Linguistically Mapping Aśoka: A Dialectometric Approach to the Major Rock and Major Pillar Edicts

Patrick Zeitlhuber
University of Vienna
patrick.zeitlhuber@gmail.com

Abstract

The edicts of the emperor Aśoka were inscribed in different Middle Indo-Aryan language varieties as well as Greek and Aramaic in the 3rd century BCE. These Middle Indo-Aryan varieties have been variously categorized into three or four dialect groups. In this paper, these classifications are reassessed by applying methods of dialectometry. Dialectometry is a branch of quantitative linguistics which aims at measuring the differences between languages and language varieties. I will examine Aśoka's Major Rock and Major Pillar Edicts by calculating the Levenshtein distance and aggregating the results. This is followed by hierarchical clustering and multidimensional scaling in order to determine the most suitable grouping of these language varieties. After triangulating the results, the dialect classification will be projected onto a geographical map, therefore showing the clear regional distribution of these dialect groups.

Keywords: Aśoka, dialectometry, quantitative linguistics, Middle Indo-Aryan, inscriptions

1 Introduction

The edicts of the emperor Aśoka constitute the earliest extant (decipherable) evidence of written culture in South Asia. They were issued in the years after Aśoka's coronation, which is commonly dated to 268 BCE. Strikingly, these edicts were not inscribed in Sanskrit, but in different language varieties of Middle Indo-Aryan (MIA) as well as Greek and Aramaic and served diverse purposes and functions.

Figure 1 shows a map of the 42 edict sites,[1] which extend over the territory of the four modern states of Pakistan, India, Nepal, and Afghanistan.[2] Of these, the inscriptions in Afghanistan as well as in Taxilā in Pakistan were written in either Greek, Aramaic or both. 174 edicts were composed in various MIA language varieties. Commonly, these inscriptions are divided into Major Rock Edicts, Minor Rock Edicts, Major Pillar Edicts, Minor Pillar Edicts, Cave Sites, and various edicts (comprising the Pāṅgurāria Separate Pillar Edict and the Bhābrū Stone Inscription).

The MIA varieties in the inscriptions show clear dialectal differences. That becomes even more obvious as the edicts are often transfers of a certain text from one variety into another. The original language is certainly an administrative language from the Eastern part of Aśoka's realm (Oberlies, 2003, 165-166). Schneider (1978) even attempted to reconstruct the original archetype of the Major Rock Edicts in this administrative language.

---

[1] A list of the abbreviations for all the Aśokan edict sites is provided in appendix A.

[2] The maps in this paper were created with `QGIS`, which is free and open-source. The coordinates for the Aśokan sites were taken from Falk (2006; 2013) and verified and amended, if necessary, with Google Maps. For better geographical referencing, river courses have been marked, which, however, represent the modern conditions.

At least to my knowledge, the MIA language of the Aśokan edicts has no commonly accepted denomination apart from "Aśokan inscriptional language" and similar ones. In this paper, I opt to call them—in accordance with the nomenclature of most other MIA languages—*Āśokī*.

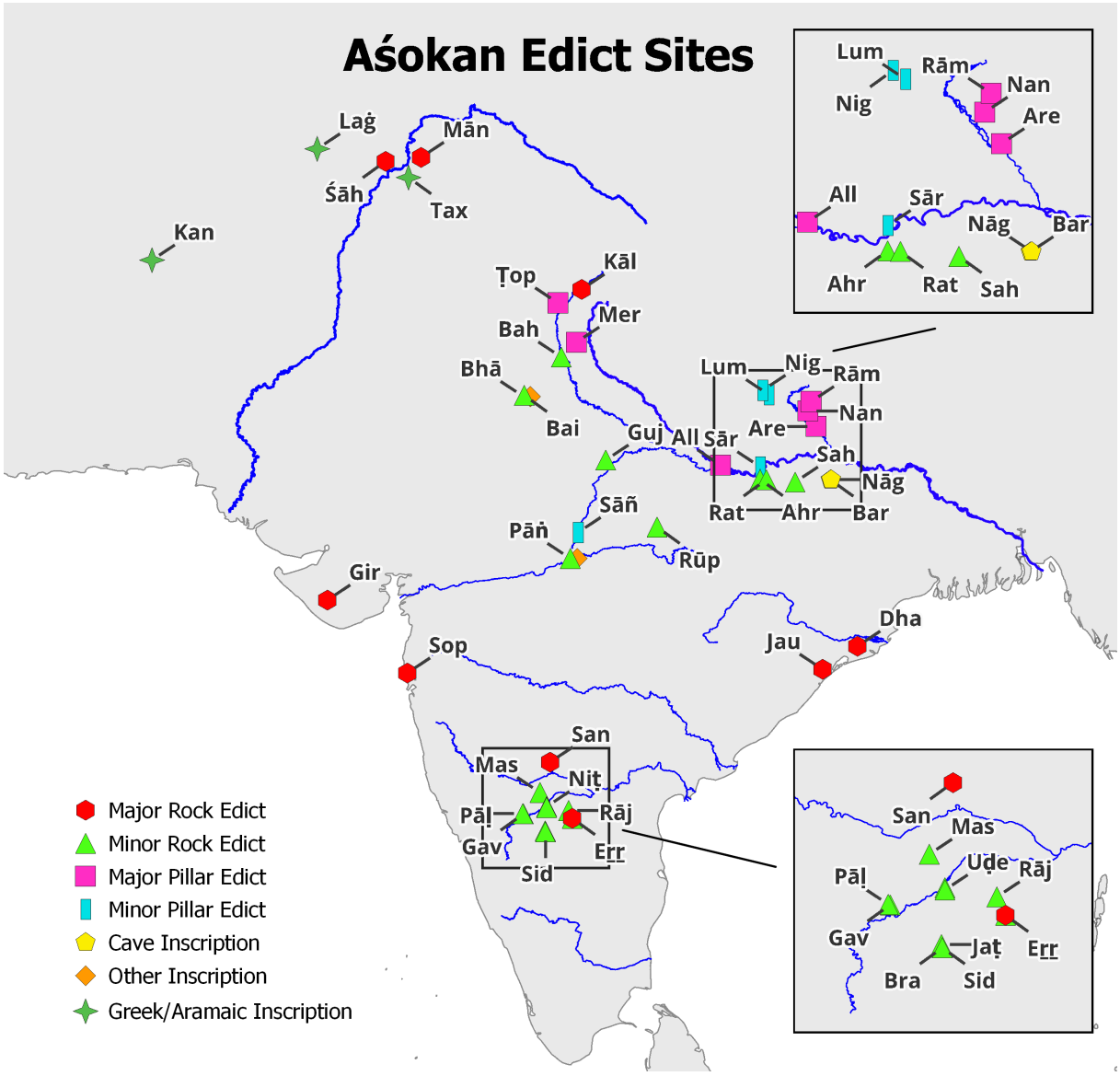

Figure 1: map of the Aśokan edicts

The classifications of the Āśokī varieties in the literature are based on qualitative linguistic analysis and vary between three and four dialect groups. Researchers picked certain linguistic characteristics which they deemed representative and distinctive, on which grounds they grouped these language varieties together. The following tables provide a sample of these different classifications.

Table 1 shows the dialect assessment by Salomon (1998, 73-76) and Oberlies (2003, 165). Both describe three dialects with the same members. The Northwestern group is constituted by Śāh and Mān, the Western by Gir and Sop, and the Eastern by Kāl, Dha, Jau, and Err (and for Oberlies also together with all the other Āśokan inscriptions).

	Northwestern	**Western**	**Eastern**
Salomon (1998)	Śāh, Mān	Gir, Sop	Kāl, Dha, Jau, E<u>rr</u>
Oberlies (2003)	Śāh, Mān	Gir, Sop	Kāl, Dha, Jau, E<u>rr</u> "all other rock edicts, pillar edicts" (p. 165)

Table 1: classifications into 3 dialects

Four dialects with slightly different members are postulated by Sen (1960, 7–11), Misra & Misra (1982, 9-10), and Bubeník (1996, 8), which can be seen in table 2. The names of the dialects are slightly different with each of these researchers. The terms in the first row before the comma are used by Sen and Misra & Misra, after the comma by Bubeník. All of them agree that the Northwestern (North-West) group is formed by Śāh and Mān and the Southwestern (West) by Gir. They disagree, however, on the exact classification of the other inscriptions. Unfortunately, Sop and E<u>rr</u>, which form a separate dialect for Bubeník, are not mentioned by Sen and Misra & Misra. Bubeník groups those two together, whereas they are clearly separated by Salomon and Oberlies. Otherwise, Sen and Misra & Misra separate Kāl, Dha, and Jau, which Bubeník considers members of the same dialect.

	Northwestern, North-West	**Southwestern, West**	**Middle Eastern, South/South-West**	**Eastern, Center/East**
Sen (1960)	Śāh, Mān	Gir	Kāl, Ṭop, Nāg[3]	Dha, Jau "all the Minor Rock Edicts and Pillar Edicts, the Cave Inscriptions" (p. 11)[4]
Misra & Misra (1982)	Śāh, Mān	Gir	Kāl, Ṭop, Nāg[5]	Dha, Jau
Bubeník (1996)	Śāh, Mān	Gir	Sop, E<u>rr</u>	Kāl, Dha, Jau

Table 2: classifications into 4 dialects

In this paper, I will apply quantitative methods to reassess these dialect classifications. To be precise, I will draw upon the methods of dialectometry, which is a well-established methodology, first and foremost in Romance and German(ic) variationist linguistics. Dialectometry was devised in the 1970s and 80s out of a desire to reassess prevailing dialect classifications, which were often based on small sets of subjectively chosen linguistic features, neglecting the major part of the concerned varieties. Nerbonne (2009, 177) summarizes it by stating:

> By focusing exclusively on single features or small combinations of these, variationists, including dialectologists, sometimes fail to isolate signals of provenance clearly. The signals are often so complex, even misleading, that they resist analysis using simple, single-featured methodologies.

[3]Sen also mentions the Jogīmara Cave inscriptions here. Even though they are from the Mauryan period, they are not Aśokan (Salomon, 1998, 76).

[4]Sen also mentions the Mahāsthān stone plaque inscription, the Sohgaurā copper-plate inscription, and the Hāthīgumphā inscriptions of Khāravela here. The former two are from the Mauryan period, but they are not Aśokan. The latter are from the Śuṅga period even (Salomon, 1998, 76, 142).

[5]Misra & Misra follow Sen in listing the Jogīmara Cave inscriptions here.

The aim of dialectometry is to make language measurable. This is achieved by quantifying linguistic differences either between dialects and varieties of one language or between related languages. By and large, two major schools of thought can be differentiated: the "Salzburg school" established by Hans Goebl (2010) and the "Groningen school" centered around John Nerbonne (2010). The main differences regard certain epistemological and methodological approaches. Both schools work predominantly with data from linguistic atlases.

The Salzburg school takes data and taxates it according to linguistic phenomena on the levels of phonetics, morphology, syntax, and lexicon. The similarity between these taxates is calculated with different algorithms (Goebl, 2010). The Groningen school uses predominantly the Levenshtein distance either in its original version or with various modifications. A taxation is not necessary but the data need to be arranged so only appropriate linguistic items are compared (Nerbonne, 2010). Common to both schools is that the attained measurements are arranged in a distance (or similarity) matrix which is the basis for further analyses. Hierarchical clustering has proven to be a useful method for both. The results are then projected onto a geographical map. Proponents of the Salzburg school create maps by using different clustering methods, Euclidian proximity, skewness, arithmetic mean, standard deviation etc. (Scherrer and Stoeckle, 2016; Goebl, 2010). Apart from clustering, the Groningen school applies multidimensional scaling and bipartite spectral graph partitioning (Nerbonne, 2010; Heeringa, 2004; Wieling and Nerbonne, 2011).

Especially in German variationist linguistics, different dialectometric approaches have been developed which are not based on distance matrices, e.g. factor analysis and principal component analysis (Pickl and Pröll, 2019). Of course, the methods of dialectometry are not restricted to horizontal variation, i.e. language variation in space. It is also possible to measure differences and distances between vertical varieties like dialects, regiolects, and standard language (Kehrein, 2012). This is an all but exhaustive list of all the different approaches that have been applied in dialectometry.

2 From the Data to the Map

2.1 Data Preparation

In order to determine the number of Āśokī dialects, I chose dialectometry as a viable methodology. After the digitization of the relevant data, I arranged the texts of the Major Rock Edicts (MaRE) and Major Pillar Edicts (MaPE) in a data frame as correspondence sets so each row equals one location and each cell in a column contains all variants of a certain variable from that location (see table 3 as an example). As the name of the variable has no influence on the distance measurements, the Sanskrit equivalent of the MIA wordforms serve as a reference point. Multiple variants are indicated by using | as delimiter.

The MaREs of Sopārā and Sannati had to be excluded as these are only extant in fragments. For the remaining MaREs and the MaPEs, I selected all the wordforms that are attested in at least 75 % of these edict sites. This is less based on statistical reasons than on practicality. This approach led to 66 wordforms that were suitable for further comparison. The next step was the philological and linguistic interpretation of these tokens.

Apart from Śāh and Mān, which were inscribed in Kharoṣṭhī, all the edicts were written in Brāhmī. The Aśokan Brāhmī script indicates vowel length but not geminate consonants. Moreover, anusvāra is often omitted. Judging from the inscriptions, the law of two morae (von Hinüber, 2001, 117-118) had already had its effect on the Āśokī dialects. In contrast to Old Indo-Aryan (OIA), in MIA no long vowel could precede a geminate or a consonant cluster. OIA VCC resulted either in MIA VC or VCC.[6] Both possibilities are attested in different lexemes at different Aśokan sites. For the linguistic interpretation of the data, this means that whenever a word in the Brāhmī edicts was written with a short vowel followed by a single consonant sign and this particular form can be traced back to OIA VCC, it can safely be assumed that

[6] V = long vowel, V = short vowel, C = consonant

Skt	bhavati	rājā	kariṣyanti
All	-	lājā	kacchanti
Are	hoti	lāja|lājā	kacchanti
Dha	hoti	lājā|lāja	kacchanti
Err	hoti|hoti	lāja|lājā	kacchanti
Gir	bhavati|hoti	rājā	kāsanti|kassanti
Jau	hoti	lājā	kacchanti
Kāl	hoti	rājā|lājā	kacchanti
Mān	hoti|bhoti	rājā	kaṣṣanti
Mer	hoti	lājā|lāja	-
Nan	hoti	lāja	kacchanti
Rām	hoti	lāja	kacchanti
Śāh	bhoti|hoti	rāyā|rājā	kaṣṣanti
Ṭop	hoti	lājā|lājā	kacchanti

Table 3: example correspondence set

that consonant has to be understood as a geminate. However, the opposite may also be the case when a consonant cluster following an etymologically short vowel got simplified and the vowel underwent compensatory lengthening, e.g. OIA *varṣeṣu* > *vāsesu* (Gir MaRE03), next to *vassesu* (Kāl MaRe03). Especially the variety of Gir was very prone to this kind of sound change.

Yet another difficulty presents itself with regard to the Kharoṣṭhī inscriptions of Śāh and Mān. Just like Brāhmī, geminates were not indicated and anusvāra often omitted (or sometimes added in unetymological positions). Apart from that, Aśokan Kharoṣṭhī does not designate the quality of vowels. When it comes to sound clusters like OIA VCC or VCC, they appear in Kharoṣṭhī as VC. It is, therefore, impossible to tell whether the vowel was shortened or the consonant degeminated. In agreement with the phonetic interpretations in the "Dictionary of Gāndhārī" on `gandhari.org` (Baums and Glass, 2002 ongoing), these cases were treated as retaining the etymological vowel length.

Another peculiarity worth mentioning are the inscriptions from Kāl. In these, the signs for *s*, *ṣ*, and *ś* are used without any clear distinction. Bubeník (1996, 9) claims, "The three sibiliants of OIA survive [...] to a certain degree in the Center (Ka)", i.e. Kāl. I tend to disagree with this statement. Sometimes the sibilant signs appear in etymological positions but in most cases there is no obvious reason. It is likely that the scribe considered these signs to be graphical variants and used them indiscriminately or according to taste to represent one and the same sibilant phoneme.

Further challenges for the linguistic interpretation concern scribal errors, orthographic peculiarities, and inconsistent spellings. Moreover, it is imperative that only cognates are compared with each other which will be elaborated on in the next section.

2.2 Distance Measurement

For the calculation of the linguistic distances between the language varieties of the Aśokan sites, the package `dialectR` for the software `R` (Shim and Nerbonne, 2022) was utilized. The function `distance_matrix` allows the creation of a distance matrix by applying the Levenshtein distance.

The Levenshtein distance (or: edit distance) measures the number of modifications that are necessary to transform one string into another by either insertion, deletion, or substitution (Kruskal, 1983, 215-219). It is the main method of measurement used by the Groningen school of dialectometry. Nerbonne (2010, 481) states that "[e]arlier work in dialectometry analyzed the data at a nominal level, where each pair of linguistic items was measured as the same or different, while the application of Levenshtein distance allows numeric characterizations per

pair of pronunciations to be obtained." Discussing the advantage of this kind of measurement, Heeringa (2004, 24) argues that "[t]he Levenshtein distance is completely objective, and its results are verifiable, an advantage it shares with other computational methods, in contrast to dialect maps based on tribes and intuition", if "the data used consists of representative samples of the varieties."

Another important aspect regarding distance measurements is the fact that only cognates in different varieties should be compared. It would be possible to use the Levenshtein distance to calculate the difference between two words that are etymologically unrelated. However, this would yield methodologically and epistemologically incorrect results. Therefore, it is a prerequisite that already the data preparation is carried out with sound philological and linguistic knowledge.

In order to illustrate a measurement with the Levenshtein distance, the variable *YATHĀ* will serve as an example in table 4:

Gir (MaRE12)	y	a	th	ā	
E<u>rr</u> (MaRE12)		a	th	a	
	1			1	= 2

Table 4: Levenshtein distance example

Two modifications are necessary to get from *yathā* to *atha*: the deletion of word-initial *y* and the substitution of *ā* by *a*. Hence, the absolute number of changes is 2. Yet, the parameters for the function `distance_matrix` can be set to normalize the length of strings so the penalty of the modification is calculated in relation to the total number of characters by setting `alignment_normalization = TRUE`. In the example above, this means these 2 modifications are divided by the sum of the string length of 4, which equals a relative difference of 0.5.[7]

Of course, these are still only two variants. For a useful distance measurement, a matrix needs to be calculated that compares all the variants of a certain variety with all the variants in every other variety for every variable. Herein lies the value of dialectometry as it is not based on certain single features but it combines all the distance values of all the variables in all the varieties. This step is called aggregation (Nerbonne, 2010).

	All	Are	Dha	E<u>rr</u>	Gir	Jau	Kāl	Mān	Mer	Nan	Rām	Śāh	Ṭop
All	0,0	2,8	3,6	3,9	10,4	3,3	6,1	10,9	1,6	2,9	3,1	13,9	2,5
Are	2,8	0,0	5,3	5,7	13,6	4,4	8,2	12,7	2,7	0,6	0,6	16,4	4,0
Dha	3,6	5,3	0,0	4,8	11,9	2,6	6,2	11,2	3,6	5,1	5,2	14,9	3,8
E<u>rr</u>	3,9	5,7	4,8	0,0	12,9	4,2	5,9	12,0	4,7	5,6	5,5	15,5	4,8
Gir	10,4	13,6	11,9	12,9	0,0	10,9	14,4	13,2	12,3	13,6	13,8	12,3	13,0
Jau	3,3	4,4	2,6	4,2	10,9	0,0	5,2	10,0	3,6	4,3	4,6	13,4	4,1
Kāl	6,1	8,2	6,2	5,9	14,4	5,2	0,0	12,1	5,3	8,0	8,6	15,7	6,2
Mān	10,9	12,7	11,2	12,0	13,2	10,0	12,1	0,0	9,8	12,3	12,7	6,8	12,5
Mer	1,6	2,7	3,6	4,7	12,3	3,6	5,3	9,8	0,0	2,4	2,6	12,8	2,4
Nan	2,9	0,6	5,1	5,6	13,6	4,3	8,0	12,3	2,4	0,0	0,5	16,4	3,7
Rām	3,1	0,6	5,2	5,5	13,8	4,6	8,6	12,7	2,6	0,5	0,0	16,6	4,1
Śāh	13,9	16,4	14,9	15,5	12,3	13,4	15,7	6,8	12,8	16,4	16,6	0,0	16,5
Ṭop	2,5	4,0	3,8	4,8	13,0	4,1	6,2	12,5	2,4	3,7	4,1	16,5	0,0

Table 5: distance matrix of 13 MaREs and MaPEs

[7]The dental voiceless aspirate is represented by the digraph *th* in this illustration but in the calculation it is considered one element to reflect its phonological status. As one of the anonymous reviewers pointed out, due to the use of IAST aspirates like *th* are treated by the Levenshtein algorithm as two string elements not reflecting the phonological status of aspirates. Even though the examples in this paper are presented in IAST, for the calculation I have resorted represent aspirates with capital letters and non-aspirates with lowercase letters.

In this manner, the Levenshtein distance was calculated for the Āśokī data set containing 66 tokens from 13 locations with MaREs and MaPEs.[8] As a result, a distance matrix is created by calculating the Levenshtein distance between all the variants at a certain location with all the variants at another location for every variable. Table 5 represents the resulting distance matrix but for a more concise display the values were rounded to one decimal.

It is rather difficult to make sense of a plain distance matrix. Hence, further processing of the distance measurements is necessary. Certain methods of illustration have proven useful, first and foremost creating dendrograms on the basis of hierarchical clustering (section 2.3) as well as multidimensional scaling (section 2.4). The results from both approaches can be used to create maps (section 2.5).

2.3 Hierarchical Clustering

To project the linguistic distances onto a map, it is necessary to reduce the distance matrix to a value matrix (Scherrer and Stoeckle, 2016, 101). One means to accomplish that is clustering, whereby one of the most frequently used approaches is agglomerative hierarchical clustering.

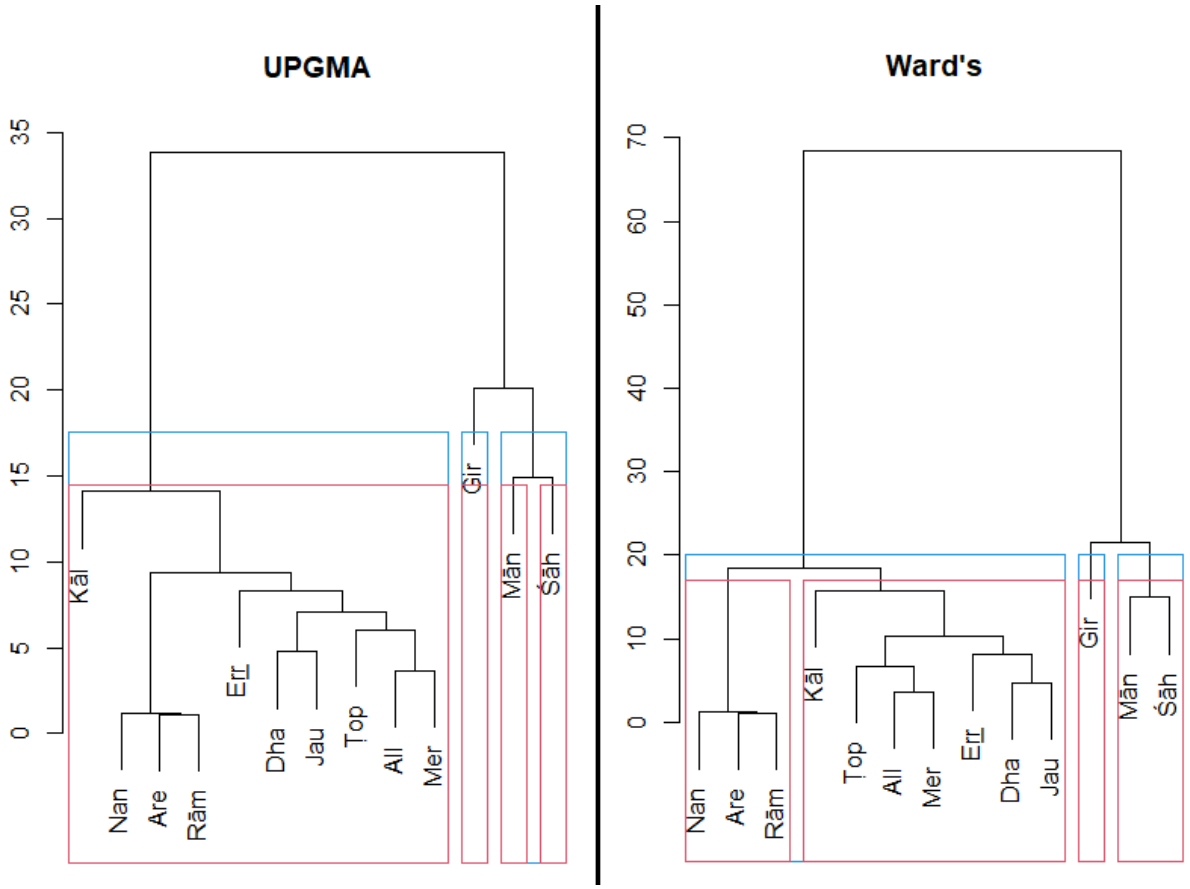

Figure 2: hierarchical clustering with 3 (blue) and 4 (red) clusters

`R` provides the built-in function `hclust` for that. For this paper, the agglomeration methods of UPGMA (= unweighted pair group method with arithmetic mean; called `"average"` in `R`) and Ward's minimum variance method (called `"ward.D2"` in `R`) were chosen in order to be able to compare the validity of the results. Most frequently, the results obtained by hierarchical clustering are plotted as a dendrogram.

From a linguistic point of view, a dendrogram allows for a grouping of dialects. The branches show which varieties are linguistically closer to each other. The distances between the branches

[8] The parameters in `R` were set like this: `distance_matrix(dataset, funname = "leven", alignment_normalization = TRUE, delim = "|")`.

give valuable clues about the most suitable dialect categorization.

Figure 2 shows a dendrogram based on UPGMA to the left and Ward's method to the right. What can be clearly seen from both plots is that there are two major linguistic clusters. One is constituted by Gir, Mān, and Śāh, the other by Kāl, Nan, Are, Rām, Ṭop, All, Mer, Err, Dha, and Jau. The blue lines indicate which locations are grouped together when the number of clusters is set to be three. Even then, the ten locations on the left from Kāl to Mer form one cluster, Gir alone a second, and Mān with Śāh a third. These clusterings hold true with UPGMA as well as Ward's.

Strikingly, the agglomeration into four clusters, illustrated by the red lines, yields different results depending on the selected method. With UPGMA, it does not lead to a subdivision of the location from Kāl to Mer. It rather assigns Gir, Mān, and Śāh to separate clusters each.

When selecting four clusters with Ward's method, however, Gir constitutes one of its own, while Mān and Śāh remain in one cluster together. Nan, Are, and Rām are grouped together and separated from another cluster comprising Kāl, Ṭop, All, Mer, Err, Dha, and Jau.

Consequently, the grouping of the language varieties of these locations into three clusters seems valid as both agglomeration methods agree in this respect. The subdivision into four clusters remains questionable, however. Shim & Nerbonne (2022, 23) state that hierarchical clustering methods are rather unstable and need to be validated with other methods, e.g. multidimensional scaling. In the following section, two different forms of multidimensional scaling will be applied.

2.4 Multidimensional Scaling

Nerbonne (2010, 487) describes multidimensional scaling (MDS) as "a statistical technique aimed at representing very high dimensional data in a smaller number of dimensions." This is accomplished by assigning the calculated distance values points in a coordinate system, usually either in two or three dimensions. These coordinates yield a plot that can be read as a linguistic map that depicts the linguistic distances each and every point has from the other (Embleton et al., 2013, 14). To put it simply, MDS is one form of graphical representation of a distance matrix.

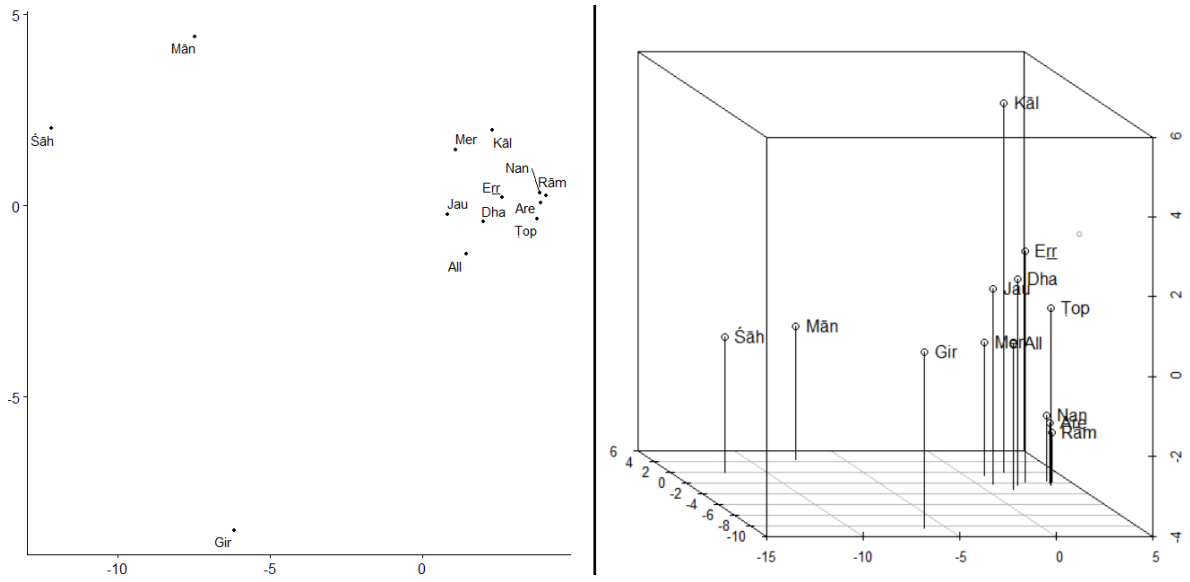

Figure 3: 2D and 3D multidimensional scaling

The left plot in Figure 3 shows a projection of MDS onto two-dimensional space. As with hierarchical clustering, Kāl, Nan, Are, Rām, Err, Dha, Jau, Ṭop, All, and Mer are very close to each other. Gir as very far down. Even though Śāh and Mān are nearer to each other than to any other variety, they show nevertheless some considerable linguistic differences.

The 3D illustration on the right side of Figure 3 depicts the same distances on the horizontal

axis but it gives more details about the distances between points that are very close to each other. This is in accordance with the dendrogram in Figure 2, in which Kāl is the highest point on this branch and Nan, Are, and Rām the lowest. The distances between the cluster to the right still remain the same to Gir, Śāh and Mān.

Compared with the dendrogram in Figure 2, it can be claimed with certainty that Kāl, Nan, Are, Rām, Err, Dha, Jau, Ṭop, All, and Mer form a cluster and, therefore, constitute one dialect group. Gir is set so far apart from any other language variety, hence, it must be assumed that it forms a dialect of its own.

Not as straightforward is the classification of Śāh and Mān. The notion of these two as individual dialect groups would be supported by hierarchical clustering with UPGMA, not by Ward's method though.

In both two- and three-dimensional scaling, Nan, Are, and Rām appear rather close to Kāl, Ṭop, All, Mer, Err, Dha, and Jau. Consequently, it does not seem advisable to divide these varieties into separate clusters as suggested by Ward's method in Figure 2.

2.5 Mapping Linguistic Distances

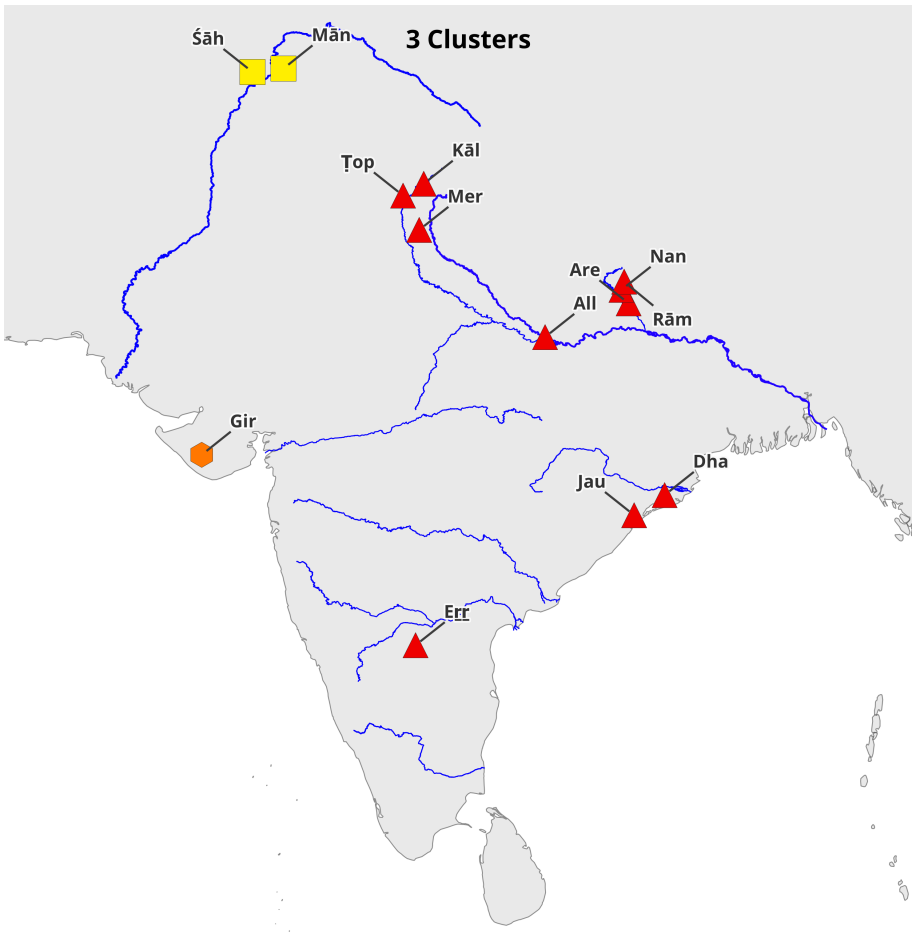

Figure 4: map with 3 clusters of Āśokī

Based on the results of hierarchical clustering and MDS, it is sensible to divide the language varieties of the Major Rock and Major Pillar Edicts into three clusters. These dialect groups can be projected onto a geographical map in order to illustrate the geographical dimensions. For these purposes, the coordinates of the Aśokan sites and the clusters obtained by the above-described methods were imported into `GQIS`, which allowed for the creation of this map of linguistic clusters (Figure 4).

Seeing the three dialect clusters in geographical space, it becomes clear that there is an obvious relation between geographical and linguistic distance. I will, consequently, follow Salomon and Oberlies in referring to the dialects according to their geographical provenance. Northwestern Āśokī (yellow squares) is constituted by Śāh and Mān although there is some variation between these two varieties. Gir is clearly set apart and is the only representative of Western Āśokī (orange hexagon). The varieties of Kāl, Nan, Are, Rām, Ṭop, All, Mer, E<u>rr</u>, Dha, and Jau form Eastern Āśokī (red triangles), which is the best and most widely attested dialect.

Circling back to Figure 3, the coordinates assigned by MDS to the Aśokan sites based on the distance matrix mirror the geographical distribution of the dialect groups as a whole on the map, but the distribution of individual varieties is different. The distances of the Northwestern, Western and Eastern dialect in the MDS plots more or less reflects their geographical distance. Even though this is mere coincidence, it is a striking one indeed.

3 Discussion

With regard to Table 1, the dialect classification of Salomon (1998) and Oberlies (2003) can be affirmed. The grouping in Table 2, however, does not seem to be valid compared with the dataset for which the measurements in this paper were made. Even with the two different options of four clusters in Figure 2, there is no reason for separating Kāl from Dha and Jau as Sen (1960) and Misra & Misra (1982) suggested, nor E<u>rr</u> from Kāl, Dha and Jau as proposed by Bubeník (1996). There may be arguments for this division with a focus on single linguistic characteristics. Based on the dataset for this paper, there is no evidence for this differentiation from the point of view of dialectometric aggregation. It is possible, however, that these results might change with an expanded data set containing more wordforms and linguistic phenomena.

The wide prevalence of Eastern Āśokī creates an epistemological issue. As Salomon (1998, 75) pointed out:

> But it must also be understood that they [i.e. the Aśokan inscriptions] do not provide anything like a real dialect map of the time. For the geographical distribution of the dialects—especially of the eastern dialect—can hardly correspond with linguistic reality; the eastern dialect was obviously not the mother tongue of residents of the far north and the central south, though it was used for inscriptions (Kālsī, E<u>rr</u>aguḍi, etc.) in those regions.

Hence, I want to emphasize that the aim of this paper is not to present a dialect map of the 3rd century BCE. Figure 4 is supposed to be a map of a linguistic clustering of the language varieties used in the Aśokan inscriptions regardless of whether or not they are an authentic reproduction of speech habits of speakers of that time.

Still, this study has some limitations. The data set with 66 tokens is rather small. Due to the fact that the Levenshtein distance is applied to whole strings, it was necessary to include only those wordforms that are attested on all or most of the sites. To base the analyses on a broader linguistic foundation, it will be necessary to utilize some other kind of comparison—perhaps plain word stems (which comes with its own challenges).

Another option would be to chose an approach like Goebl (2010) and taxate the data according to linguistic phenomena. Methods not relying on distance matrices like the ones Pickl & Pröll (2019) use might be a viable endeavour. In the future, other methods of mapping will be explored like multidimensional scaling maps (Nerbonne, 2010). Furthermore, it is my desideratum to expand the dialectometric approach to include all the MIA Aśokan edict sites.

Acknowledgements

I want to thank Dr. Philipp Stöckle for providing his expertise and patiently answering all my methodological and technical questions. I am also grateful to Prof. Alexandra N. Lenz and the Austrian Academy of Sciences for giving me the time to work on this paper. And a big thank you goes to Assoc. Prof. Hannes A. Fellner for his on-going support.

References

Stefan Baums and Andrew Glass. 2002-ongoing. A dictionary of Gāndhārī (online). `https://gandhari.org/dictionary`.

Vít Bubeník. 1996. *The structure and development of Middle Indo-Aryan dialects*. Motilal Banarsidass, Delhi.

Sheila Embleton, Dorin Uritescu, and Eric S. Wheeler. 2013. Defining dialect regions with interpretations: Advancing the multidimensional scaling approach. *Literary and Linguistic Computing*, 28(1):13–22.

Harry Falk. 2006. *Aśokan sites and artefacts. A source-book with bibliography*. Number 18 in Monographien zur indischen Archäologie, Kunst und Philologie. Philipp von Zabern, Mainz am Rhein.

Harry Falk. 2013. Remarks on the Minor Rock Edict of Aśoka at Ratanpurwa. *Jñāna-Pravāha Research Journal*, 16:29–48.

Hans Goebl. 2010. Dialectometry and quantitative mapping. In Alfred Lameli, Roland Kehrein, and Stefan Rabanus, editors, *Language and Space. An International Handbook of Linguistic Variation. Volume 2: Language Mapping*, number 30.2 in Handbücher der Sprach- und Kommunikationswissenschaft, pages 433–457, 2201–2212. De Gruyter Mouton, Berlin, New York.

Wilbert Jan Heeringa. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, University of Groningen.

Roland Kehrein. 2012. *Regionalsprachliche Spektren im Raum. Zur linguistischen Struktur der Vertikale*. Number 152 in Zeitschrift für Dialektologie und Linguistik. Beihefte. Franz Steiner Verlag, Stuttgart.

Joseph B. Kruskal. 1983. An overview of sequence comparison: time warps, string edits, and macromolecules. *SIAM Review*, 25(2):201–237.

Satya Swarup Misra and Haripriya Misra. 1982. *A historical grammar of Ardhamāgadhī*. Ashutosh Prakashan Sansthan, Varanasi.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

John Nerbonne. 2010. Mapping aggregate variation. In Alfred Lameli, Roland Kehrein, and Stefan Rabanus, editors, *Language and Space. An International Handbook of Linguistic Variation. Volume 2: Language Mapping*, number 30.2 in Handbücher der Sprach- und Kommunikationswissenschaft, pages 476–501. De Gruyter Mouton, Berlin, New York.

Thomas Oberlies. 2003. Aśokan Prakrit and Pāli. In Danesh Jain and George Cardona, editors, *The Indo-Aryan languages*, Routledge Language Family Series, pages 161–203. Routledge, London et al.

Simon Pickl and Simon Pröll. 2019. Ergebnisse geostatistischer Analysen arealsprachlicher Variation im Deutschen. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum. Ein internationales Handbuch der Sprachvariation. Volume 4: Deutsch*, number 30.4 in Handbücher der Sprach- und Kommunikationswissenschaft, pages 861–879. De Gruyter Mouton, Berlin, Boston.

Richard Salomon. 1998. *Indian epigraphy: a guide to the study of inscriptions in Sanskrit, Prakrit and the other Indo-Aryan languages*. South Asia Research. Oxford University Press, New York et al.

Yves Scherrer and Philipp Stoeckle. 2016. A quantitative approach to Swiss German. Dialectometric analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, 24(1):92–125.

Ulrich Schneider. 1978. *Die großen Felsen-Edikte Aśokas. Kritische Ausgabe, Übersetzung und Analyse der Texte*. Number 11 in Freiburger Beiträge zur Indologie. Otto Harassowitz, Wiesbaden.

Sukumar Sen. 1960. *A comparative grammar of Middle Indo-Aryan*. Number 1 in Special Publications of the Linguistic Society of India. Linguistic Society of India, Poona.

Ryan Soh-Eun Shim and John Nerbonne. 2022. dialectR: Doing dialectometry in R. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 20–27, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Oskar von Hinüber. 2001. *Das ältere Mittelindisch im Überblick. 2., erweiterte Auflage*. Number 20 in Veröffentlichungen der Kommission für Sprachen und Kulturen Südasiens. Verlag der Österreichischen Akademie der Wissenschaften, Wien.

Martijn Wieling and John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, 25:700–715.

A Abbreviations of Aśokan Sites

Ahr	Ahraurā	All	Allāhābād
Are	Arerāj	Bah	Bahāpur
Bai	Bairāṭ	Bar	Barābār
Bhā	Bhābhrū Stone Inscription	Bra	Brahmagiri
Dha	Dhaulī	E<u>rr</u>	E<u>rr</u>aguḍi
Gav	Gavīmaṭh	Gir	Girnār
Guj	Gujarrā	Jaṭ	Jaṭiṅga-Rāmeśvara
Jau	Jaugaḍa	Kāl	Kālsī
Kan	Kandahār	Laġ	Laġman
Lum	Lumbinī	Mān	Mānsehrā
Mas	Maski	Mer	Merāṭh
Nāg	Nāgārjuni	Nan	Nandangaṛh
Nig	Niglīvā	Niṭ	Niṭṭūr
Pāḷ	Pāḷkīguṇḍu	Pāṅ	Pāṅgurāriā
Rat	Ratanpurvā	Rāj	Rājula-Maṇḍagiri
Rām	Rāmpūrvā	Rūp	Rūpnāth
Sah	Sahasrām	Sāñ	Sāñcī
San	Sannati	Sār	Sārnāth
Śāh	Śāhbāzgaṛhī	Sid	Siddapur
Sop	Sopārā	Tax	Taxilā
Ṭop	Ṭoprā	Uḍe	Uḍegoḷam