مجلة العلوم الإنسانية
ISSN 1112-9255
العدد السابع / الجزء(2) - جوان 2017

OEB Univ. Publish. Co.

Amazigh Lexicostatistics: A Correlation Study between Geographic distances and Lexical Differences

أمازيغ ليكسيكوستاتيستيكش: دراسة الترابط بين المسافات الجغرافية والاختلافات المعجمية

Bouri Hadj,Tlemcen University,Algeria.

تاريخ التسليم:(2017/01/22)، تاريخ القبول:(2017/04/19)

Abstract :

This study aims at measuring the impact of geographic distance on the linguistic difference. A five word questionnaire was designed to elicit Berber lexis based on the Swadesh list. A levenshtein algorithm was applied to measure the linguistic distance between five North African Berberophone regions: Algeria, Morocco, Mauritania, Tunisia and Libya. Also, a correlation experimental design was administered to conduct the study and measure how language varies according to the geographical landscape of the region. The results show a positive influence of the geographical distance on Berber linguistic diversity. The implementation of inferential statistics along with the levenshtein algorithm helps in understanding how the Berber language intersects with its geographical landscape.

Keywords: Geolinguistics, Levenshtein distance, Dialectometry, Amazigh language, Lexicostatistics, Language Variety

ملخص :

تهدف هذه الدراسة إلى قياس أثر المسافة الجغرافية على الاختلاف اللغوي. وقد تم تصميم استبيان من خمسة كلمات أمازيغية على أساس قائمة سواديش. تم تطبيق لوغاريتم ليفنشتاين لقياس المسافة اللغوية بين خمس بلدان شمال أفريقية: الجزائر والمغرب وموريتانيا وتونس وليبيا. كما تم تصميم نموذج تجريبي للارتباط لإجراء الدراسة وقياس مدى اختلاف اللغة تبعا للمسافة الجغرافية للمنطقة. وأظهرت النتائج تأثيرا إيجابيا للمسافة الجغرافية على التنوع اللغوي الأمازيغي. تطبيق الإحصاءات الاستنتاجية جنبا إلى جنب مع لوغاريتم ليفنشتاين يساعد في فهم كيفية تقاطع اللغة الأمازيغية مع المشهد الجغرافي.

الكلمات المفتاحية: الجغرافية اللسانية، قياس ليفنشتاين، القياس اللهجي، اللغة الأمازيغية، الإحصاء المفرداتي، التنوع اللغوي

Introduction :

Understanding the complexity of Berber lexical diversity is vitally important to the variationists, if newly adopted approaches in dialectometry are applied. Contemporary Studies on the phylogenetics of Berber represent a growing field in the field of Berber linguistics. In the last few decades, there has been a surge of interest in the effects of geographical and ethnic variables on the linguistic variation of the Berber language. In the literature on lexical variation of Berber, the relative importance of the word has been subject to considerable discussion by a number of scholars. To date, however, there has been little experimental evidence on how the large geographical distance affects the lexical diversity of Berber language. This paper offers a new model that combines a nonexperimental correlation method and a data mining process under the name of the leveishtein algorithm to understand fully how language varies between the different North African Berber communities. The purpose of this investigation is to explore the relationship between language difference and geographic distance amid 29 North African Berber communities. This paper begins by a highlight on the literature review related to previous studies and approaches of linguistics on Berber dialects. and It will then go on to a practical part where a correlation and a dendrogram analyses on Berber varieties are conducted.

1 - **Literature Review :**

Studies on Berber dialectometry represent a growing field in linguistics and language of the minorities. Both concepts of physical distance and linguistic diversity are central to the study of the impact of geographical features on language variation. Traditionally, Hans Goebl(2008, 2014, 2010, 1982) has subscribed to the belief that language variation according to geographical distance can be measured. Traditionally, linguistics scholars have subscribed to the belief that language boundaries or isoglosses with all its types are geolinguistic zones that can be measured and the amount of linguistic diversity can be delimited. Since the appearance of dialectometry, imaginary linguistic boundaries have been subject for further clarification. The impact of geographical distance on language diversity was a key issue in dialectology, traditional linguists tried to study the relationship between language and geography with poor approaches and sketchy methodological steps. Both the choice of the linguistic features as well as the sampling of remote populations subdued inadequacy. However, knowing the importance of the impact of geographical elements in determining linguistic variation is primordial.

Results from earlier studies demonstrated a strong and consistent association between geography and language. It has been observed that the larger the geographical distance is the diverse the linguistic features will be. What we know about language variation and geography comes from accounts by Goebl and many other dialectologists till Peter Trudgill. To date, there has been little agreement about how best to design linguistic atlases and how to approach in scientific research the linguistic diversity in remote geographical areas. Also, there is a current paucity of high-quality research on nonexperimental research in Berber dialectometry. Previous studies have failed to consider the geographical element as an independent variable that modifies considerably the linguistic variation of the Berber language; however, there has been no empirical evidence that clarifies the lexicostatistics of the Berber language and its clear distribution on the geographical distance. Previous studies in Berber dialectology have suffered from several conceptual and methodological

weaknesses. Many sociolinguists from the Maghreb have highlighted linguistic variation broadly and mentioned the isoglosses between different regional varieties either in traditional inaccurate maps, as in the works of André Basset (Chaker, 1995), or researchers were unable to collect and draw geolinguistic data from the vast and complex geographical area of the North African countries. This is why the extent to which geographical distance affects linguistic features of Berber is still poorly highlighted by many Dialectologists.

In this context, this paper comes to investigate the design and the implementation of lexicostatistical as well as geographical techniques to understand fully how Berber correlates with geographical features of the Maghreb region. This study seeks to answer the following specific hypothesis: In the Maghreb region, Berber's lexical features vary due to geographical distance. This study draws on two theoretical frameworks: First a levenshtein algorithm was applied on 5 words list questionnaire based on Swadesh list (Zastrow, 2011) which are "expected to be culturally neutral and stable over time, a real influence is kept to a minimum and diachronic conclusions are potentially justified" (Jack Grieve, 2011). In this questionnaire informants write the equivalent of the word in Arabic in their local Berber dialect. Also a non-experimental study was conducted were the researcher sought to find a correlation between geographical distance and linguistic diversity. The experimental work presented here provides one of the first investigations into how to measure Berber language variation according to geographical distance

.2-**Material and Methods:** Recently, a considerable literature has grown up around the theme of dialectometry in general (Haimerl, 2006; Heeringa, 1970; Mucha & Haimerl, 2005; Nerbonne & Kretzschmar, 2003; Szmrecsanyi, 2008). Dialectometry has been studied extensively since the last decade of the twentieth century and, as a concept in computational linguistics, it is widespread among scholars in northern Europe, Germany the USA and other parts of the world. It is also fundamental to contemporary linguistics since computational tools have given larger perspectives to linguistic studies. Nerbonne and Heeringa are major contributors in this field with their numerous scientific articles. "In dialectometry, the dialect data collected mostly in language lexicon or dialect dictionaries are analysed by means of quantitative methods (statistics, information theory, etc.) with the aid of electronic data processing systems and methods"(Zastrow, 2011). The aim is to make the linguistic structures between the individual dialects of a language visible. The levenshtein algorithm is one of the key components of dialectometry. Evidence suggests that geographical distance is among the most important factors for a diverse language. In recent years, researchers have shown an increased interest in Berber dialectometry. Lafkioui (2008) has been attracting considerable interest since the beginning of 2000. One advantage of using computational approaches to study dialect variation is that it allows the synthetic quantitative analysis and apprehension of linguistic atlas using geolinguistic and numerical taxonomies. Both geolinguistic and statistical calculations are displayed on charts using VDM Visual Dialectometry designed in 2000 by Edgar HAIMERL (Hans Goebl, 2010; Jeszenszky & Weibel, 2015).

Berber Dialectometry: Eighteen informants from all the five North African countries: Algeria, Morocco, Tunisia, Mauritania and Libya, were recruited for this study, as in the map bellow (figure 1). As table one shows, we have chosen eleven

regions in Algeria, in Morocco three, in Tunisia two, and in both Mauritania and Libya One region for each.

TABLE 1. North African Countries and the Berberophone regions selected in this research

Algeria	**Morocco**	**Tunisia**	**Mauritania**	**Libya**
Benisnous	Figuig	Djerba	Mederdra	Ghat
Boussemghoun	Ksour	Ghadames		
Chawia	Rif			
Ghardaia				
Kabil				
Menaceur				
Moughel				
Ouedghir				
Sfisifa				
Tamantit				
Touareg				

Data were collectedusing a questionnaire where the informants were asked to fill the appropriate Berber word in front of the equivalent Arab one. The list of 5 words is selected from the Berber vocabulary under the criteria set by Swadesh list, as above-mentioned. The method applied in this study in String Edit Distance Tokenized. The local incoherence is 0.64 which means that the results drawn from these regions are different due to the large geographic distances between the North African Berber regions and the Lower the values for Local incoherent are the similar the results will be.Also the Cronbach's Alpha of the questionnaire is 0.83 percent which means that the validity of the questionnaire is high and can be trusted as a tool of collection and measurement of the linguistic data aggregated.

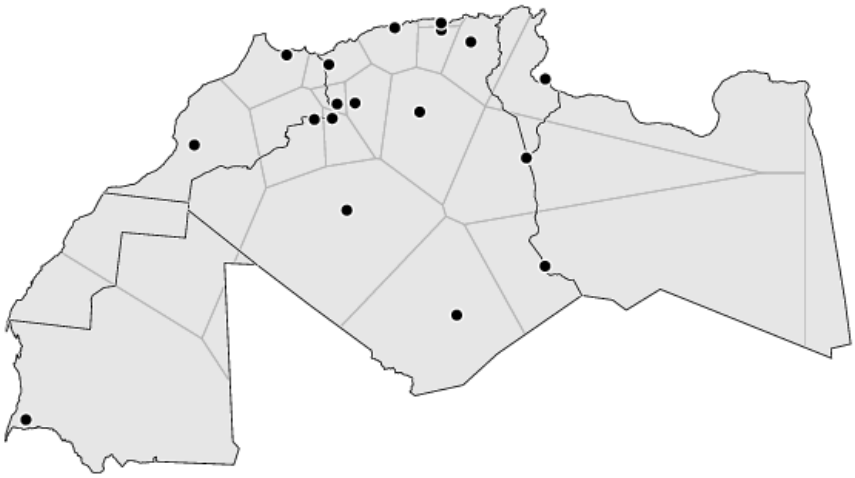

FIGURE 1. The Main North African Berber region where the research was conducted.

LEVENSHTEIN DISTANCE:

The collection of data was conducted over the course of the growing period of the second semester of 2016. All the work on the computer was carried out using R (Team, 2017) software for statistical analysis and GABMAP (John Nerbonne, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, 2011) an online software for dialectal data mining and visualization. In order to understand how geographical distance regulates linguistic variation, a levenshtein algorithm was applied to compare between the linguistic strings: where addition, omission or substitution of sounds were given the value of 1 between each location and the rest of the other geographical areas at the level of each string that is to say each lexical item is used as a basis to compare between all the selectedBerberophone regions of North Africa. As Tables 2 displays an example of how the operation processes at the lexical level of the Berber word *camel*. This comparison is made between four regions where the linguistic distance between the Chawia and the Menaceuris 1 and Djerba and Ouedghiris2. This function was repeated with all the 5 lexemes in a binary comparison between all the 18North African regions.

TABLE 2.Binary distance matrix of the lexeme "Camel" between four regional dialects

chawia — menaceur

a	l	r		m	
a	l	r	u	m	
			1		**1**

djerba — ouedghir

a	l	r	a	m	
a	l	g	o	m	
		1	1		**2**

Correlation:

After conducting a Levenshtein analysis, a linguistic distance matrix was designed, this latter was plotted with another geographical distance matrix to picture how linguistic variation between the different Amazigh regions correlate with geographical distance.A statistical analysis was used based on nonexperimental studies a correlation experimental design was administered. The aim from this latter is to interpret how linguistic data are influenced bygeographical distance. A plot with local regression and asymptotic regression was designed and the value seen on the plot chart shows the value of a: 0.03243 and b: 0.32503. The value of these two numbers is very small which indicates that there is a very low signal ratio in the data. Also the value of c equals 23.87622 which mean that linguistic variation is measurable over a large geographic distance. These statistical results can be clearly seen on figure 2 where linguistic difference is plotted with geographic distance.A curve linear

positive relation between the geographic and the linguistic element was found were the higher the geographical distance is the more different linguistic differences between the Amazigh varieties will be.

	benisnous	boussemgho	chawia	djerba	figuig	ghadames	ghardaia	ghat	kabil	ksour	mederdra	menaceur	moughel	ouedghir	rif	sfisifa	tamantit	touareg
benisnous	0																	
boussemgho	0.196687	0																
chawia	0.213858	0.150808	0															
djerba	0.194161	0.136835	0.03367	0														
figuig	0.160179	0.072549	0.130111	0.097619	0													
ghadames	0.288674	0.237688	0.178019	0.162194	0.243146	0												
ghardaia	0.258688	0.156593	0.14785	0.131099	0.163903	0.119801	0											
ghat	0.243434	0.303263	0.321951	0.318332	0.285864	0.364616	0.323461	0										
kabil	0.219192	0.12116	0.148419	0.131415	0.133796	0.247045	0.242843	0.287807	0									
ksour	0.17129	0.072549	0.148629	0.116138	0.0185185	0.26069	0.181447	0.285864	0.115278	0								
mederdra	0.481088	0.477152	0.453792	0.440324	0.456197	0.453691	0.446296	0.518398	0.462735	0.464969	0							
menaceur	0.138889	0.151292	0.144841	0.145515	0.144865	0.239905	0.222455	0.272908	0.0951178	0.154125	0.449989	0						
moughel	0.17129	0.072549	0.148629	0.116138	0.0185185	0.26069	0.181447	0.285864	0.115278	0	0.464969	0.154125	0					
ouedghir	0.269589	0.167361	0.18878	0.17144	0.222631	0.242827	0.221958	0.346873	0.130882	0.212827	0.432326	0.190815	0.212827	0				
rif	0	0.196687	0.213858	0.194161	0.160179	0.288674	0.258688	0.243434	0.219192	0.17129	0.481088	0.138889	0.17129	0.269589	0			
sfisifa	0.17129	0.072549	0.148629	0.116138	0.0185185	0.26069	0.181447	0.285864	0.115278	0	0.464969	0.154125	0	0.212827	0.17129	0		
tamantit	0.182509	0.118939	0.145931	0.146015	0.130214	0.256038	0.226459	0.242657	0.140144	0.140018	0.464695	0.109804	0.140018	0.200316	0.182509	0.140018	0	
touareg	0.188492	0.279623	0.29382	0.276143	0.229787	0.352009	0.308983	0.220862	0.269311	0.23959	0.479341	0.234804	0.23959	0.298927	0.188492	0.23959	0.17735	0

Table4. Average geographic distance matrix among Amazigh regions

	benisnous	boussemgho	chawia	djerba	figuig	ghadames	ghardaia	ghat	kabil	ksour	mederdra	menaceur	moughel	ouedghir	rif	sfisifa	tamantit	touareg
benisnous	0																	
boussemgho	245.673	0																
chawia	759.632	691.94	0															
djerba	1144.74	1011.87	438.198	0														
figuig	284.544	145.354	836.628	1143.57	0													
ghadames	1153.26	949.628	678.031	426.531	1045.98	0												
ghardaia	541.036	344.715	458.136	684.671	463.431	612.994	0											
ghat	1559.59	1321.09	1243.84	982.392	1367.02	576.896	1047.88	0										
kabil	622.081	597.568	167.367	605.178	742.209	809.535	447.437	1358.58	0									
ksour	822.01	872.14	1553.6	1880.13	736.576	1749.56	1198.13	1953.37	1434.08	0								
mederdra	2416.19	2367.9	3041.26	3241.39	2224.44	2946.46	2599.41	2825.23	2964.95	1650.61	0							
menaceur	398.843	449.547	408.441	838.417	581.629	976.433	463.342	1485.79	246.603	1220.76	2795.93	0						
moughel	299.223	231.462	922.472	1237.1	94.2145	1137.31	557.635	1442.32	820.545	643.291	2153.25	643.78	0					
ouedghir	632.837	621.347	185.517	622.998	765.24	841.396	483.226	1393.15	37.2457	1449	2989.25	246.407	841.527	0				
rif	227.636	440.372	974.568	1370.05	411.091	1376.33	763.336	1759.89	827.31	675.013	2314.15	588.194	369.182	832.837	0			
sfisifa	212.529	94.7988	778.997	1106.44	80.3434	1038.79	438.182	1391.2	675.526	779.468	2295.87	504.766	145.047	696.744	370.382	0		
tamantit	773.932	567.092	1104.43	1255.29	490.155	987.088	646.299	1086.87	1074.84	870.37	1986.22	995.942	508.064	1107.38	876.772	562.498	0	
touareg	1485.59	1242.31	1444.11	1328.02	1229.54	905.837	1090.98	531.781	1507.42	1645.8	2309.3	1552.51	1277.7	1544.65	1640.04	1281.55	802.802	0

Also the correlation is not so strong since the dispersion of the localities is not stable as shown in figure2. The plot also shows that the majority of the localities in the region, especially those in the far North, under research nearly share the same linguistic traits.

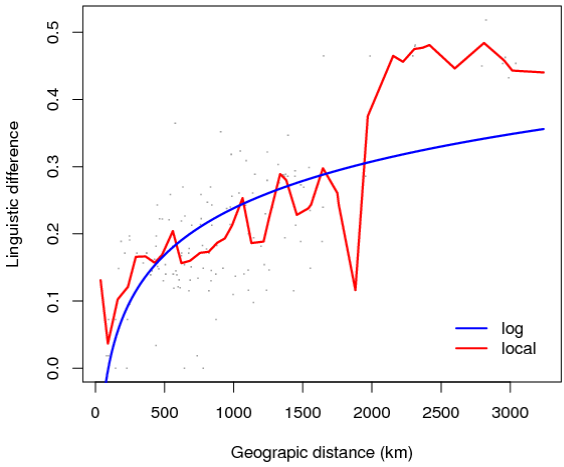

FIGURE 2. Geographic and linguistic distance between the North African localities

Conclusion:

The present study was designed to determine the effect of geographical distance on Amazigh linguistic difference. One of the more significant findings to emerge from this study is that Amazigh varieties as the Tshawit or the Targui differ tremendously because of the vast distances that separate them. The same case can be inferred from the other remote and dispersed varieties This study produced results which corroborate the findings of a great deal of the previous work in dialectometry led by Goebl, Haimerl and Heeringa. Besides, the data we processed and the results we obtained must be interpreted with caution because we have relied only on a five word corpus which is not enough to over generalize our result on a huge community living in a so vast geographical area as the North of Africa. Therefore, geographical distance could be a major factor causing lexical variation, but other factors should be interfering as the ethnic and social elements which were not taken into account in this study.

References :

-Case, A. R., & Dialectometry, O. F. (2009). New insights into the use of vdm: some preliminary stages, 2, 23–35.

- Chaker, S. (1995). Dialecte. In Encyclopédie berbère (Vol. XV, pp. 1–5).

- Goebl, H. (2008). Brève Introduction Aux Problèmes Et Méthodes De La Dialectométrie. Revue Roumaine de Linguistique, 1–2, 87–106.

- Goebl, H. (2014). L ' Impact De La Polynymie Des Cartes D ' Atlas Sur Le Résultat. Linguistique Romane et Linguistique Indoeuropéenne, 243--260.

- Goebl, H. (1982). Atlas, Matrices Et Similarities: Petit Apercu Dialectometrique. Computers and the Humanities, 16, 69–84.

- Goebl, H. (2010). Dialectometry: Theoretical Prerequisites, Practical Problems, And Concrete Applications (Mainly With Examples Drawn From The “Atlas Linguistique De La France”, 1902-1910). Dialectologia. Special Issue, 1, 63–77.

- Haimerl, E. (2006). Database Design and Technical Solutions for the Management , Calculation , and Visualization of Dialect Mass Data. Literary and Linguistic Computing, 21(4), 437–444. https://doi.org/10.1093/llc/fql037
- Heeringa, W. (1970). Measuring Dialect Pronunciation Differences using Levenshtein Distance. University of Groningen.
- Jack Grieve. (2011). The use of spatial autocorrelation statistics for the analysis of regional linguistic variation. In A. Z. and A. Lüdeling (Ed.), Proceedings of Quantitative Investigations in Theoretical Linguistics 4 (pp. 34–36). Humboldt-Universität zu Berlin.
- Jeszenszky, P., & Weibel, R. (2015). Measuring boundaries in the dialect continuum. Proceedings of the AGILE.
- Lafkioui, M. (2008). Dialectometry Analyses Of Berber Lexis. Folia Orientalia, 44, 71–88.
- Mucha, H., & Haimerl, E. (2005). Automatic Validation of Hierarchical Cluster Analysis with Application in Dialectometry. In Classification the Ubiquitous Challenge (pp. 513--520). Springer.
- Nerbonne, J., & Kretzschmar, W. (2003). Introducing Computational Techniques in Dialectometry. In Computers and the Humanities (Vol. 37, pp. 245–255). Netherlands.: Kluwer Academic Publishers.
- Szmrecsanyi, B. (2008). Analyzing aggregated linguistic data, 1–27.
- Zastrow, T. (2011). Neue Analyse- und Visualisierungsmethoden in der Dialektometrie. akultät der Eberhard Karls Universität.