV_{jk}) and 641 proximity values (according to the Euclidean proximity). Corpus (of the similarity measurement): 1117 working maps (phonetics) Algorithm of visualization: MEDMW (6-tuple).
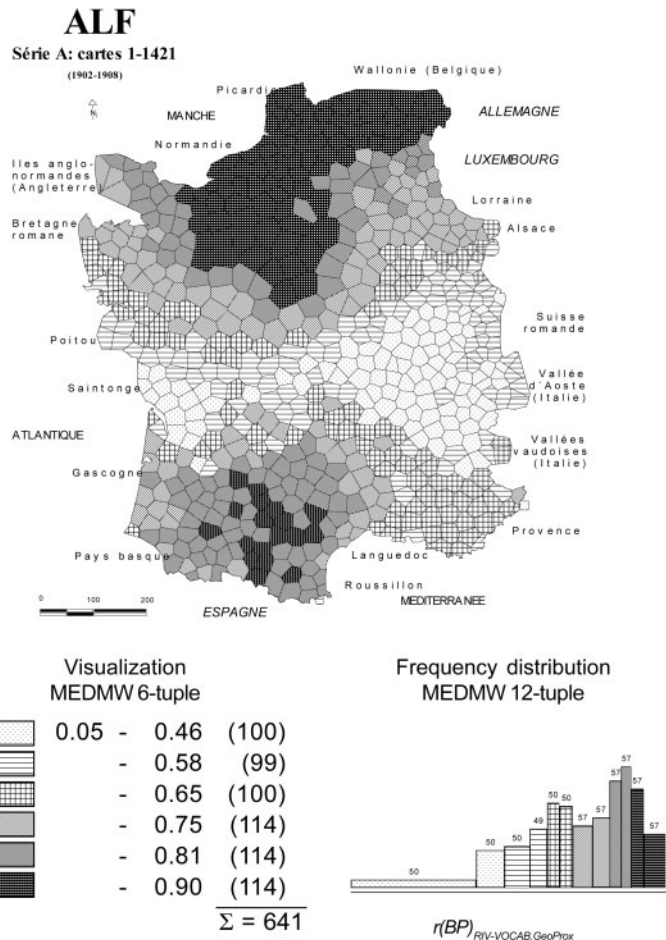
Map 15: Choropleth map of the correlation values [according to r(BP)] between 641 similarity values (according to RIV$_{jk}$) and 641 proximity values (according to the Euclidean proximity). Corpus (of the similarity measurement): 471 working maps (vocabulary)
Algorithm of visualization: MEDMW (6-tuple).

Map 16: Choropleth map of the correlation values [according to r(BP)] between two pairs of 641 similarity values (according to RIV$_{jk}$): phonetics (1117 working maps) and vocabulary (471 working maps)
Algorithm of visualization: MEDMW (6-tuple).

Recent Advances in Salzburg Dialectometry