Dialectologia 23 *(2019), 151-171.*
ISSN: 2013-2247

Received 30 April 2017.

Accepted 13 October 2017.

SUBSTANTIATING TABARI REGIONAL DIALECTS IN CENTRAL-EASTERN ALBORZ VIA AGGREGATE ANALYSIS OF WORD PRONUNCIATIONS

Sayfolah MOLLAYE PASHAYE

Payame Noor University (PNU)*

s_mollaye@pnu.ac.ir

Abstract

Current study applied a computational measure of pronunciation differences to a database of 62 word pronunciations from 425 sites throughout Northern Slopes of central-eastern Alborz Mountains based on *Iran National Dialect Atlases database*. The result is a comprehensive view of the aggregate pronunciation differences among all the sites, which were certified by two measures of validity: *Cronbach's α* (0.81) and *Cophenetic Correlation Coefficient* (0.76). This study aims to contribute, therefore, to Tabari dialectology, as well as to the testing of the computational technique now implemented as a web application (GABMAP), for Iranian Languages; besides, Substantiating traditional dialectology of Māzandarān via contrastive analysis of dialect maps.

Keywords

dialectometry, quantitative analysis; Tabari (Māzandarāni) language, Alborz regional dialects, dialect map

JUSTIFICANDO LOS DIALECTOS REGIONALES DE TABARI EN EL ÁLBORZ CENTRO-ORIENTAL A TRAVÉS DEL ANÁLISIS AGREGADO DE LAS PRONUNCIACIONES DE PALABRAS

Resumen

Este estudio ha aplicado una medida computacional de diferencias de pronunciación a una base de datos que contenía 62 pronunciaciones de palabras en 425 enclaves situados a lo largo de las laderas septentrionales de las montañas del este y centro-este del Alborz basada en la base de datos de los Atlas dialectales nacionales de Irán. El resultado ofrece una visión completa de las diferencias de pronunciación

* Department of Linguistics, Payame Noor University (PNU), P.O. Box 19395-3697, Tehran, Iran.

agregadas entre todos los lugares, que fueron certificadas por dos medidas de validez: α de Cronbach (0,81) y coeficiente de correlación de Cofenética (0,76). Este estudio tiene como objetivo contribuir, por lo tanto, a la dialectología de Tabari, así como a la prueba de la técnica computacional ahora implementada como una aplicación web (GABMAP) para las lenguas iraníes; y además justificar la dialectología tradicional de Māzandarān a través del análisis contrastivo de los mapas de dialectales.

Palabras clave

dialectometría, análisis cuantitativo, lengua Tabari (Māzandarāni), dialectos regionales de Alborz, mapa dialectal

1. Introduction

The systematic study of all forms of dialect, but especially regional dialect, is called "dialectology" or "dialect geography" (Crystal 2008: 143). In recent years, computational techniques enable the incorporation of large amounts of dialectal material into studies of language variation. Kessler (1995) first introduced the use of the Levenshtein distance as a tool for measuring linguistic distances among the pronunciations of language varieties. He applied Levenshtein string edit distance algorithm to the comparison of Irish dialects. Kessler also applied clustering to check whether the edit distances could delineate dialect areas. Nerbonne, Heeringa & Kleiweg (1999) attempts for analyzing Dutch led to introduction of multi-dimensional scaling (MDS) as a means of further analyzing the average pronunciation differences. Later the same techniques were successfully applied to Sardinian (Bolognesi & Heeringa 2002), Norwegian (Gooskens 2004, Heeringa 2004), German (Nerbonne & Siedle 2005), and Azeri (Asadpur 1390 A.P.)[1] varieties. Palander et al. (2003) used cluster analysis to demonstrate variation in a transitional dialect zone in eastern Finland. Prokic et al. (2007) have conducted aggregate phonetic measurements of Bulgarian dialects to find phonetic changes responsible for dialect varieties. Leinonen (2010) describes the geographic variation in vowel pronunciation across the Swedish language area by carrying out an acoustic aggregate analysis of vowels.

[1] In dates given, A.P. denotes the Iranian solar calendar, which is official in Iran and Afghanistan.

Resuming this line of work, current paper investigates the present relations of some Tabari dialects, the nonstandard regional language of the Māzandarān province. The study will apply the computationally inspired Levenshtein distance to measure dialect differences, via automatic processing of entire phonetic transcriptions. In addition to clustering, current study shall submit the distance measurements to consistency analysis, line maps, and classification map. To set up the project, however, was challenging; first, because there was no digitized language material for Tabari dialects; second, because Tabari dialects had not been processed earlier with computational tools; therefore, it is a serious issue to see how well the methods developed primarily for other language families perform for Tabari, an Indo-Iranian language. In this view, two measures of quality are applied, Cronbach's α and Cophenetic Correlation Coefficient.

The structure of the paper is as follows: the next section describes the traditional divisions of Tabari dialects, focusing on east-central Alborz. Section 3 deals with the data source and the preparation of the data. In Section 4 the Levenshtein distance measurement and line maps are presented. Section 5 contribute to further analysis of the distances computed in Section 4 via clustering and classification map. Section 6 proposes conclusions.

2. Tabari Dialect Scholarship

Tabari is the regional language of Māzandarān, a province stretched off the south Caspian Seashores, "with a population of about three million" (Lewis et al. 2003). Modern Tabari, also called Māzandarāni, is known since the first half of the nineteenth century, when several European travelers, scholars, and diplomats undertook the task of documenting the language. The geographical domain of Tabari, roughly within the present administrative boundaries of the province, has remained almost unchanged over the past millennium. It is still spoken in the historical cities as well as in modern industrial centers. Most speakers, however, dwell in a series of loosely knit villages spread over the plains of Māzandarān. They also live in individual mountainous

settlements in the central-eastern Alborz, as far south as the suburbs of Tehran (Borjian 2004, Windfuhr 1989). Among the living Iranian languages, Tabari boasts one of the longest written traditions (from the 10th to 15th centuries), roughly matching that of New Persian. This status was achieved during the long reign of the independent and semi-independent provincial rulers in the centuries after the Arab invasion (Borjian 2006: 9).

In spite of its long history, numerous speakers, and vast regional territory very little scholarship is available on Tabari dialect geography. Tabari linguistics literature has offered different ideas about the geographical distribution of Tabari dialects, mostly relying on geographic criteria and linguist's intuition such as Naǰaf-Zādeh-Bārforush (1368 A.P.), Dehgān (1368 A.P.), and Schmitt (1989). The oldest academic dialect classifications have performed by Jahāngiri (1352 A.P.). He determined the number and geographic distribution of dialects in Māzandarān based on isogloss method. Momeni (1374 A.P.) draw so-called hand-made linguistic map of northern slop of central Alborz Mountain range. The most recent work, however, is *A Dictionary of Tabari* [Farhang-e vāžegān-e Tabari] edited by Nasri-Ashrafi (1381 A.P.). This voluminous publication is in fact a comparative glossary containing lexical units from almost all major urban and rural centers of the region. The glossary itself fills only four volumes (pp. 1-2113). In order to illuminate the geographical domain of Tabari, authors outlined 12 different dialect region (p. XXXVIII); namely 1) East Astarābād; 2) West Astarābād, Myānkāleh peninsula, Behshahr & Galugāh; 3) Sāri & Hezārjarib; 4) Juybār, Qā'emshahr, & Savādkuh; 5) Bābol & Band-e Pey; 6) Amol & Lārijānāt; 7) Nur & Nowshar; 8) Abbāsābād; 9) Chālus; 10) Tonekābon; 11) Damāvand & Qasrān; 12) Eastern Gilān. Except three of these main divisions — i.e. 1st (East Astarābād), 11th (Damāvand & Qasrān), and 12th (Eastern Gilān) regions — almost all other regions fall within spatial domain of current research. The process of establish the database have taken the form of detailed surveys using questionnaires and tape-recorded interviews. In this work, like other traditional dialectology studies, accomplished based on isogloss and dialect boundary notions, regionally distinctive words (distinct in form, sense or pronunciation) were the center of attention, as well as syntax variations.

Dialectologia 23 *(2019), 151-171.*
ISSN: 2013-2247

Figure 1. The map sketches 12 different Tabari dialect regions roughly as maintained in Nasri-Ashrafi (1381 A.P, XXXVIII). Upon new administrative divisions 1st, 11th and 12th dialect regions located out of Māzandarān Province. Almost all other dialect regions fall within spatial domain of the research.

3. The Database

3.1 LAI Language Material

The dialect data was originally extracted from oral interviews of National Project of the Linguistic Atlases of Iran (LAI) database, which cover almost the entire Māzandarān province area. The LAI started in 1974 in the form of a joint endeavor by the then 'Iranian Academy of Language' and 'National Geographic Organization'. The project suspended in 1978 due to sociopolitical revolution in contemporary history of Iran, then, revived in 2001 (Parmun 2007: 14-15). Fieldworkers interviewed NORM (Non-mobile Old Rural Man) speakers in spatial domain of current research through October 2002 to February 2006. Because there was no language material for Sāri and Qā'emshahr included, later the author himself gathered new data in the same procedure to compensate for the missing language material. Finally, there are altogether 425 interviews at research disposal. In the present paper, the author extract words from these interviews that he then compares in pronunciation.

3.2 Digitization

When the researcher started the task, the data was available only in orally-recorded form. Since current research methodology wished to analyze the data computationally, digitization became a very important subtask. The data was converted to the IPA (the alphabet of the International Phonetic Association, see Handbook IPA 2003), because it is used within the Levenshtein-based application the researcher used i.e. GABMAP <http://www.gabmap.nl/>. The digitization step involved phonetic transcription of entire words into tab-separated text file with UTF8 Unicode format. Where it was needed to extrapolate, the researcher always did this conservatively, e.g. using little phonetic detail. Table 1 presents the interpretation of vowel symbols, and Table 2 presents the interpretation of the consonant symbols, to show how current study interpreted the Tabari phonetic transcription system in terms of equivalents in the International Phonetic Alphabet (International Phonetic Association 2003).

symbol	symbol description	Vowel interpretation
i	Latin small letter I	front, close, unrounded, long
ɪ	Latin letter small capital I	immediate front, close-mid, unrounded, short
ɐ	Latin small letter turned A	front, open-mid, half-spread, short
ə	Latin small letter schwa	central, half-spread, short
ɔ	Latin small letter open O	back, open-mid, half-rounded, short
a	Latin small letter A	front, open, spread, short
u	Latin small letter U	back, close, rounded, long
ɒ	Latin small letter turned alpha	back, open, rounded, long
ɔʷ	Latin small letter open O + small W	diphthong
ː	triangular colon	length marker

Table 1. The IPA vowel symbols and their interpretation used in the Tabari language material transcription.

symbol	symbol description	Consonant interpretation
p	Latin small letter P	labial, stop, voiceless, aspirated, fortis
b	Latin small letter B	labial, stop, voiced, aspirated, lenis
t	Latin small letter T	dental, stop, voiceless, aspirated, fortis
d	Latin small letter D	dental, stop, voiced, lenis
c	Latin small letter C	pre-palatal, stop, voiceless, aspirated, fortis
ɟ	Latin small letter dot-less J stroked	pre-palatal, stop, voiced, lenis
ʔ	Latin letter glottal stop	glottal stop, voiceless, aspirated, fortis
s	Latin small letter S	alveolar, fricative, voiceless, fortis
z	Latin small letter Z	alveolar, fricative, voiced, lenis
ʃ	Latin small letter esh	alveolar-palatal, fricative (spread), voiceless, fortis
v	Latin small letter V	labiodentals, fricative, voiced, lenis
f	Latin small letter F	labiodentals, fricative, voiceless, fortis
x	Latin small letter X	uvular, fricative, voiceless, fortis
ɣ	Latin small letter gamma	uvular, fricative, voiced, lenis
h	Latin small letter H	glottal, fricative, voiceless, fortis
ç	Latin small letter C with cedilla	alveolar-palatal, affricative, voiceless, fortis
ʝ	Latin small letter J with crossed-tail	alveolar-palatal, affricative, voiced, lenis
r	Latin small letter R	alveolar, trill, voiced, lenis
m	Latin small letter M	labial, nasal, voiced, lenis
n	Latin small letter N	alveolar, nasal, voiced, lenis
ŋ	Latin small letter ENG	velar, nasal, voiced, lenis
l	Latin small letter L	alveolar, lateral, nasal, voiced, lenis
y	Latin small letter Y	palatal, glide, voiced, lenis

Table 2. The IPA consonant symbols and their interpretation used in the Tabari language material transcription.

3.3 Data Preparation

Current research method relies on transcriptions of entire words, which it took from the recorded interviews as best it could. The study digitized a set of 62 words, which were instantiated in every site. The selected lexical corpus represents different parts of speech (namely nouns, adjectives, and adverbs) and singular word forms. They glean local words in accordance with the so-called semantic fields, such as kinship, luminary body, everyday life, human body, animal husbandry, flora and fauna, agriculture, culture, household, and crafts. Table 3 shows Phonemic transcription of the subset of 62 words, abstracting away from the phonetic variation found among the

dialect sites. Note that the list below is therefore not meant to suggest the range of phonetic variation found in Tabari, nor is it meant to provide phonetic detail about any single variety. For example, final devoicing (e.g. /saɟ/→/sac/ [dog]), which is very common in Tabari dialects, is not represented, nor are vowel reductions (e.g. /dɛccə/→/dɛcə/ [home]).

1	father	/pɪyer/	22	day	/ruz/	43	timber	/çuː/
2	mother	/mɒr/	23	wind	/vɒ/	44	leaf	/ɟɛlɒm/
3	brother	/bərɒr/	24	fire	/taʃ/	45	flower	/ɟɔl/
4	sister	/xɒxər/	25	soil	/ɟɛl/	46	tree	/dɒr/
5	son	/ricɒ/	26	rain	/vɒrəʃ/	47	barley	/ɟɔʷ/
6	daughter	/dɛtər/	27	snow	/varf/	48	peanut	/bɒdɛm/
7	breed	/vaçə/	28	stone	/saŋ/	49	yoke	/ɟəft/
8	uncle(paternal)	/ʔɒmi/	29	eye	/çəʃ/	50	house	/sɛrə/
9	uncle(maternal)	/dɒyi/	30	moth	/tɛc/	51	kiln	/cɛlə/
10	aunt(p.)	/ʔammə/	31	toung	/zəvun/	52	wood	/himə/
11	aunt(m.)	/xɒlɛ/	32	blood	/xun/	53	bread	/nun/
12	wife	/zənɒ/	33	bovine	/ɟɔʷ/	54	cheese	/panir/
13	Husband	/ʃi/	34	goat	/bəz/	55	prayer	/nəmɒz/
14	son inlaw	/xɛʃ/	35	lamb	/varə/	56	masculine	/nar/
15	daughter-in-law	/pɛsərzan/	36	sheep	/mɪʃ/	57	female	/mɒ/
16	grandpa(p.)	/ɟappɒ/	37	wolf	/vərɟ/	58	sugary	/ʃirin/
17	grandpa(m.)	/ɟappɒ/	38	mule	/ɣɒtər/	59	away	/dir/
18	sun	/ʔəftɒb/	39	dog	/saɟ/	60	near	/tan/
19	moon	/muŋ/	40	hen	/cərc/	61	pregnant(man)	/saŋin/
20	star	/ʔɛssɒrə/	41	rooster	/təlɒ/	62	pregnant (animal)	/ʔɒbɛssən/
21	night	/ʃɔʷ/	42	pigeon	/cutər/			

Table 3. Phonemic transcription of the 62 Tabari words, common for all of 425 sites, which formed the database of the study.

3.4 Geographic References

For the present experiment, the researcher sought the transcribed pronunciations of a common set of words throughout northern slopes of Alborz Mountains off the south Caspian seashores in Māzandarān Province. Alborz is a mountain range, 900 km long and 60 to 120 km wide, forms a crescent open to the north in northern Iran. It stretches from the borders of Azerbaijan and Armenia in the northwest to the southern

end of the Caspian Sea, and ending in the east at the borders of Turkmenistan and Afghanistan.

Alborz range divides into three parts. First, the western part located between the valley of the Sefid-rud and the Kandovān pass. Next, the central Alborz, between the Kandovān and Gaduk passes constitutes the highest tract of the chain; the majestic cone of Damāvand, the highest elevation in Iran and in the Middle East, with an estimated altitude of 5,678 meters, dominates all these peaks. Finally, to the east of Telār river, the Alborz has a different physiognomy. It runs from west-southwest to east-northeast, but it is narrower (60-80 km) and, except for the Shāhkuh massif, is composed of interlocking low ridges. The spatial domain of the research mainly locates in northern slopes of central Alborz. Settlement of the Alborz goes back to ancient times, as several discoveries of prehistoric tombs suggest (Morgan 1896). Linguistically the populations of the Alborz form part of the north Iranian group. In the north and in the high valleys of the southern slope, languages of the Gilaki and Māzandarāni type predominate.

In LAI, the sites were selected with respect to two main criteria: maximally complete coverage of the area covered by the atlas, and a representative number of varieties and sub-varieties (Parmun 2007: 28). In order to standardize geographic references, current study determined 425 sites' GIS coordination and locate them by Google Earth software < http://www.google.com/earth/>. See the Map in Figure 2. Its higher elevations, in the Alborz Range forest steppe region, are arid with few trees, but its northern slopes, in the Caspian mixed forests region, are lush and forested. This would explain why there are less mountainous settlement than one would expect; thereupon, sites are not distributed uniformly within spatial domain of the research.

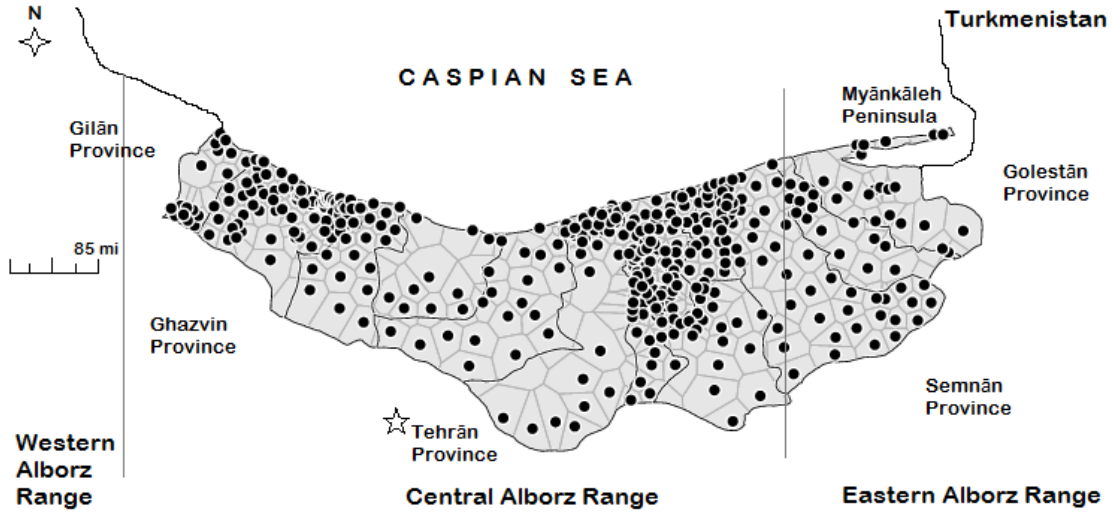

Figure 2. The distribution map of the 425 selected sites on northern slops of central-eastern Alborz Mountain Range in Māzandarān. cf. Figure 1.

4. Measuring Pronunciation Distance

4.1Method

Levenshtein distance is a technique to compare a pair of strings (words) and to assay their distance from each other. Two dialects are compared by comparing the pronunciation of the same words in the two dialects and then averaging the distances of the pairs of words (Osenova et al. 2007: 11). The Levenshtein distance is a numerical value of the cost of the least expensive set of insertions, deletions or substitutions that would be needed to transform one string into another (Kruskal 1999). The simplest technique is phone string comparison. In this approach, all operations have the same cost, e.g., 1. In Kessler's approach when two phones are basically equal but have different diacritics, they are regarded as different phones. So [a] versus [a:] costs 1 unit just as [a] versus [p] costs 1 unit (Heeringa 2004: 18). One may ask how many operations of a simple sort are required to transform one string into another. In the following chart it is illustrated how one dialectal pronunciation of Tabari *klin*, namely [taʃcalə], is transformed into another, [cɛlə]. By writing each derivation step, we can see the operations at work:

t	**a**	**ʃ**	**c**	**a**	**l**	**ə**	
	a	ʃ	c	a	l	ə	delete [t]
		ʃ	c	a	l	ə	delete [a]
			c	a	l	ə	delete [ʃ]
			c	**ɐ**	**l**	**ə**	substitute [ɐ] for [a]
1	1	1	0	1	0	0	=**4**

Since the four operations above are indeed minimal, the distance between the two strings is the sum of the cost of the operations, four. This is naturally rough; in versions that are more sensitive gradual segment distances are used as weights. The segment distances are based on comparison of feature values or acoustic measurements. In fact, Heeringa (2004: 186-194) shows that the phone-based methods outperform most of the methods which are using gradual segment distances as operation weights. The researcher uses this simple version of the Levenshtein algorithm in this paper.

Using the phone string comparison Kessler calculated Levenshtein distances not only when words are phonetic variants of each other, but also when they lexically differ. He called this the all-word approach. However, when he used the feature string comparison, not only the all-word approach was used, but also an approach was used in which the Levenshtein distance is only applied when words are phonetic variants of each other. Kessler called this approach the same-word approach (Heeringa 2004: 18). All-word approach is in process here.

4.2 Results

The initial result of quantitative analysis of pronunciation via Levenshtein algorithm is a “site × site” distance table, in which half of the values simply repeat (due to the symmetry of the distance measures). Given a sample of 62 word pronunciations from two sites, we average the distances of all pairs of corresponding words to obtain an estimate of the aggregate pronunciation difference between any two sites. We repeat this for all ((425 × (425−1)) / 2 = 90,100) pairs of sites. An important issue is

whether the data sample is large enough for us to extract a reliable signal. As a measure of reliability current study used Cronbach's α method, for which a widely accepted threshold is 0.70 (for details see Heeringa 2004: 170-173). The study results show a value of 0.81 for the set of 62 words. The researcher therefore views the data sample large enough to provide a reliable view of pronunciation differences.

The distance table is too big to be inserted here; even though, it would be hard to read and draw any conclusion directly upon it. Instead, beam maps provide an excellent aggregate view of the data. In principle, a line is drawn between each two sites. Darker lines signal pairs of sites that are more linguistically similar to each other. Darkness tonality of the lines depends straight on Levenshtein index in distance table. Particularly coherent areas are normally immediately visible as dark collections, and boundaries appear as lighter-colored swaths.

In Figure 3, the connections between the dialects are shown based on the Levenshtein distances for 62 words. At this interpretive level, a dialect continuum stretches horizontally out from far east through to far west, within which some divisions can be detected. Four main divisions emerge clearly with vertical direction in the map. It therefore implies that they did not maintain any variation between plain and Alborz highland slops. The first dialect border roughly overlaps the central-eastern Alborz border. Actually, if we compare the maps in Fig.3 and 1, we shall conclude that the first division emerged in Fig.3 consists of dialect regions 1, 2, and 3 in Figure 1, which we may name Eastern Tabari Dialect. Then, the largest division in the central part of the plain covers cities of Juybār, Qā'emshahr, Savādkuh, Bābolsar, Bābol, Mahmudābād, and Âmol. Finally, the Next two divisions cover the narrow regions between Alborz and Caspian seashore. The beam map also suggests that the western dialect groups are closer and more coherent than two other divisions. Most speakers, we have to bear in mind, dwell in a series of tight knit villages spread over the plains of Māzandarān. Adjacent distance and contiguous contact, therefore, may explain the recognized coherence.

Dialectologia 23 *(2019), 151-171.*
ISSN: 2013-2247

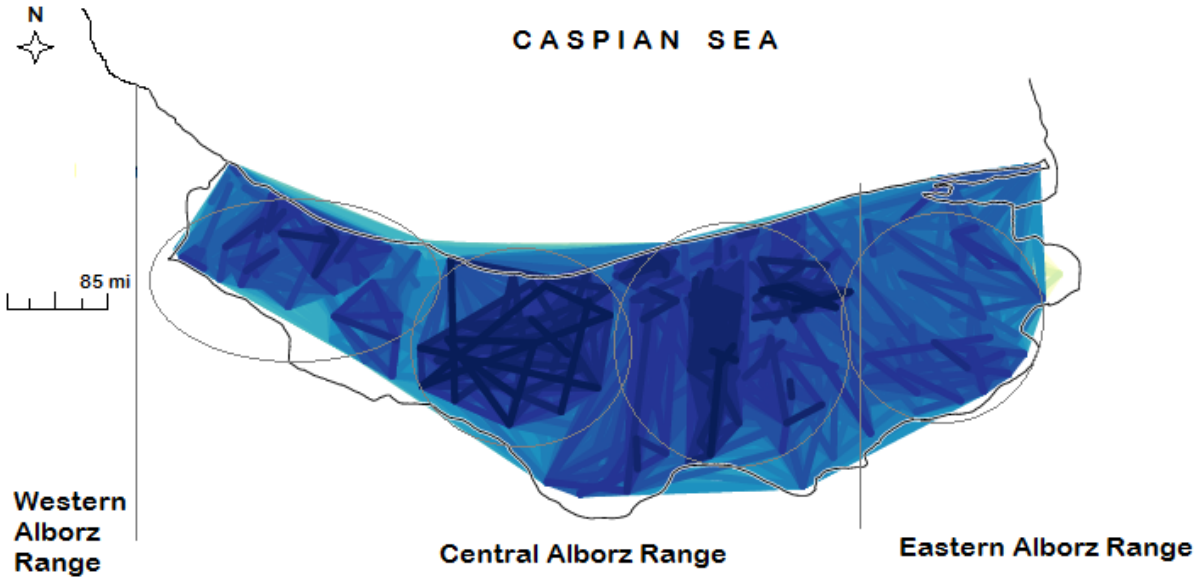

Figure 3. The beam map, which outlines four regional dialects, shows aggregate pronunciation distance of 62 words between 425 sites. The darker the line, the more similar the sites and vies versa. Cf. Figure 1.

Figure 4 outlines a network map of local language varieties based on the Levenshtein distances for 62 words. Network maps connect only adjacent sites; again, darker lines indicate linguistically close varieties, while lighter lines indicate more remote ones. They offer a less complete, but also a clearer illustration of the local linguistic differences measured. Network map of Figure 4 shows three focal areas. The first one locates around Sāri, Qā'emshahr, Bābol, and Âmol urban areas (cf. Figures 1 & 3). Next focal area situates close to the first, in conjunction with Nur urban area. Although the procedure in current study was to select NORMs, the network designation of these two focal areas around urban regions reminds us to take social factors into account. It seems urban area bears testimony to the variations.

Further, other peripheral sites in eastern, southern, and western parts make loosely knit networks. Though one can hardly sketch a focal area westward, there is another focal area, however, in southeast rural highlands of which transitional area stretches westward.

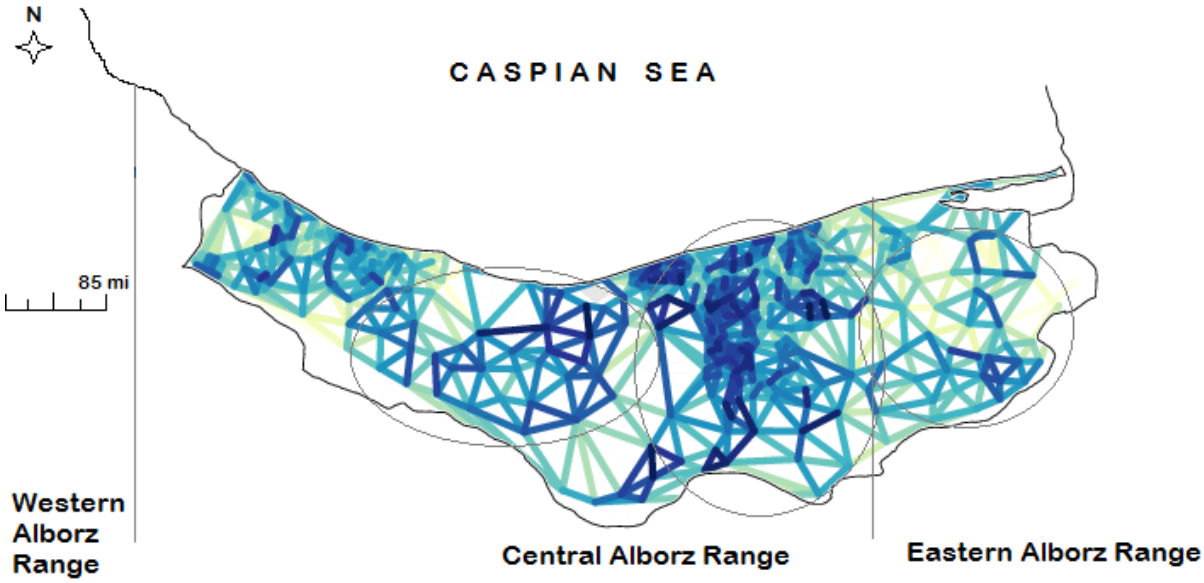

Figure 4. The network map, which outlines 3 focal areas, shows local language varieties between 425 sites based on aggregate pronunciation distance of 62 words. The darker the line, the more similar the sites and vies versa.

5. Cluster analysis

From the two line maps, discussed above, we have already obtained some useful information about the Tabari dialect regions and focal areas, but it is not sufficient for substantiating dialect groups and understanding the amount of their linguistic varieties. For that reason, it is better to continue the analysis by exploring cluster analysis in order to divide the dialects into similarity classes. The resulting classification allows us to compare our findings with the results of dialectological scholarship, which has focused on the identification of dialect areas.

5.1 Method

Clustering is simply the process of dividing a set of elements into groups. In this case, the elements are sites and we want to divide them into groups based on their linguistic similarity. To start with, each site is a cluster of its own. The two sites that have the smallest linguistic distance in the distance table are merged into a new cluster. Then the difference is calculated between that new cluster, and all remaining sites. Based on

the new distances, again, the objects with the smallest difference are merged. It goes on, until all sites are merged into one big cluster. The history of the clustering procedure is displayed in a 'dendrogram'. Dendrogram is a tree incorporating all the elements as its leaves, in which more similar elements are grouped lower in the tree.

A disadvantage of clustering is that it is statistically not stable, meaning that small differences in input data may lead to large differences in results (Kleinberg 2003, Prokić & Nerbonne 2008). We can measure the quality of a clustering result by comparing the distances in the dendrogram (e.g. the number of nodes which have to be traversed to move from one site to another in the dendrogram) to the distances in the pronunciation distance table via a Cophenetic correlation coefficient. This is simply the Pearson correlation coefficient of the two distances (Osenova et al. 2007: 14). The clustering technique that current study used obtained a Cophenetic correlation coefficient (0.76). The dendrogram obtained with this method explains $(0.76)^2 \times 100 = 57.7\%$ of the variance of the original Levenshtein distances.

5.2 Results

Cluster analysis is applied to the distance matrix with the pair-wise aggregate linguistic distances between 425 sites. The dendrogram in Fig.5 shows history of the clustering discriminating six clusters. The scale distance shows average Levenshtein distances as a fraction. The tree structure explains 57.7% of the average Levenshtein distances. Although there are many so-called "matrix updating algorithms" (see Jain & Dubes 1988), we may be content with Weighted Average Method. The leaves on the far left correspond to the individual sites in the sample, which are gradually fused into sub-clusters as one follows the diagram to the root on the right. Six clusters are distinguished in the dendrogram.

A few similar sites that are found to be the most distinct sites are grouped lower in the tree, which surprisingly spot southeast rural highlands. In the following lines we will focus in more detail on this distinct groups, namely the Gāleshi group, via the classification map in Fig.6 resulting from the dendrogram. The next group is composed

of sites located in Chālus, Tonekābon, Rāmsar and few western sites of Nowshar. Finally, the most populated group that covers a noticeable area contains the rest.

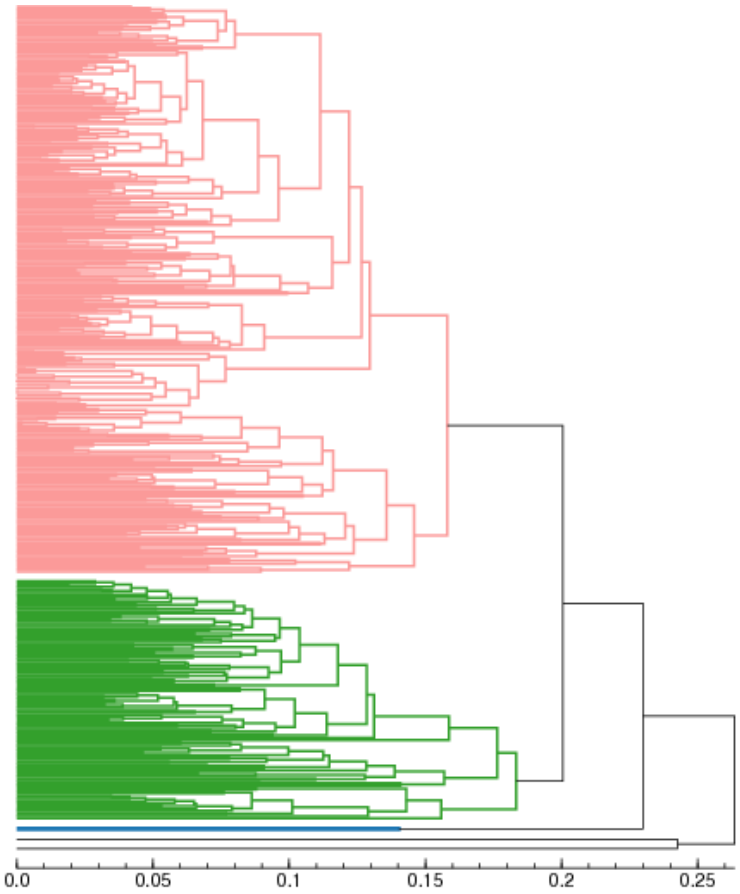

Figure 5. A dendrogram resulting from clustering the aggregate distance table with weighted average method. The scale distance shows average Levenshtein distances as a fraction. The tree structure explains 57.7% of the average Levenshtein distances.

A map can be created by coloring the respective geographic area of each cluster with a distinct color. The colors in these maps are arbitrary. Similarity of colors does not imply linguistic similarity, but each distinct color simply denotes one cluster. The classification map in Figure 6 shows the varieties assigned to dialect groups. It represents the projection of five clusters of the dendrogram in Figure 5 onto central-eastern Alborz geography. In other words, this classification map is a cartography of northern slops of central and eastern Alborz language varieties. The dendrogram on Fig.5 compactly shows the five most significant clusters of Tabari dialects. The colors and numbers that indicate clusters in the dendrogram correspond to the numbered areas in the classification map to facilitate the comparison.

Dialectologia 23 *(2019), 151-171.*
ISSN: 2013-2247

The classification map in Figure 6 suggests that eastern part is a rather uniform area, where only some scattered sites are distinguished. it shows some distinct varieties in the far east, the southern Galugāh county, signaling a relic area. The sites 10, 13, 18, 21 make leaves of the small group lower in the tree, which already elucidated relic areas in beam and network maps. A flash back to database reveals that linguistically naive respondents of respective sites affirm their language variety Gāleshi. According to diachronic, geographic, and dialectological parameters Widfuhr (2009: 14) classifies Gāleshi, Gilaki, Māzandarāni (former name Tabari), and Gorgāni (extinct since 16th century) sub-groups of Caspian Dialects. The sedentary nuclei of Gilakis or Māzandarānis who have resisted the nomads and repopulated the whole range from the north slopes are numerous in the sub-arid central and western Alborz and in the basins at medium altitudes near the Caspian forest (Planhol 1968: 37). Therefore, Gāleshi, spoken by mountain herdsmen along the Alborz, resist change and take his innovation because of physical barriers (e.g. isolated valleys of Alborz) and/or conservative ways of rural life.

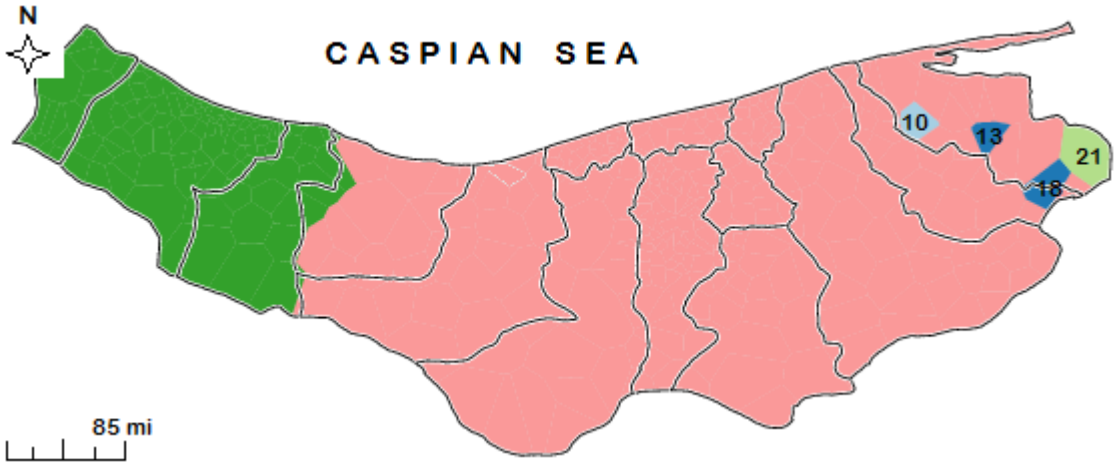

Figure 6: The classification map represents the projection of five most significant clusters of the dendrogram in Figure 5 onto northern slops of central and eastern Alborz geography (cf. Figure 1).

6. Conclusions and Prospects

In this paper, the researcher applied dialectological techniques that had already been successfully applied to Germanic, Romance, and Slavic language to the Tabari (or

Māzandarāni), an Indo-Iranian language; because, the Levenshtein distance is completely objective, and its results are verifiable, an advantage it shares with other computational methods (Heeringa 2004: 24).

Current study analyzed the relations among some Tabari dialects based on the 26350 pronunciation distances calculated via Levenshtein algorithm between all pairs of 425 Tabari varieties. The data includes 62 word pronunciations for all sites, selected and digitized from the LAI database. Though, we should bear in mind the fact that the study procedure is based exclusively on a limited number of word pronunciation, and not all on morphology or the lexicon, in researcher's opinion, this work paints a faithful picture of Tabari dialects. Two measures of validity approve the opinion: Cronbach's α (0.81>0.70) and Cophenetic correlation coefficient (0.76) are certified.

As it has been discussed in section 2, spatial domain of our research traditionally includes different dialect regions. In general, the important distinction between western and eastern dialects is present in almost all Tabari dialect scholarship. In current study findings, the most significant border remains the east-west border. If we compare the beam, network, and classification maps — in Figures 3, 4, and 6 respectively — with the map driven from Nasri-Ashrafi (1381 A.P: XXXVIII) in Figure 1, then we can see aggregate-driven dialect borders approximately coincide traditional borders in vertical direction.

At one hand, the west versus east border, however, is not straightforward. The traditional nine distinct divisions melt into two of aggregate analysis. Although some sub-clusters and peripheral dialect divisions cannot be ignored unnoticed. some small exceptions in eastern region which is a compact group that is distant from the surrounding dialects. This group seems to behave divergently for tending to replace half-open /ɔ/ for close /u/. At the other hand, current research also confirms Borjian's (2006) claims about the need to introduce a north-south or plain-highland division. In other words, the study results suggest that the western dialects are more cohesive than the eastern ones, and that within the western dialects, the northern regions are distinguished from southeast rural highlands (see Figure 4).

The East shows weaker cohesion among dialects, at the same time, Gāleshi variety are also distinguished. One explanation could be that more migrations have

taken place in the East than in the West, especially in new industrial towns. Moreover, Gāleshi variation did not show up in line maps of Figure 3 and 4, which is distinguished sharply within eastern dialects through clustering. The dendrogram in Figure 5 emphasizes the distinction. Scattered Gāleshi sites in Figure 6 — which divides the area almost neatly into two parts — may reflect the fact that a number of migrations have taken place in this area. This group seems to behave divergently with respect to phonetic process, like vowel reductions and consonant shift, as well as Persian and Arabic loan-words. As mentioned before, in vertical direction, rural areas in highlands tend to differ linguistically from plain dialects especially in urban areas.

7. Acknowledgments

The author would like to thank Mazdak Anusheh, Yadollah Parmun, and Faryar Axlaqi (LAI project) for their kind help in language material. I also owe thanks to Therese Leinonen (University of Groningen) for her kind help with the GABMAP, and Peter Kleiweg (University of Groningen) for the software facilities.

References

BOLOGNESI, R. & W. HEERINGA (2002) “De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten”, *Gramma/TTT: tijdschrift voor taalwetenschap*, 9(1), 45-84.

BORJIAN, H. (2004) “Māzandarāni: Language and people (the state of research)”, *Iran and the Caucasus*, 8 (2), 289-328.

BORJIAN, H. (2006) “The Oldest Known Texts in New Tabari: The Collection of Aleksander Chodzko”, *Archiv Orientální*, 74(2), 153-171.

CRYSTAL, D. (2008) *A Dictionary of Linguistics and Phonetics*, Oxford: Blackwell Publishing.

GOOSKENS, C. (2004) “How well can Norwegians identify their dialects?”, *Nordic Journal of Linguistics*, 28(01), 37-60.

HEERINGA, W. (2004) *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, Groningen Dissertations in Linguistics, 46.

IPA (1999) *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* Cambridge: Cambridge University Press.

JAIN, A. K. & R. C. DUBES (1988) *Algorithms for clustering data*, Englewood Cliffs, New Jersey: Prentice Hall.

KESSLER, B. (1995) "Computational dialectology in Irish Gaelic", in *Proceedings of the 7th conference of the European Association for Computational Linguistics*, Dublin: EACL, 60-67.

KLEINBERG, J. (2003) "An Impossibility Theorem for Clustering", *Advances in Neural Information Processing Systems* (NIPS 2002), 15, 463-470.

KRUSKAL, J.B. (1999) "An Overview of Sequence Comparison", in D. Sankoff & J. Kruskal (eds.), *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison,* Stanford: CSLI, V-XV.

LEINONEN, T. (2010) *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects* (Unpublished doctoral dissertation), Groningen: University of Groningen Dissertations in Linguistics, 83.

LEWIS, M. Paul, Gary F. SIMONS & C. D. FENNIG (eds.). (2013) *Ethnologue: Languages of the World,* Seventeenth edition, Dallas, Texas: SIL International.

MORGAN, J. de (1896) *Mission scientifique en Perse IV*, Paris.

NASRI-ASHRAFI, J. (ed.) (1381A.P./2002) *A Dictionary of Tabari*, 5 volumes, Tehran: Eḥyā-e ketāb.

NERBONNE, J., W. HEERINGA & P. KLEIWEG (1999) "Edit Distance and Dialect Proximity", in D. Sankoff & J. Kruskal (eds.), *Time Warps, String edits, and Macro molecules; The Theory and Practice of Sequence Comparison,* Stanford: CSLI, 1-14.

NERBONNE, John & Christine SIEDLE (2005) "Dialektklassifikation auf der Grundlage Aggregierter Ausspracheunterschiede", in *Zeitschrift für Dialektologie und Linguistik*, 72(2), 129-147.

OSENOVA, P., W. HEERINGA & J. Nerbonne (2010) "A quantitative analysis of Bulgarian dialect pronunciation"; *Zeitschrift für Slavische philology*, *66* (2), 425-458.

PALANDER, M. & L. L. OPAS-HANNINEN & F. TWEEDIE (2003) "Neighbours or enemies? Computing Variants Causing Differences in Transitional Dialects", Computers and Humanities, 37,: 359-372.

PARMUN, Y. (2007) *The National Project of the Linguistic Atlas of Iran*, Tehran: ICHHTO.

PLANHOL, X. (1968) "Elbourz et chaînes pontiques: deux franges montagneuses du Proche-Orient**",** *Acta geographica*, 71 (Janvier-Mars), 11-13.

PROKI´C, J. (2007) "Identifying Linguistic Structure in a Quantitative Analysis of Dialect Pronunciation", *Proceedings of the Student Research Workshop*, Prague: Association for Computational Linguistics, 61-66.

PROKIĆ, Jelena & John NERBONNE (2008) "Recognizing groups among dialects", International *Journal of Humanities and Arts Computing*, Special Issue on Language Variation, 2(1-2), 153-172-

SCHMITT, R. (ed.) (1989) *Compendium Linguarum Iranicarum*, Wiesbaden: L. Reichert.

WINDFUHR, G. L. (1989) "New Iranian languages: Overview", in R. Schmitt (ed.), *Compendium linguarum Iranicarum,* Wiesbaden: L. Reichert, 246-249.

WINDFUHR, G. (2009) *The Iranian Languages*, London: Routledge.