The Rif Berber language continuum: an algorithmic geolinguistic study

Mena B. Lafkioui*

Aan collega en vriend Jacques van Keymeulen,
met wie ik de passie voor taalvariatie deel.

1. Introduction

The study presented in this article examines from an algorithmic geolinguistic perspective a corpus of lexical material of Rif Berber, which forms a language continuum covering the Rif area, which is located in North, Northwest and Northeast Morocco. In doing so, the study offers quantitative classifications of the Berber varieties of the Rif area and hence verifies the numerous qualitative classifications provided in the *Atlas linguistique des varieties berbères du Rif* (Lafkioui 2007), the ALR henceforth. This study builds further on the methods and results obtained from the algorithmic classifications of Rif Berber's lexis discussed in Lafkioui (2008, 2009), which provide evidence for the validity of the Levenshtein distance calculating method, also called edit distance, especially when the phone strings are tokenised in pair-wise alignments. Furthermore, among the many techniques to analyse and visualise aggregate distances, Multi Dimensional Schaling – MDS henceforth – was proven to be the best suited for studying language continua, which is the case of Rif Berber (Lafkioui 2007, 2008). The MDS technique has also the advantage to visualise the aggregates as well as the degree of their intra and inter linguistic divergence. Moreover, it is one of the most stable techniques, compared to classical clustering, for instance (Lafkioui 2008; Nerbonne et al. 2011). In this study, I will continue using these techniques, which draw on Kleiweg's free software tools (See http://www.let.rug.nl/kleiweg/ L04/) as well as on the more

* Université Sorbonne Paris Cité/LLACAN-UMR 8135.

recent web application GABMAP (Nerbonne et al. 2011), in addition to the data conversion programmes developed for this purpose and for which I am grateful to Bart Cocquyt for his assistance.

In section 1, I will present an overview of the Rif Berber geolinguistic area. In section 2, a presentation of the map and of the corpus used in this study will be given. Section 3 will be dedicated to the algorithmic classifications of the Rif Berber lexis and will discuss the results obtained, with a special focus on the techniques that allow identifying organic sites in the data. A general conclusion and a list of references cited in the study will end the article.

2. The Rif Berber geolinguistic continuum

The Rif is that region of Morocco stretching from the Strait of Gibraltar in the West to the Algerian frontier in the East, and from the Mediterranean See in the North to the corridor of Taza in the South, where Moroccan Arabic is mostly spoken. The Rif Berber varieties belong to the Northern Berber language type and thus are part of the large Berber language family, which forms a branch of the Afro-Asiatic language phylum (Lafkioui 2017). The Rif area consists of two main regions which are predominantly Berber-speaking: the small isolated area of Ghomara (Camps and Vignet-Zunz 1998; Colin 1929) and the extensive territory where Rif Berber, aka Tarifit, is spoken and which has the shape of a continuum. This Rif Berber continuum is delimited (see Figure 1):

- In the West, by the varieties of the Ktama group, which belong to the so-called Senhaja varieties.
- In the South, by the koinè of Gersif, which is the ultimate geographic point where Rif Berber is spoken before reaching the corridor of Taza.
- In the East, by the varieties of Iznasen, which have spread to the regions of Arabic-speaking varieties towards the Moroccan-Algerian border.

The Ghomara geolinguistic area is entirely detached from the Rif Berber continuum by the Arabic varieties of the Jbala, whose great impact on the Ghomara Berber varieties has played an important role in their linguistic distinctiveness (Mourigh 2016).

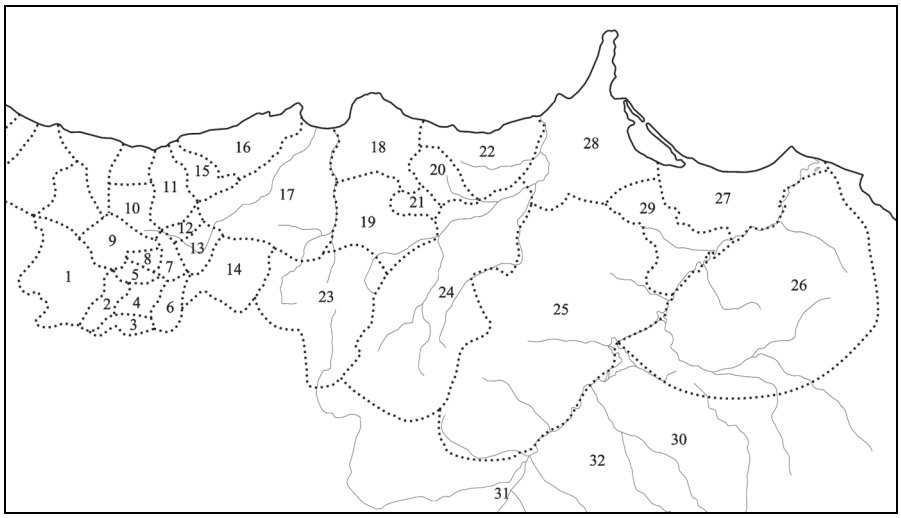

Figure 1: Map of the Berber-speaking groups of the Rif Berber continuum

1	Ktama	12	Targist	23	Igzennayen
2	Taɣzut	13	Ayt Mezduy	24	Ibḍalsen
3	Ayt Bušibet	14	Ayt Ɛammart	25	Ayt Buyeḥya
4	Ayt Ḥmed	15	Ayt Iṭṭeft	26	Iznasen
5	Ayt Bunsar	16	Ibeqquyen	27	Ikebdanen
6	Ayt Bšir	17	Ayt Weryaɣel	28	Iqelɛiyen
7	Zerqet	18	Ayt Temsaman	29	Wlad Settut
8	Ayt Ḫennus	19	Ayt Tuzin	30	Ayt Buzeggu
9	Ayt Seddat	20	Ayt Wlišek	31	Gersif
10	Ayt Gmil	21	Tafersit	32	Tawrirt
11	Ayt Bufraḥ	22	Ayt Sɛid		

Table 1: Berber-speaking groups of the Rif Berber continuum

Despite some classification challenges that come with geolinguistic continua, due to the gradual and hybrid variation of the forms and functions of which they are composed, my qualitative and quantitative research on the Rif Berber continuum clearly demonstrates that the further West we go, the more we detect features belonging to the "Senhaja" group (1-13), some of which are also attested in Ghomara Berber; such as, the lack of grammatical gender distinction in plural verb conjugations. Whereas towards the East, the "Zenet" traits predominate (especially 26, 27, 30-32), which are also found elsewhere in North Africa, such as in the Aures in East Algeria, in Djerba in South Tunisia and in Zouara in North Libya (Lafkioui

2007, 2017). Typical Zenet phenomena are, for instance, the palatalization of the velars *k*, *kk*, *g*, and *gg* (e.g. *šəm* instead of *kəm* 'you') and the conditioned absence of the prefixal vowel (e.g. *fus* instead of *afus* 'hand'). It should be mentioned, however, that the ethnolinguistic notions of Senhaja and Zenet, frequently used in Berber linguistics, are somewhat problematic because not only they tend to disappear in the language practices of Berber speakers, but they can also be strongly criticized, stigmatised even, which is particularly the case of the term Senhaja, even though a recent revalorisation of local ethnic and cultural identity can be observed in some local activist circles.

3. The Rif Berber geolinguistic map and lexical corpus

The data examined in this study mainly come from the ALR (Lafkioui 2007), of which the basic map with its 141 georeferenced points, belonging to 32 Rif Berber-speaking groups (Figure 1), is extracted and presented in Figure 2. These points are a selection of the 452 points that were examined and selected by their degree of linguistic variation and comparativeness in the ALR. Initially, the survey points were selected on the basis of the principle of equidistance, which divides the inquiry field into several grids to which were assigned points that could match with localities on the field. The greater the variation, the more the grids were reduced.

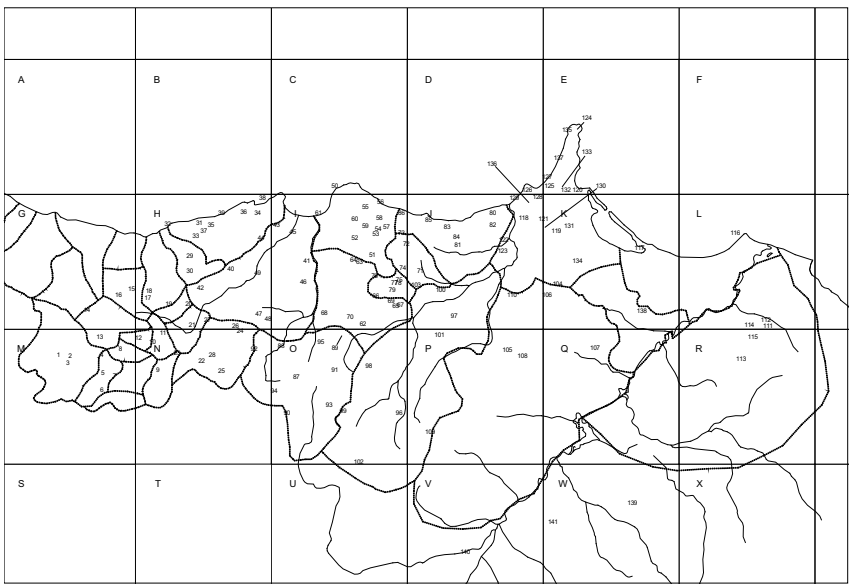

Figure 2: Map of the selected georeferenced points of the Rif (ALR)

As for the digital data that are compared and classified in this study, they consist of 169 lexical items regarding the human body, kinship, animals, colours, numbers, along with a subset of various nouns and verbs. This lexical selection amounts to 287459 tokens and stems from a vast geolinguistic corpus built by means of specific methodological procedures concerning data gathering, their systematisation, and their archiving (Lafkioui 2007, 2015). The data examined here are also based on numerous linguistic, sociolinguistic and ethnographic fieldwork investigations in the Rif area, which started in 1992 and of which the last one was in autumn 2017.

Due to the digital nature of the ALR, algorithmic classification was possible, although an adaptive conversion to the data formats used by the RuG/L04 software and by GABMAP was necessary, involving also a systematic conversion to UTF-8 for the geolinguistic data and to KML[1] for the geographic data, which was a time and energy consuming task.

The tokenized and pair wise aligned lexical data used for this research meet well the criteria for good quality data, as is indicated by the two relating measures: Cronbach's α, which has a value of 0.99 here, taking into account that the closer to 1 the better the score, with a minimum of 0.7. As for the local incoherence measure, the data has a value of 0.90 (the closer to 0 the better the score), while, for instance, in the study of Nerbonne & Kleiweg (2007), which aims at providing a yardstick for dialectology research, values range from 1.75 to 2.05.

4. Algorithmic classifications of the Rif Berber lexis

Compared to the algorithmic classifications of the Rif Berber lexis presented in Lafkioui (2008, 2009), this study provides similar classification outcomes, notwithstanding the significant augmentation (from 62 items to 169 items) and diversification of the data. The first classification that accounts for this general similarity is projected onto the Classical MDS map (with r= 0.99) in Figure 3, which aggregates the lexical differences measured between the various geolinguistic sites considered.

This MDS classification not only confirms that the Rif Berber varieties form a language continuum, which means that one variety gradually merges into another variety when they are contiguous, but also that this continuum contains some major subdivisions, which correspond to Eastern Rif Berber (cream, salmon and pink aggregates), Central Rif Berber (light and dark orange aggregates), Central-Western Rif Berber (fuchsia and blue aggregates), and Western Rif Berber (green

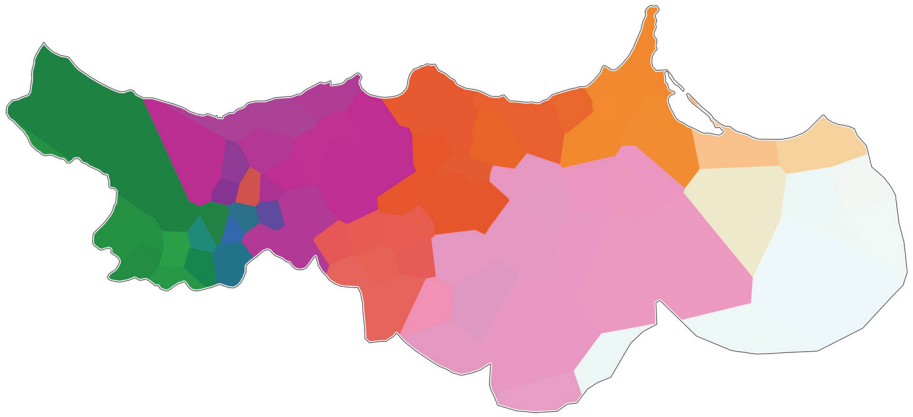

Figure 3: Classical MDS Map of Rif Berber lexis

aggregates). These aggregates are validated by means of a corresponding MDS scatter plot, a stable technique that GABMAP offers for this purpose and which usually stands for more than 80% of the variation in the data. The distances as measured by the plot show a high correlation with the distances given in the linguistic distance table, with a value of r=0.98.

Another technique that confirms that these obtained aggregates are very stable and so accounts for an adequate aggregate display of the data is the probabilistic clustering technique, which basically consists of constantly adding quantities of noise while clustering and maintaining the cophenetic distance of the sites compared (Nerbonne et al. 2008). Even after 0.8 of noise added – while the default extra noise is 0.2 – the aggregates remain stable. The dendrogram displayed in Figure 4 represents a clustering with a noise level of 0.2 added.

The stability of these major aggregates of the Rif Berber lexis is also verified by other algorithmic classification techniques, as is shown in Figures 5 and 6, which present the results of a clustering classification based on the following weighted average algorithm (GABMAP):

$$d_{k[ij]} = \left(\frac{1}{2} \times d_{ki}\right) + \left(\frac{1}{2} \times d_{kj}\right)$$

In doing so, the Berber data corroborate that this algorithm has the advantage of delivering consistent and representative clusters, as it allocates equal weight to the clusters that merge, despite the unequal number of sites that make up each cluster. Note that these clusters are also validated by means of the GABMAP cluster validation technique, which draws on MDS and its two dimension plots.

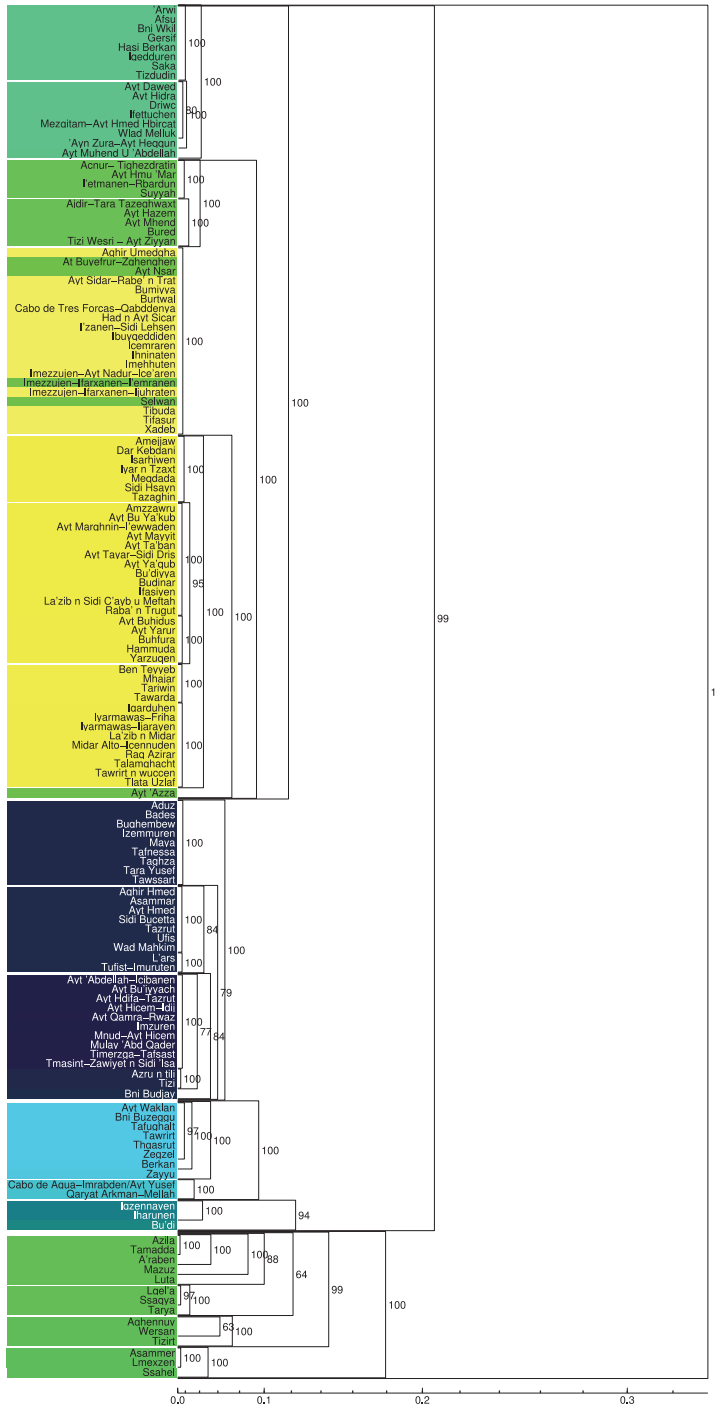

Figure 4: Probabilistic dendrogram of Rif Berber lexis

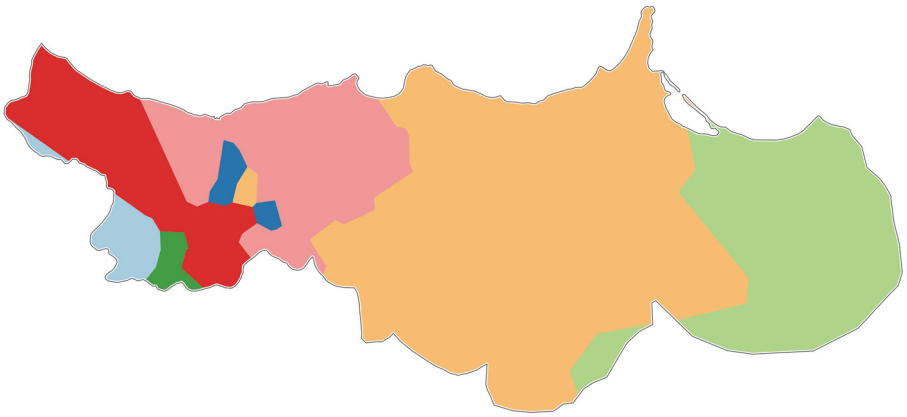

Figure 5: Weighted average cluster analysis map of Rif Berber lexis

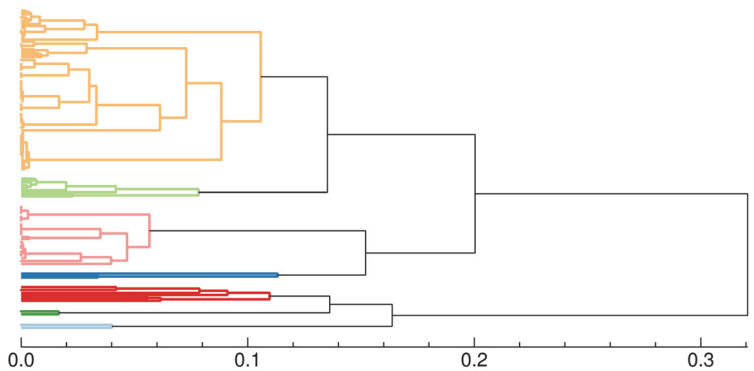

Figure 6: Weighted average cluster analysis dendrogram of Rif Berber lexis

The dendrogram in Figure 6 also shows at glance that the main subdivision of the Rif Berber continuum is made between the Western Rif Berber varieties (red branch as boundary) and the rest of the vaste continuum. The second most important subdivision, on the other hand, is made between the Central-Western Rif Berber varieties (pink branch as boundary) and the remaining large part of the Rif Berber continuum. All this evidence substantiate Lafkioui's findings (2008, 2009).

In addition, my many linguistic and sociolinguistic surveys in the Rif area clearly indicate that daily contact is important in the spread of linguistic variants (see e.g. Lafkioui 2011, Forthcoming). This is also the case of lexical variation, which is diffused through networks of neighbouring sites in the Rif area, as is displayed on the difference map in Figure 7.

As one can see on this map, several networks are formed and spread over the entire Rif Berber continuum. The darker the colour of the lines, the more alike are

Figure 7: Difference map of adjacent sites for Rif Berber lexis

the linguistic sites they connect and hence the stronger the linguistic networks they constitute and the more frequent and intense the contacts between the Berber speakers involved. Furthermore, when all sites of the area are taken into account, as in Figure 8, one not only distinguishes the same core aggregates as in Figure 7 (considering the adjacent sites only), but also perfectly witnesses the whole continuum structure emerging.

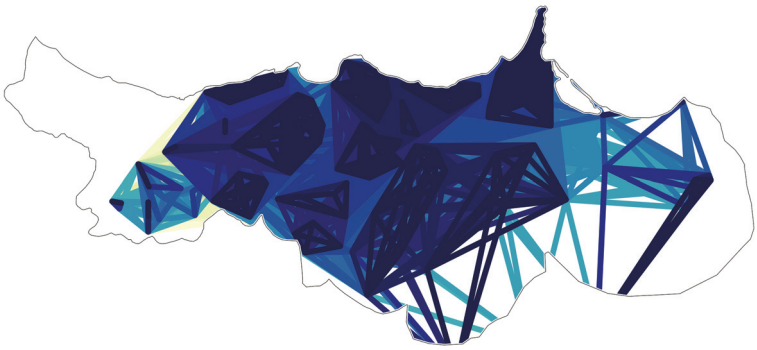

Figure 8: Difference map of all sites for Rif Berber lexis

Finally, this algorithmic study also examines systematically which lexical items are accountable for the major geolinguistic differences attested in the Rif Berber area. In other words, the study objectively identifies which lexical features primarily determine the formation of the different aggregates. In order to do so, the two quantitative measures of representativeness and of distinctiveness, which are provided by GABMAP, are employed. The main outcomes of this data mining task point to a set of lexemes functioning as cluster discriminators, which belong to various semantic fields, although slightly higher scores are observed for the fields concerning the human body and the expression of time and space.

5. Conclusion

The present study and its outcomes fully support the algorithmic classifications of the Rif Berber geolinguistic area as presented in Lafkioui (2008, 2009) and, accordingly, demonstrate that this area has the makeup of a language continuum, which includes a number of stable core aggregates corresponding to the following geolinguistic subdivisions: Eastern Rif Berber, Central Rif Berber, Central-Western Rif Berber, and Western Rif Berber. Moreover, given that no substantial difference has been detected between the aggregates resulting from this extensive examination and those aggregates obtained in Lafkioui (2008, 2009), which were based on a smaller lexical corpus, one can confidently infer that the latter corpus is representative, and hence well selected, since it contains most of the items that identify the stable aggregates of the Rif Berber continuum.

References

Camps, G., & J. Vignet-Zunz (1998). GhomâRa. *Encyclopédie berbère | Gauda – Girrei* 20: 3110-19.

Colin, G.S. (1929). Le parler berbère des Ghmara. *Hespéris*, 173-208.

Lafkioui, M. B. (2007). *Atlas linguistique des variétés berbères du Rif*. Köln: Rüdiger Köppe Verlag.

Lafkioui, M. B. (2008). Dialectometry analysis of Berber lexis. *Folia Orientalia* 44: 71-88.

Lafkioui, M. B. (2009). Analyses Dialectométriques Du Lexique Berbère Du Rif. In: D. Ibriszimow, R. Vossen, & H. Stroomer (eds.), *Studien zur Berberologie/Etudes Berbères* 4. Köln: Rüdiger Köppe Verlag, 133-50.

Lafkioui, M. B. (2011). How system-internal linguistic factors indicate language change and diffusion. A geolinguistic analysis of Berber data." *Dialectologia et Geolinguistica* 19: 62-80.

Lafkioui, M. B. (2015). Méthodologie de recherche en géolinguistique. *Revue Corpus* 14: 139-64.

Lafkioui, M. B. (2017). "Rif: La Langue (Rifain/Tarifit)." *Encyclopédie Berbère* 41: 6916-56.

Lafkioui, M.B. (Forthcoming). Geolinguistic complexity in Berber: Structural and algorithmic perspectives. *Dialectologia et Geolinguistica*.

Mourigh, K. (2016). *A grammar of Ghomara Berber (North-West Morocco)*. Berber Studies, volume 45. Köln: Rüdiger Köppe Verlag.

Nerbonne, J., R. Colen, C.S. Gooskens, T. Leinonen, & P. Kleiweg. (2011). Gabmap – A web application for dialectology. *Dialectologia*, no. II: 65-89.

Nerbonne, J. & P. Kleiweg (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14: 148-66.

Nerbonne, J., P. Kleiweg, W. Heeringa, & F. Manni. (2008). Projecting dialect differences to geography: Bootstrap clustering vs. noisy clustering. In: Ch. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (eds.), *Data analysis, machine learning, and applications. Proc. of the 31st Annual Meeting of the German Classification Society*. Berlin: Springer, 647-54.

◈ ◈ ◈

Notes

[1] KML is an open standard officially named the OpenGIS® KML Encoding Standard (OGC KML). It is maintained by the Open Geospatial Consortium, Inc. (OGC). The complete specification for OGC KML can be found at http://www.opengeospatial.org/standards/kml/.