Dialectologia. **Special issue, I** *(2010), 63-77.*
ISSN: 2013-22477

DIALECTOMETRY: THEORETICAL PREREQUISITES, PRACTICAL PROBLEMS, AND CONCRETE APPLICATIONS (MAINLY WITH EXAMPLES DRAWN FROM THE "ATLAS LINGUISTIQUE DE LA FRANCE", 1902-1910)

Hans GOEBL
Salzburg University
Hans.Goebl@sbg.ac.at

Abstract

This paper documents the many taxometric and cartographic achievements of the Salzburg school of dialectometry. It discusses the following topics: 1) problems of measurement of linguistic atlas data (with particular emphasis on Romance linguistic atlases), 2) establishment of a data matrix, 3) choice of a 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 and dendrograms (and their spatial projection). The ultimate purpose of these highly sophisticated cartographic techniques (choropleth and isarithmic 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 Romance linguistic geography ("géographie linguistique"), 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.

Key words
dialectometry, *Atlas Linguistique de la France*, geolinguistics.

1. What is Dialectometry (DM)?

The primary objective of DM is the synthetic quantitative exploration and interpretation of linguistic atlas data or similar data collections. Actually, DM is based on one crucial theoretical hypothesis assuming that HOMO LOQUENS – who through all ages has also been a HOMO SPATIALIS settling all over the globe – manages this space he lives in through his speech and language(s) in a specific way. As all dialects (or languages) used by communities of speakers occur in some kind of spatial distribution, the question of the linguistic management of space by HOMO LOQUENS is indeed a universal one. It is also of genuine interdisciplinary relevance, as other human sciences (anthropology, ethnography, population genetics, etc.) ask analogous questions concerning spatial management. Within the scope of philology, the results elaborated by dialectometry are also of diachronic relevance. In addition, a great number of synchronically important results concerning the various aspects of communication and interaction in space can be found.

Methodologically, DM is characterized by the following formula:

DM = geolinguistics + numerical taxometry (or classification). The statistical procedures used by DM for quantitative data synthesis are also commonly used in the natural, social and human sciences. Nowadays, they are subsumed under the generic term of "data mining". DM intends (with relatively simple statistic means) to uncover in linguistic atlas data lower and higher ranking structural patterns which had hitherto been hidden to the observer at first glance. After the statistical analysis, the patterns and structures are displayed on maps and discussed dialectometrically.

For a rapid execution of the numerical and classificatory computations and their corresponding mappings or visualization, a computer program called "Visual DialectoMetry" (VDM) was created in 2000 by my senior research assistant, Edgar HAIMERL. All the maps in this contribution were generated with VDM and reset in their final cartographic form by my present collaborator, Slawomir SOBOTA.[1] As a consequence, the Salzburg variety of DM focuses on a great number of (mostly colourful) maps of different types.

[1] My warmest regards and thanks to these two gentlemen for their remarkable and motivated assistance.

2. A quick overview of the methodological repository of Salzburg DM

In what follows, the essential methodological advances of Salzburg DM will be presented and illustrated. A DM project always starts with the selection of an appropriate linguistic atlas and the adequate processing of its data. This stage is still executed manually, according to the well-established methodological principles of Romance linguistics (concerning phonetics, morphology and the lexicon). The output is a great number (p) of "working maps" (with N inquiry points, or 'sites'): see Map 1. The content of these p "working maps" is condensed in a data matrix with the dimensions N x p. Only after this process can the data (or information) processing of DM with its quantitative measures begin.

First, the similarities between (pairs taken from) the N vectors have to be measured. As the dialectometrician can in principle use different similarity measures, he has to select the appropriate index, keeping in mind his theoretical hypotheses on dialectal similarity. In this particular instance, the "Relative Identity Value" (RIV_{jk}) has proved to be very useful. It measures the percentage of pairwise matchings between the discrete nominal (or qualitative) types of those linguistic features that are registered in the data matrix. A squared similarity matrix (with the dimensions N x N) contains all these measurements and gives the total number of all the variations in quantitative form. Thus, the relevant geolinguistic information is transferred from the qualitative to the quantitative level. Experience has shown that many linguists have difficulty understanding this transition, although it is commonplace in natural, human and biological sciences.

By a simple transformation (similarity values + distances values = 100) an appropriate *distance* matrix can be calculated from the *similarity* matrix. Finally, the measurement values stored in the two matrices have to be interpreted (according to the objectives of dialectology/geolinguistics) as part or as a whole. In this process, a number of maps were established according to specified aims, each of them corresponding to different segments of the two matrices. Therefore, the inner logic of DM imperatively requires the parallel availability of many analyses with their resulting mappings.

2.1. From the linguistic atlas (ALF) to the data matrix

In the years between 1997 and 2000, 626 original maps of the ALF were analyzed ("taxated") in the Salzburg research project. Its objective was the taxating of the phonetic, morphological and lexical variation on these ALF maps. A data matrix was created with 1,687 "working maps" and 641 sites. It is easy to explain this number of 641 (instead of 638): three artificial sites (or "false dialects") corresponding to the standard forms of French, Italian and Catalan were added to the 638 original sites of the ALF.

2.2. The geometrical adjustment of the linguistic atlas grids

The results of our DM analyses were printed out on two types of maps: 1) choropleth maps as a puzzle of different *spatial* signatures and 2) isogloss or beam maps (called also isarithmic maps), relying on a network of *linear* signatures. Both types can only be adjusted in accordance with the principles of Delaunay triangulation and Voronoi polygonisation. These two procedures are cartographic essentials and correspond to normal international standards.

2.3. From the data to the similarity or distance matrix

The "Relative Identity Value" (RIV_{jk}) is usually applied for the measurement of the similarities between (a pair of) sites which relies on the percentage of the similarities of nominal distinctive features. We also apply the "Weighted Identity Value (with the weight 1)" [$WIV(1)_{jk}$] to underscore rare features rather than those linguistic types (taxates) which are found everywhere. This view corresponds to the ideas of many competent linguists who think that rare and therefore "more important" language features should be privileged over frequent ones, as they might be considered "trivial".

2.4. On the organization of the choropleth and isarithmic (isogloss and beam) maps

In the process of the visualization of the measurement values derived from the similarity or distance matrix, a given numerical variation has to be transformed

algorithmically to a well-shaped iconic variation (here: of colours, otherwise, of shadings and hatchings). As this problem has been discussed over the last 200 years, there is a great deal of previous experience. We know that the human capacities of pattern recognition work best with 6-8 colour steps, and most maps are established in agreement with this circumstance. All maps are centred according to the arithmetic mean: the *cold* coloured polygons correspond to measurement values below the arithmetic mean, and the *warm* coloured polygons to the measurement values above the arithmetic mean.

The bottom left of each map displays a numerical legend and the bottom right the respective histogram (always with N or N-1 measurement values). Such histograms refer therefore to the quantitative nature of the respective frequencies and are of statistical interest. The number of the histogram bars is always the double number of the respective colour steps. The curves above each histogram represent the Gauss distribution (i.e. a normal distribution), calculated by means of two statistical index values (the arithmetic mean and the standard deviation) of the relative frequency distribution. It is also of statistical relevance, mainly for comparison with the histogram's shape.

3. Introduction to the similarity maps: see Map 2.

The heuristic instrument of the similarity map represents the typo-diagnostic basis of Salzburg DM. Formally, each similarity map relies on one of the N vectors of the similarity matrix. One of the N measurements values is always 100.

The similarity map gives all the information concerning the spatial stratification of the dialectal similarities of respective N-1 measurement values of the linguistic atlas being investigated (symbolized by: *k*) with regard to an already selected site (symbolized by *j*). In the case of Map 2 this is the locality la Chapelle-Yvon, Département Calvados (= ALF-P. 343). Actually, all similarity maps show a more or less harmoniously structured iconic profile, reflecting the decrease of dialectal similarity (in general) with the increasing distance to the reference point. Nevertheless, each dialectal landscape has its own characteristics. Thus, Map 2 shows a "typical" Norman iconic profile.

Linguistically, each similarity map contains information on the geo*linguistic* modalities regarding the *position of a certain dialect within the whole inquiry grid*. Questions about the position of a dialect regarding its environment have often been investigated over the last 150 years in Romance and German linguistics. With DM, precise answers can be given, which respond, moreover, to the underlying spatial metaphor of such questions.

An analogous non-linguistic interpretation may help the observer understand the deeper meaning of the message conveyed by the similarity maps. The following examples show obvious analogies:

1) telephone analogy :

Mathematical logic, which is at the basis of the similarity maps, allows every one of them to be considered as a general record of the phone calls (active and passive) of one of the N participants within a telephone network. Therefore, Map 2 indicates, where individual participants phoned very often, seldom or hardly ever.

2) missionary analogy :

Under the assumption that each reference point can be compared with a single "missionary" eager to spread his ideology by diffusion, the similarity maps show where and to what extent he succeeded.

It is important to realize that the iconic message of a similarity map is *quantitative*. The six colour steps which visualize the spatial decrease of the RI-values reflect a (coarsely structured) spatial pseudo-continuum. This implies that the boundaries between the different colour steps do not have the character of language boundaries. With VDM software, we are now able to visualize each similarity map in ten different forms (from 2 to 20 colour steps). Furthermore, we can insert three iconic algorithms (MINMWMAX, MEDMW and MED), in order to comply with the diagnostic demands of the observer.

4. The honey comb map ad the beam map: see Maps 3 and 4.

The honey comb map corresponds methodologically and cartographically to the traditional isogloss synopsis and is widely used in modern language departments. The beam map, which represents a new geolinguistic instrument, is (in cartographical, taxometrical and linguistic respects) the logical reversion of the honeycomb map. In the following table, compare the basic methodological principles:

	beam map	**honeycomb map**
geometrical preparation of the basic grid	triangulation	polygonization
visualized variable	similarity values (calculated with RIV_{jk})	distance values (calculated with RDV_{jk})
linguistic meaning	evidence of interpunctual contact ("friendship")	evidence of interpunctual distances ("conflict")

Note also that these two types of maps give only information about existing similarities or distances *between neighbouring* inquiry points, hence the generic term "*inter*point maps". For maps 3 and 4, the iconic syntax relies on 1,791 linear signatures (= polygon or triangle sides), which can vary according to thickness or darkness. In both Maps the following principle applies: "the thicker, the bluer (Map 3)/redder (Map 4)". Metaphorically, the blue polygon sides symbolize conflict, whereas the red triangle sides illustrate contact.

A look at the honeycomb map (Map 3) gives evidence of clear boundary phenomena that structure the ALF space, but also of a lack of continuous lines, or – in the traditional sense – "border lines". One actually sees more or less strong honeycomb phenomena: thus, on Map 3, the North-South division can be detected as well as the special position of Francoprovençal (in the East) and of Catalan in the Roussillon (in the South).

In contrast, the amassment of thick red triangle sides on Map 4 represents compact interaction spaces: this concerns the central area of the Domaine d'Oïl in the North and some less prominent areas in the South (the Languedoc, the Provence). One also recognizes a clearly distinguishable transitional zone between the North and the South,

indicating that between these ALF inquiry points there are far fewer affinities than in the core zones of the North and the South.

If the honeycomb maps and the beam maps are highly suggestive and self explanatory, they are nonetheless with respect to taxometrics superficial. Consider that with the well-known combinatory formula N/2 (N-1), 205,120 measurement values are found in the similarity or distance matrix (with the dimensions 641 x 641) upon which Maps 5 and 6 are based. Of all these values, only 1,791 (i.e. .87%) are recorded or visualized, whereas in the establishment of Maps 5 and 6, 100% of the measurement values were recorded. This comparison gives rise to two observations: 1) general: in principle, taxometrical analyses can reveal more or less deeply embedded structures, 2) in this special instance: the interpoint analysis remains rather superficial. It can therefore be considered a useful instrument of initial geolinguistic classification, but is not the appropriate heuristic means for deeper structural analyses of dialect networks.

5. The parameter maps: the example of the synopsis of skewness values: see Map 5

It has been suggested that the histograms of the similarity maps of a similarity matrix refer to the *statistical* nature of the underlying similarity distributions. A comparison of many histograms reveals that their shape varies from any large dialect area to any other large dialectal area. It is therefore logical to set up a synopsis of these "characteristic parameters" of a similarity distribution (such as minimum, arithmetic mean, maximum, medium or standard deviation) as well, and to apply it to the recognition of geolinguistic patterns. A linguistically relevant characteristic parameter is the "skewness value", as it measures the symmetry of a similarity distribution.

Theoretically, this means that if the respective (similarity or other) distribution is completely symmetrical, the skewness value is 0; it adopts a negative value if the respective distribution is skewed to the right, and it is positive if it is skewed to the left. A look at Map 5 shows that the *spatial* distributions of the blue (= negative) and the red (= positive) skewness values are spatially clearly distributed.

What is the genuine *geolinguistic* relevance of the synopsis of the skewness values? This question leads us to the concept of *Sprachausgleich* [linguistic compromise or exchange], which was first introduced in German linguistics. It

corresponds to the various phenomena of linguistic contact and compromise between large and small dialect areas with ultimately more or less hybridized zones.

Applied to Map 5 this means:

1) The dark blue zones represent an intensive linguistic compromise, indicating those zones where strong processes of hybridization took place.
2) The red zones represent a very weak language compromise and refer to zones which were kept almost out of the general process of language exchange.

On Map 5, a dark blue belt (or ditch) surrounding the Domaine d'Oïl and a dark blue semi-circle surrounding Francoprovençal can be detected. These two circular phenomena are 1) the consequence of the general centrifugal irradiation of the language type of the Langue d'Oïl, and 2) the result of those "rearguard actions" of the Latinity of Lugdunum/Lyon (the historical basis of Francoprovençal) with the old Latinities of the Langue d'Oïl (in the North), and of the Langue d'Oc (in the South).

The three red shaded "bulwarks" in the South on Maps 5 (Gascoigne, the Languedoc-Roussillon and the Provence) represent, on the other side, those landscapes which were hardly involved in the process of general linguistic exchange. The transitions between the dark blue and red nuclear zones of the two maps are loosely structured.

The striking structure of Map 5 obviously reflects the linguistic history of France since 500 AD, which was determined by supraregional dynamics in the North and the polycentric concurrence of several active regional sub-centres in the South.

6. The dendrographic DM: see Map 6

Since August Schleicher published his famous family tree of the Indo-European languages in 1863, family trees have been a metaphorically fruitful heuristic instrument in many fields of linguistics. Remember the frequent use of the genealogical tree for the reconstruction of linguistic parental relationships, especially in lexicostatistics. As it is quite easy with taxometrical methods to generate binarily structured trees on the basis of a similarity or distance matrix (with different algorithms). This heuristic instrument was

also included in DM analyses. In DM, it is of utmost importance to reset the resulting fragmentation of a genealogical tree directly into space, i.e. to project it spatially on a map ("spatialization"). With VDM, both procedures – the construction of the tree (with subsequent colouring) and the spatialization of the colouring – are realized without difficulty.

A terminological note: we refer to the branches of a tree which are marked by a significant colouring and their correspondences on the map as "dendremes".

Again, dendrographic DM yields a highly suggestive graphic output, but it requires a solid knowledge of the statistical processes in the tree generation. In brief: the specific algorithm we used in this instance (the Ward algorithm) belongs to the class of "hierarchic-agglomerative procedures". By a successive fusion or agglomeration of two very similar elements of the similarity matrix – beginning with the unconnected N elements ("leaves" of the future tree) – a binary hierarchy is generated. In this process, the number of the agglomerated clusters (= dendremes) is progressively reduced, leading to the final isolation of two big clusters, which can be united at the stem (or the root) of the tree. The inner quantitative heterogeneity of the dendremes is much greater next to the root (or the stem) of the tree than next to the leaves.

Geolinguistically, the tree can again be interpreted diachronically and synchronically. The diachronic interpretation simulates the progressive fragmentation of a given linguistic area, assuming that this linguistic area was originally homogeneous. Thus, Map 6 shows those two macro-dendremes/choremes (below nodes 1 and 2) which reflect a basic opposition between the North (dendremes/choremes A-D) and the South (dendremes/choremes E-G) of the Gallo-Romania. In the same way, the next fragmentation (nodes 3 and 4) divides the North of the Gallo-Romania again into two parts (dendreme/choreme D versus dendremes/choremes A-C), and so on.

The synchronic interpretation of the genealogical trees concentrates on spatial classification and regionalization, where the position of the dendremes in the branches of the tree indicates their relative reciprocal similarity. Note that on Map 6, the agglomerated dendremes B and C are more similar to each other than to dendreme A, or to dendreme D.

It is interesting to note the uniformity and the great spatial coherence of the choremes. In summary: dendrographic DM is able to reveal relatively deep geolinguistic

patterns, but requires a good understanding of the statistical-mathematical foundations of tree generation.

7. In lieu of a conclusion

This short presentation has only been able to convey a first impression of the manifold (and promising) applications of DM and – through the 6 coloured maps – to introduce the possibilities of the DM program VDM. Needless to say, it is freely available to all dialectometrically minded (geo)linguists. For comprehensive use, future adepts should nevertheless take advantage of (three or four days of) practical training in Salzburg. Several colleagues have already taken advantage of this opportunity, but as of the present only from Romance countries. Should this offer successfully reach a more varied group of linguists, thus spreading the provenience of DM specialists worldwide, this contribution would have reached its primary aim.

8. References

ALF: GILLIÉRON, Jules & EDMONT, Edmond (eds.) *Atlas linguistique de la France*, Paris: Champion 1902-1910, 10 vols. (reprint: Bologna: Forni 1968).

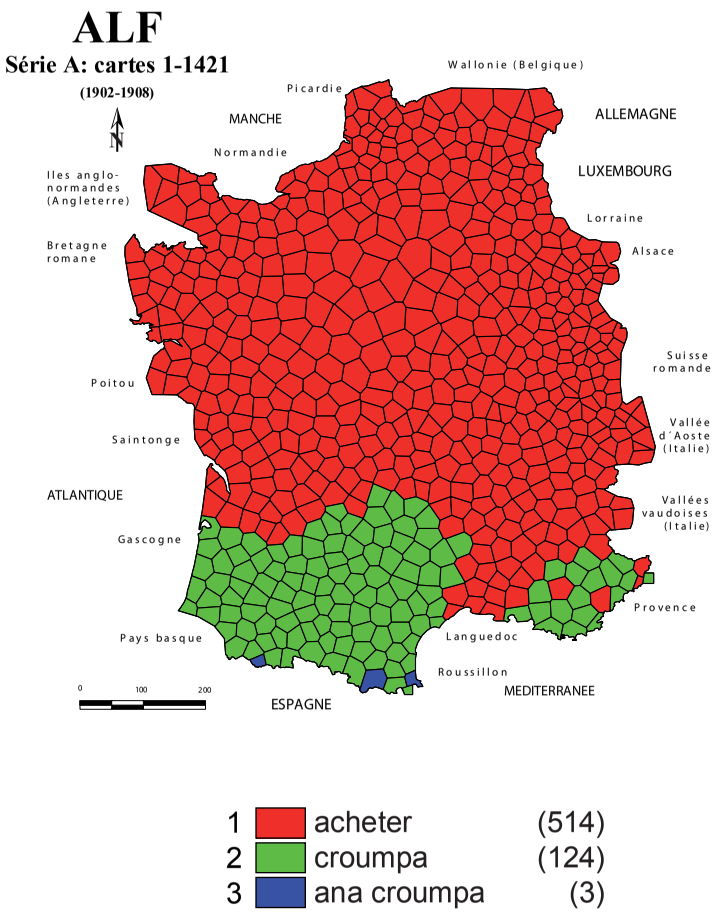
GOEBL, Hans (1984) *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF* [*Dialectometric studies: based on geolinguistic data taken from the AIS and ALF*], Tübingen: Niemeyer, 1984.

GOEBL, Hans (2003) "Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur" [*Some dialectometric considerations on the data of the «Atlas linguistique de la France (ALF)»: quantitative relations and deep structures*], *Estudis Romànics*, XXV, 59-121.

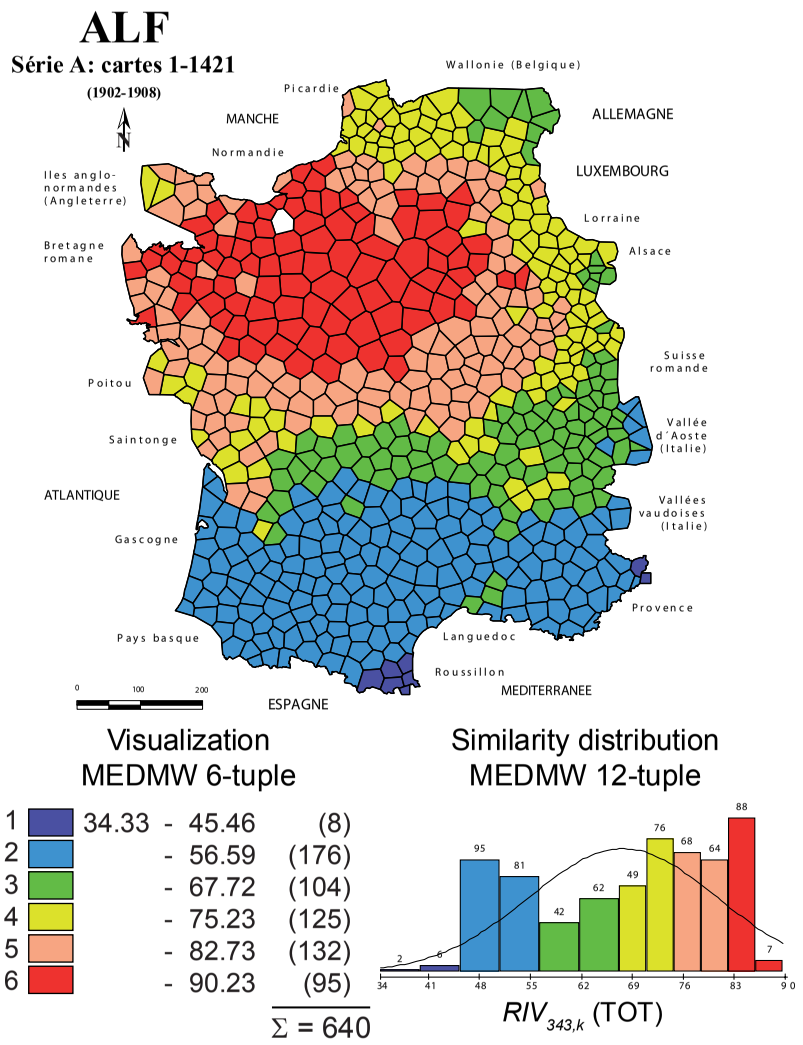
GOEBL, Hans (2005) "La dialectométrie corrélative. Un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme" [*The correlative Dialectometry: a new tool for the study of the dialectal management of space by man*], *Revue de linguistique romane*, 69, 321-367.

GOEBL, Hans (2006) "Recent Advances in Salzburg Dialectometry", *Literary and Linguistic Computing* 21/4, 411-435.

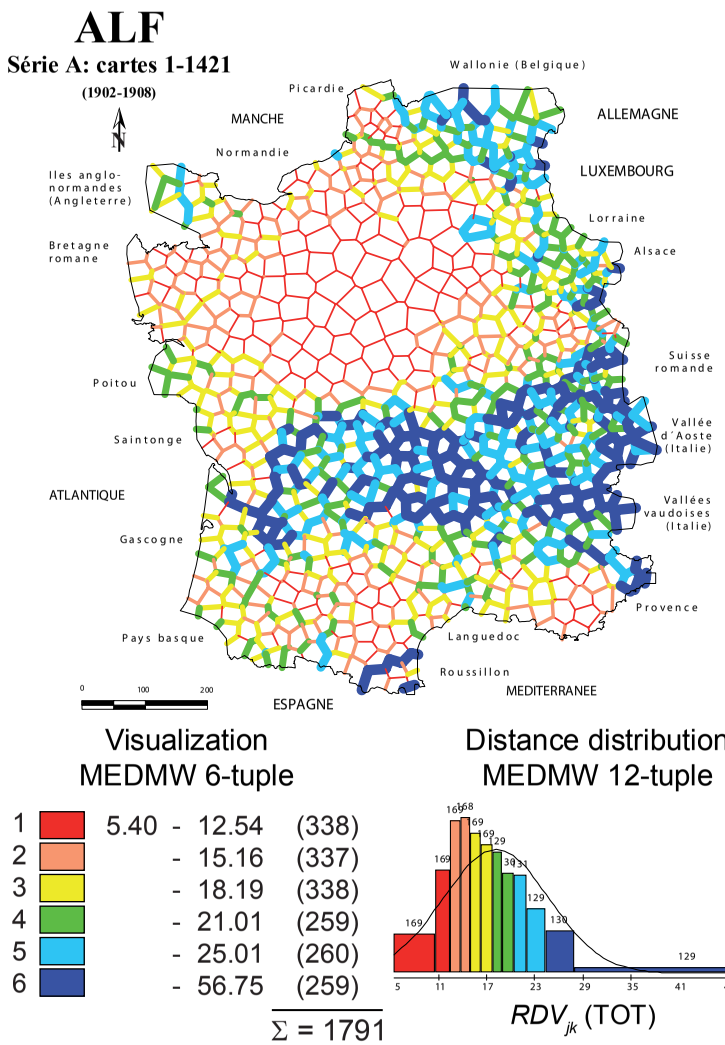
GOEBL, Hans (2007) "A Bunch of Dialectometric Flowers: a brief Introduction to Dialectometry", in SMIT, U./DOLLINGER, St./HÜTTNER, J./KALTENBÖCK, G./LUTZKY, U. (eds.), *Tracing English through Time. Explorations in Language Variation. In Honour of Herbert SCHENDL on the Occasion of his 65th Birthday*, Wien: Braumüller, 133-171.



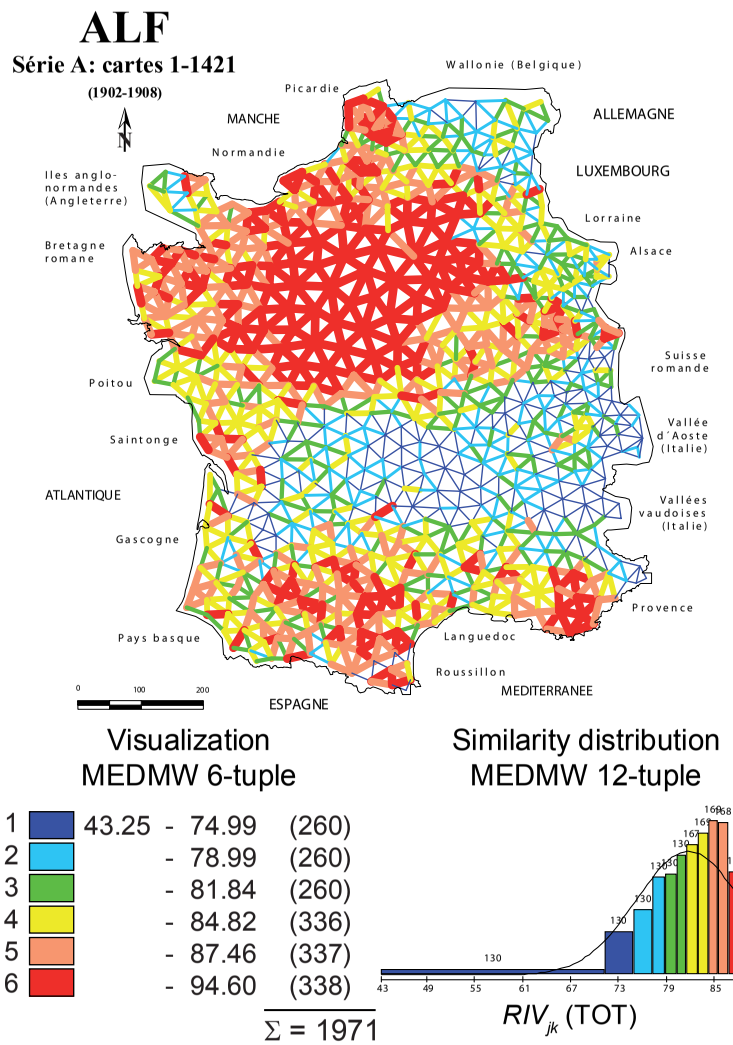
Map 1: Sample of a lexical „working map“ showing the spatial distribution of the Gallo-Romance designations of „to buy“ (following ALF 6 *acheter*)



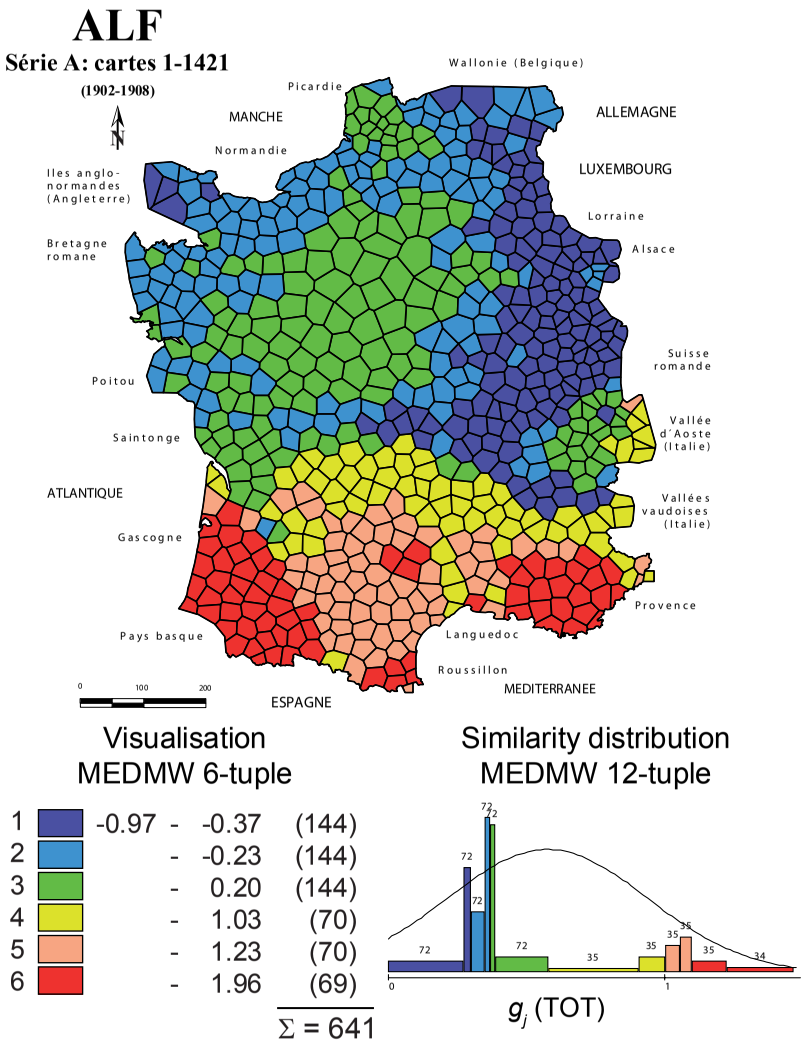
Map 2: A typical similarity profile of the northern Domaine d'Oïl: similarity map to the ALF-point 343 (La Chapelle-Yvon, Dép. Calvados) Similarity index: $RIV_{343,k}$; corpus: 1687 working maps (total corpus) Algorithm of visualization: MINMWMAX (6-tuple)



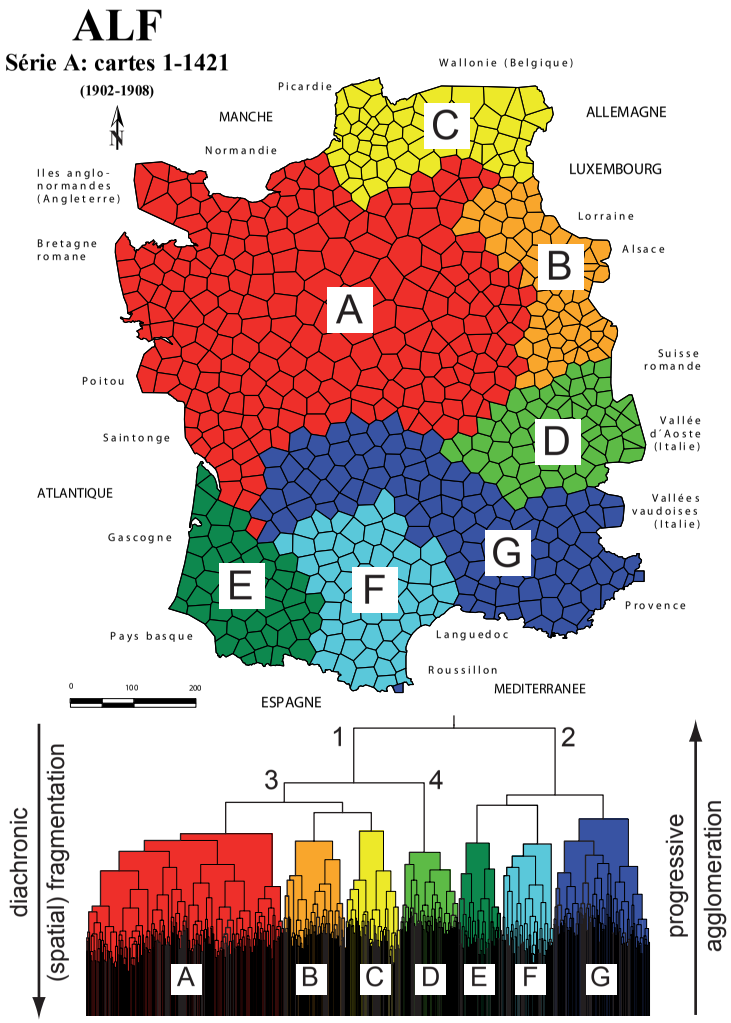
Map 3: Honeycomb map showing a synopsis of 1791 interpoint distance values
Distance index: RDV_{jk}
Corpus: 1687 working maps (ALF)
Algorithm of visualization: MEDMW (6-tuple)



Map 4: Beam map showing a synopsis of 1791 interpoint similarity values
Similarity index: RIV_{jk}
Corpus: 1687 working maps (ALF)
Algorithm of visualization: MEDMW (6-tuple)



Map 5: 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 6: Dendrographic classification of 641 dialectological objects (ALF-points) and the spatial conversion of the tree
Similarity index: RIV_{jk}; corpus: 1687 working maps (total corpus)
Dendrographic algorithm (above): hierarchical grouping method of Ward
Number of choremes corresponding to the dendremes (below): 7