Dialectometric Approaches to Korean

SIMON BARNES-SADLER
SOAS, University of London

1 Introduction

The languages spoken on the Korean peninsula have been known to vary over the area in which they are spoken since the earliest reports of Chinese ethnographers. This variation has been acknowledged throughout the history of the language, for example by the creation of specific, albeit unattested, letters for non-standard pronunciations in the *hwunmincengum haylyey* (Lee and Ramsey 2011: 159), but it is only in the twentieth century that this variation began to be recorded systematically, specifically in the framework of traditional dialectology.

Within this paradigm, studies of single or very small sets of features (often also restricted to a single traditional dialect) predominate; a systematic review of two databases of Korean research (the Korean studies Information Service System (KISS) and DBpia) revealed that of all of the papers featuring the word dialect (*pangen*) in their title published over the course of 2016, fifty percent (twelve out of twenty four) in the case of KISS and forty eight percent (fifteen out of thirty one) in the case of DBpia were single feature studies. This compares with only two papers approaching

Japanese/Korean Linguistics 24.
Edited by Kenshi Funakoshi, Shigeto Kawahara, and Christopher D. Tancredi

linguistic variation from broader perspectives on KISS and four on DBpia, that is, roughly four and thirteen percent of papers, respectively. It may be inferred from this that, while the fine-grained and in-depth understanding of the phonological, grammatical, and lexical characteristics of many Korean varieties continues to develop apace, the bigger picture of the totality of geographical variation of language over the Korean peninsula has been somewhat neglected. To address this issue and explore such variation from a more synoptic perspective, we suggest the application of the quantitative techniques of dialectometry to Korean data drawn from the Linguistic Atlas of Korea (LAK) (Lee et al. 2008).

Although the awareness of dialectometry is longstanding in Korean linguistics (see Lee 1997: 314; Choi 2001; Lee 2003: 205–7), the data available has been considered insufficient and the sub-discipline has received comparatively little attention. Therefore, the background and practice of the sub-discipline bear examining before the results of such an analysis are discussed. This paper, then, first introduces dialectometry in more depth before going on to present a dialectometric perspective on the description of linguistic variation in the Republic of Korea (ROK) and the classification of this variation into dialect areas before concluding with a brief review. [1]

2 Fundamentals of Dialectometry

Dialectometry has been broadly defined as 'the use of computational and quantitative techniques in dialectology' (Nerbonne and Kretzschmar 2013). It is perhaps more commonly understood in a more narrow sense which connotes the mass comparison of quantified aggregate linguistic variation over a particular geographical area. In this section, we provide a little more background to the field of dialectometry before going on to discuss how linguistic variation may be quantified with particular emphasis on the Levenshtein/String-edit distance, which was used in this study, and what precisely is meant by the term 'aggregation'.

2.1 Prerequisites and History

While it may be possible to trace the roots of dialectometry back to the late-nineteenth century (see Nerbonne and Kretzschmar 2013: 1), it is widely acknowledged that dialectometry proper began with Seguy's landmark paper on variation in the French of Gascony (1973). Dialectometric research was initially more widely practiced in Europe (for example, Goebl et al. 1982) before spreading to North America (for an early example, see

[1] This work was supported by Laboratory Program for Korean Studies through the Ministry of Education of the Republic of Korea and Korean Studies Promotion Service of the Academy of Korean Studies (AKS-2016-LAB-2250003)

Kretzschmar 1996), but remained a somehwat niche approach to language variation. In the early twenty-first century, the increasing availability of sufficiently powerful computers has led to a massive increase in the possibilities and accessibility of dialectometric research, especially with the ready availability of such specialist dialectometric software as Visual Dialectometry (Goebl 2004) and Gabmap (Nerbonne et al. 2011). This paper represents an attempt to introduce this field and its possibilities for research on Korean through its application to the data collected for the production of the LAK (Lee et al. 2008).

2.2 Linguistic Distance (Levenshtein/String-edit Distance)

The 'Levenshtein' or 'String-edit' distance (Levenshtein 1966) is a numerical representation of the dissimilarity between two strings. It is derived from the number of operations (insertions, deletions, and substitutions) which must be performed on an input string in order to transform or edit it into an output string. Once values are provided for each operation, the Levenshtein distance (hereafter 'linguistic distance') may be generated algorithmically (for full details, see Heeringa 2004). In the case of the current study, all operations are assigned the same value of 'one'. We may take the comparison of the forms of *mogi* 'mosquito' recorded in Ganghwa Island, Gyeonggi Province, Muju, North Cheolla Province and Mungyeong, North Gyeongsang Province as an example. These dialect forms are [mogi], [mogu] and [məgɛŋi], respectively. As a result of carrying out just one operation (substitution) to transform [mogi] into [mogu] the linguistic distance between these two forms is one, whereas the linguistic distance between [mogi] and [məgɛŋi] is three, since one substitution and two insertions are required. Finally, four operations (two substitutions and two insertions) must be carried out to transform [mogu] into [məgɛŋi]; therefore the linguistic distance between these two forms is four.

Incorporating the concept of linguistic distance into Korean dialectology represents a profound difference between the approach advocated here and the traditional approach to linguistic variation over the Korean peninsula. Most notably it provides us with a more nuanced picture of variation, in which the degree of difference between dialect forms is systematically measured and taken into account, rather than the earlier situation in which dialect forms were categorically classified on the basis of subjective judgement. For example, applying the concept of linguistic distance allows us to acknowledge the difference between two notional dialect forms which differ in the articulation of a specific consonant, but are otherwise identical, while also recognising that they are more similar than entirely lexically distinct dialect forms. Furthermore, this has implications for the production and interpretation of dialect maps. Whereas traditional

dialect maps are representations of the geographical distribution of the realisation of particular linguistic items over a given surveyed area, dialectometric maps tend to be somewhat more relational. They only very rarely give a clear indication as to the spatial distribution of particular linguistic items or features, but present a visual representation of the (dis)similarity of the realisations of a set of linguistic items.

Also contributing to this fundamental change in the information contained in dialectometric rather than traditional dialectological maps is the aggregation of data, which we examine in more detail below.

2.3 Aggregation

It may be argued that a degree of aggregation is implicit in much dialectological research. For example, transcription of a dialect survey using the International Phonetic Alphabet (IPA) necessarily aggregates ideolectal realisations of phonemes into the sound values which are conventionally assigned to each IPA symbol. A concrete example of this would be the way that precise formant values for vowels collected from different informants for this notional survey would not be ascertainable from the transcription; rather a reader would only be able to infer the broader articulatory characteristics of each transcribed vowel.

Here, though, we are explicit in identifying aggregation as a feature of the dialectometric approach. Whereas earlier approaches to variation focussed on single linguistic phenomena (for example the distribution of dialect forms of the word *kawi* 'scissors' over the Korean peninsula (see, for example, Kim 1974: 429) or the operation of vowel harmony in verb stems in the Central Dialect (see Kim 2001: 325)), every point of difference in transcription between every recorded item at each survey site is taken into account in calculating linguistic distance. There are arguments both for and against this approach.

The advantages of a dialectometric approach may be broadly summarised by saying that it allows us to identify areas in which language use may be characterised as 'generally' similar or dissimilar and to determine the degree of linguistic (dis)similarity observed over a continuum through the examination of potentially vast numbers of variants of linguistic items. This is an especially marked contrast with earlier single feature or isogloss-based studies which made categorical distinctions between survey sites on the basis of either single or very small numbers of features. Furthermore, it contributes to removing researcher biases in the identification of features which may be taken as representative of linguistic variation.

The sacrifice made when taking such an approach is a loss of granularity. While the strong signal provided by a large number of

comparators allows us to say with confidence which survey sites are most globally (dis)similar, precisely how the survey sites may be differentiated linguistically is not immediately apparent, as it is in more traditional dialectology.

In reality, whether aggregate, 'big picture' approaches or smaller-scale more detail-oriented approaches are to be preferred is a function of the question at hand. Thus, dialectometry may be considered suitable for examining questions of the general relationship between linguistic (dis)similarity and geographic space, such as dialect taxonomy, while traditional dialectological methodologies may be fruitfully applied in order to establish the precise nature and distribution of variation over a given space. The complementary nature of these sub-disciplines is revealed in the observation that, for a truly meaningful analysis of the spatial distribution of single dialect features, we require a knowledge of geographical linguistic variation which allows us to 'interpret individual features with respect to global patterns and… assess the importance of individual signals', rather than continuing to rely on the subjective judgement of researchers to determine the significance of (bundles of) isoglosses (Nerbonne 2009: 193).

3 Application of Dialectometry to Korean and the LAK

The foregoing, then, described the fundamental concepts of dialectometry and the pre-requisites for carrying it out. We now go on to examine how these ideas and methods may be applied in the Korean situation. The LAK data used in this analysis is comprised of data gathered from 138 survey sites spread over the territory of the contemporary ROK in the form of maps of the distribution of dialect variants of 153 linguistic items. These 21,114 items of data were then transcribed using the IPA for processing using Gabmap (Nerbonne et al. 2011). While a linguistic distance between each item of data collected for each surveyed item is calculated (1,446,309 individual linguistic distances), the most directly relevant product of this processing for our purposes is the aggregated survey site by survey site distance table (9,453 aggregated linguistic distances).

3.1 Breadth of Transcription and Hangul

In contrast with earlier dialect materials (e.g. Ogura 2009 [1940]) or those which record, for example, European languages, the materials in the LAK are transcribed exclusively using a Hangul rather than IPA based orthography. While Hangul has been rightly praised as an intellectual achievement (e.g. Sampson 1985), in its conventional application to Korean it does not record sound as systematically or with the same 'narrowness of transcription', as does the IPA. Breadth of transcription has been identified as a source of bias in assessing linguistis distance (Wieling and Nerbonne

2011). The relatively broad Hangul transcription used in the the LAK may lead to the data appearing more homogenous than it otherwise would.

A further issue occasioned by the Hangul transciption is the problem of text-processing. Most immediately relevant is the fact that Hangul is not currently supported by dialectometric software. Thus, for practical reasons it is necessary to transliterate the Hangul transcriptions of the LAK dialect forms into a broad IPA transcription. A particular issue attendant upon this transliteration is the question of whether the automatic phonological processes associated with standard Hangul orthography were to be represented or not. Ultimately, cases that could be regularly and unambiguously identified were transcribed as having undergone a select set of phonological processes in IPA. In the table below Hangul transcriptions of dialect forms are presented along with their Standard South Korean equivalents (in parentheses) and IPA transliterations. Examples include, but are not limited to, the following:

Phonological Process	**Hangul Transcription**	**IPA Transliteration**
Palatalisation of /s/	둥시기 (mengsek)	[tuŋʃi̞gi]
Fortification	못자리 (moscali)	[mo̞t͡ʃ͈aɾi]
Neutralisation	윳 (yuch)	[ju̚t̚]
Aspiration	놓다 (kilwuta)	[notʰa]
Nasal Assimilation	혹말 (mokmal)	[hoŋmal]

Thus, the data contained in the LAK were transcribed using just 33 unique characters, or 55 unique tokens once the combination of characters to represent such features as aspiration or greater articulatory tension is taken into account.

3.2 Consistency of Data

A regrettable limitation of this study is that it must be confined to the linguistic variation of the contemporary ROK. This is due to the coverage of the data contained in the LAK.

While there is no data included in the LAK for the entirety of the territory of the contemporary Democratic People's Republic of Korea, we may also note that there are several cases where data has not been collected from individual survey sites or larger areas for particular entries in the LAK. Elsewhere (Prokić et al. 2012) it has been suggested that those items which were collected from fewer than eighty percent of the total number of target survey sites should be excluded from dialectometric analysis. Since only one item to appear in the LAK has responses collected from fewer than eighty percent of the sites (*homissisi* 'agricultural festival generally held in

July celebrating the completion of weeding the fields' with dialect forms recorded for seventy percent of all survey sites), it was not considered necessary to exclude this data. It is a further advantage of aggregate approaches carried out on this scale that such a small amount of missing data as this will have only a negligible effect on the overall results of the analysis.

3.3 Beam Maps and Point Choropleth Maps

Below we present beam maps[2] which provide some basic insights into the linguistic (dis)similarity between the dialect forms used at each of the LAK survey sites. The map in Fig. 1 on the left connects each site to a small set of its geographical nearest neighbours, while the map on the right shows sites considerably more inter-connection between sites. On both maps, a darker connecting line indicates a stronger degree of linguistic affinity and a lighter colour indicates greater linguistic dissimilarity:

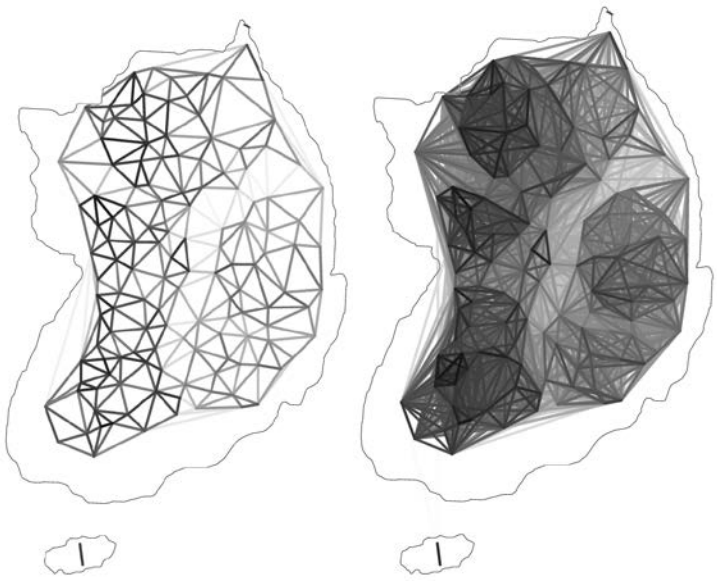

Figure 1. Beam Maps Summarising Linguistic Difference Across Survey Sites

These data may be manipulated in a wide variety of ways to produce reference point maps in the form of choropleth maps which visualise the linguistic (dis)similarity of a single survey site with all of the other survey

[2] For reasons of printing the visualisations presented here are restricted in terms of size, color, and number. A wider range of larger, full-color images are available on my personal website: https://thehackjar.com/category/publication-resources/

sites. Alternatively, these data may be represented using non-cartographic visualisations, for example plotting linguistic distance against geographic distance. When this is done for the entire data set, we note that a particular sub-linear relationship between linguistic (dis)similarity and geographical distance is revealed. This distribution of linguistic (dis)similarity over space is, in fact, cross-linguistically commonly observed and conforms to a pattern which has come to be known as 'Séguy's Law' (Nerbonne et al. 2010; Nerbonne 2010). This finding demonstrates the new insights into both specifically Korean and cross-linguistic variation made possible by the dialectometric approach.

The visualisations presented and described above summarise the aggregate linguistic variation recorded by the LAK. In the next section, we give an example of the application of quantitative, dialectometric analytic techniques to these data in order to provide a new perspective on an on-going problem in Korean dialectology.

4 Korean Dialect Taxonomy

It was noted earlier that there has been a strong tradition of recording dialect forms of Korean over the twentieth century. This data, however, has served more to document the varied linguistic forms present on the Korean peninsula rather than to inform other aspects of dialect research, for example the taxonomy of those dialects. It has even been asserted that the currently broadly accepted dialect divisions of Korean "have some basis in the characterising features of the language, but they are also to a certain extent constructed simply for convenience of description" (Lee and Ramsey 2000: 313). Taking a dialectometric approach to answering the question of how the surveyed area under examination in this paper should be classified into dialects thus represents a three-fold departure from earlier works in that:

- It has its basis entirely "in the characterising features of the language" i.e., it is empirical
- It incorporates a great deal more data than earlier taxonomies
- It uses linguistic distance rather than isoglosses to distinguish the dialects

We go into more detail about traditional approaches to dialect taxonomy and its contrasts with the dialectometric approach taken here in the following section.

4.1 Traditional Approaches

While Korean is a well-surveyed language and a vast amount of data recording dialect forms exist, somewhat fewer attempts at dialect taxonomy have been made. Following Yeon and Brown (2015: 461), it may be noted that the current broad consensus recognises six dialects, but taxonomies proposing as few as three major dialect areas (Kim 1977) and as many as nine (Coseneyenkwuhoy 1937) have also been advocated. What is more, even taxonomies which propose similar numbers of dialect areas do not necessarily agree on the precise geographical distribution of these areas.

It is worth noting that the conclusions about the number of dialect areas to be found on the Korean peninsula are as diverse as the criteria used to establish them. The number and type of linguistic features selected to be the basis of representative isoglosses varies widely between researchers and the criteria by which such linguistic features are chosen are quite unclear. For example, the well-known six-way division of the Korean peninsula into dialect areas which appears in Ogura Shinpei's 'Outline of Korean Dialects' is based on the distribution of thirteen linguistic features which take specific differences in vocabulary, phonology, and morphology into account (see Ogura 1940: 19–67; Choi 2001: 375). In contrast, other attempts have been made at establishing dialect areas over the Korean peninsula on the basis of just a single, linguistic feature (for fuller discussion see Lee 2003: 446–8).

4.2 Dialectometric Approaches: Clustering and Multi-Dimensional Scaling (MDS)

Clustering is very broadly defined as "a set of techniques for sorting variables, individuals and the like into groups on the basis of their similarity to each other" (Cramer and Howitt 2004: 43). Employing these techniques with regard to the dialectometric data described above is broadly analogous to establishing dialect areas in traditional dialectology. Once again, the main methodological differences lie in the incorporation of the concept of linguistic distance and aggregation of the data, while the concrete difference in terms of method is the employment of a (generally hierarchical and agglomerative) clustering algorithm to generate the clusters rather than the judgement of the researcher. The dialect areas established using this method may be projected onto maps for examination and represented as dendrograms. For reasons of space, here we restrict ourselves to presenting three cartographic representations of clustering carried out on the LAK data discussed in the foregoing section:

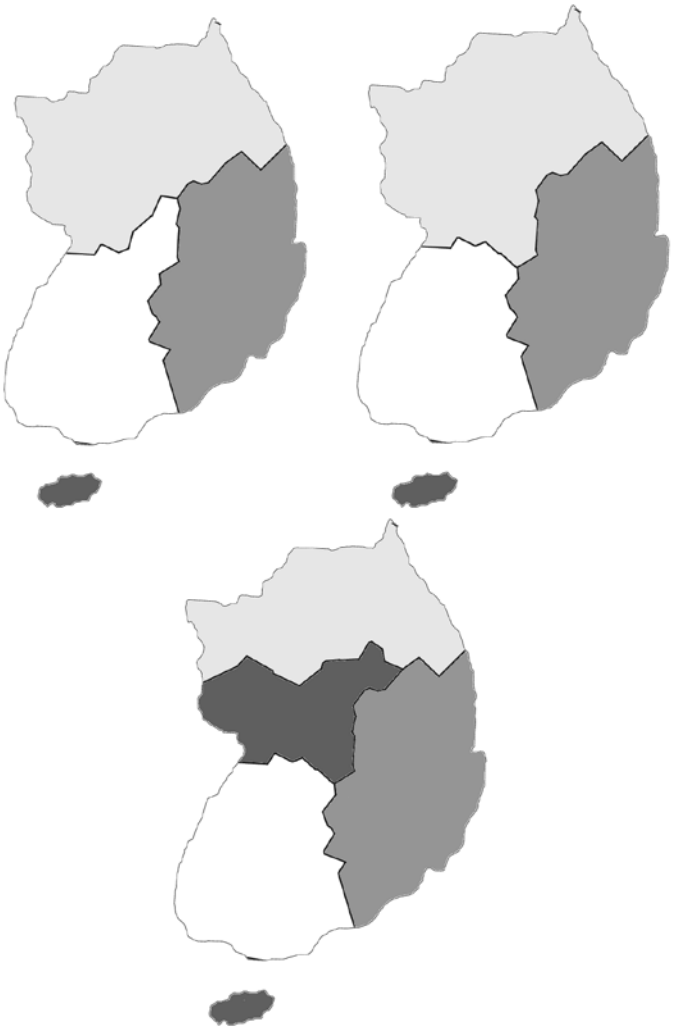

Figure 2. Maps showing clusters of LAK data using Complete Link, Group Average and Ward's Method Clustering Algorithms (top-right, top-left and bottom)

It may be easily observed that the clusters established using different clustering algorithms are not congruent with one another. Particularly striking is the fact that the dialect areas established using Ward's Method, generally considered to be the preferred clustering algorithm for quantitative dialectology (Heeringa et al. 2002: 451), groups together Jeju Island data with some areas of the traditional Central dialect area in a radical departure from all prior Korean dialect taxonomies.

Given such unexpected results, these clusters may be checked against Multi-Dimensional Scaling (MDS) plots of the data. In contrast to maps, MDS plots visualise the relationship between the data points in terms of their spatial configuration, albeit imperfectly due to the inherent characteristics of dimension reduction. In this case, points which are closer to one another are distinguished by a smaller linguistic distance, i.e., they may be considered more similar. Below we present a MDS plot of the data with points shaped to correspond with the map representing the Ward's Method clustering of the data (the darkest grey is represented by squares, white by circles, light grey is represented by crosses and the darker grey of the South east is represented by triangles):

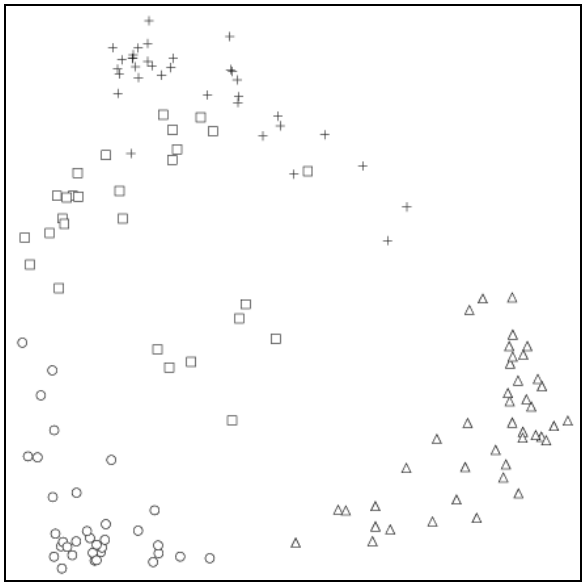

Figure 3. MDS Plot of LAK Data with Points Shaped to Represent Ward's Method Clustering

The impression that there are no wholly clear clusters in the data is inescapable from this MDS plot, which also serves to underline the degree to which linguistic variation in the ROK is continuous rather than abrupt. In order to reconcile this with the desire to establish dialect areas, in the following penultimate section we examine a more nuanced quantitative approach to the establishment of dialect areas.

4.3 Limitations and 'Fuzzy' Clustering

Simple hierarchical clustering has been subjected to a certain amount of criticism, generally due to the volatility of the results (e.g. Prokić and Nerbonne 2008). From the maps presented above it is clear that the selection of one clustering algorithm over another has serious repercussions for the grouping of data points, that is, for the establishment of dialect areas. Furthermore, these methods are also quite susceptible to statistical outliers having a disproportionate influence on the ultimate composition of the clusters. A range of techniques may be employed in order to avoid these issues. Using Gabmap it is possible to undertake so-called 'fuzzy' clustering. This method assigns particular values which represent the frequency of appearance of particular clusters in repeated clustering runs when a small, varying amount of random noise is added to the distance matrix upon which the clustering is based (see Kleiweg et al. 2004 for more detail). Larger values indicate more stable clusters which appear in a greater number of clustering runs with the largest value, one hundred, signifying clusters which invariably appear. We describe the clusters of survey sites identified by this method and provide the scores which characterise their stability below.

Especially robust clusters identified using this method include both survey sites on Jeju Island (100), the forty three sites of Gyeongsang Province (98), the nineteen sites in Gyeonggi Province (95), five sites in the South East of Gangwon Province (Pyeongchang, Myeongju, Yeongwon, Jeongseon and Samcheok) (100), and finally Jeolla Province's Muju and a group of sites in Chungcheong Province (Boeun, Okcheon, Yeongdong and Geumsan) (100). Somewhat less robust, but still appearing with great frequency was a cluster of the thirty four survey sites in Jeolla Province (excluding Muju) (83). The remaining thirty survey sites may not be grouped into clusters larger than three sites with greater than seventy per cent reliability. A further finding of note from this fuzzy clustering is that the long-standing Central dialect group emerges from this data in fewer than two thirds of clustering runs, which suggests that it lacks robustness as a grouping of survey sites, that is, as a dialect area.

In contrast to the hard clustering presented in section 4.2 and the traditional dialect taxonomy of the ROK, fuzzy clustering suggests that, rather than four dialect areas, linguistic variation over that territory would be more accurately represented by a minimum of six dialect areas. Conversely, the fact that such a high proportion of survey sites do not regularly appear in a particular cluster combined with the observation that no obvious clusters emerge in the MDS plot presented above, makes it seem reasonable to question whether a more nuanced approach to the description of dialectological variation in the the ROK is to be preferred. Such approaches are available within the frameworks of both traditional

dialectology and dialectometry, but their examination falls outside the scope of this paper.

5 Conclusion

In the preceding sections we have briefly reviewed the field of dialectometry, discussed its application to Korean and presented both a dialectometric overview of the distribution of linguistic (dis)similarity in the ROK as well as an example of how dialectometry may be applied to a specific problem in Korean dialectology - dialect taxonomy. It must be emphasised that this paper represents only a first and introductory attempt to incorporate dialectometric methodology into the study of the geographical linguistic variation of Korean. The possibilities of dialectometric approaches to contribute to the understanding of synchronic variation in Korean as new data sources (e.g. dialect corpora) become available and both dialectometric and dialectological techniques are developed and refined is not to be underestimated.

References

Choi, M-O. 2001. Hankwuk pangen kwuhoyk [The Classification of Korean Dialects]. *Pangenhak sacen* [The Dictionary of Dialectology], ed. Pangenyenkwuhoy [Dialect Research Society], 373–84. Seoul: Thayhaksa.

Coseneyenkwuhoy. 1937. *Pangencip* [Dialect Collection]. Place unknown: Sunhwacoseneyenkwupu.

Cramer, D. and D. Howitt. 2004. *The SAGE Dictionary of Statistics*. London; Thousand Oaks, CA; New Delhi: SAGE Publications.

Goebl, H., S. Selberherr, W.D. Rase and H. Pudlatz. 1982. Atlas, matrices et similarités: Petit aperçu dialetométrique [Atlases, matrices and similarities: a small diaectometric survey]. *Computers and the Humanities* 16(2): 69–84.

Goebl, H. 2004. VDM-Visual Dialectometry. Vorstellung eines dialektometrischen Software-Pakets auf CD-ROM (mit Beispielen zu ALF und Dees 1980) [VDM – Visual Dialectometry. Intorduction to a Dialectometric Software Package on CD-ROM (with examples from the ALF and Dees 1980)]. *Romanistik und neue Medien* [Romance Studies and New Media], eds. Dahmen, Wolfgang et al., 209–41. Tübingen:Gunter Narr Verlag.

Heeringa, W. 2004. Measuring Dialect Pronunciation Differences Using Levenshtein Distance. Doctoral Dissertation, University of Groningen.

Heeringa, W., J. Nerbonne and P. Kleiweg. 2002. Validating Dialect Comparison Methods. *Classification, Automation and New Media*, eds. W. Gaul and G. Ritter, 445–52. Berlin; Heidelberg: Springer-Verlag.

Kim, B-G. 2001. Cwungpwu pangen [Central Dialect]. *Pangenhaksacen* [Dictionary of Dialectology], ed. Pangenyenkwuhoy [Dialect Research Society], 322–8. Seoul: Thayhaksa.

Kim, H-G., 1974. *Hankwukpangenyenkwu* [Korean Dialect Reserach]. Seoul: Seoul tayhakkyo chwulphanpwu.

Kim, K-C. 1977. *Pangenhak* [Dialectology]. Place unknown: Cenghyangchwulphansa.

Kleiweg, P., J. Nerbonne and L. Bosveld. 2004. Geographic Projection of Cluster Composites. *International Conference on Theory and Application of Diagrams*. Berlin; Heidelberg: Springer Verlag.

Kretzschmar, W.A. 1996. *Introduction to Quantitative Analysis of Linguistic Survey Data: An Atlas by the Numbers*. Thousand Oaks, CA; London: Sage Publications.

Lee, I. 1997. *Hankwukuy ene* [The Language of Korea]. Seoul: Sinkwumunhwasa.

Lee, I, G. Jeon, G. Lee, B. Lee and M. Choi. 2008. *Hankwukenecito* [Linguistic Atlas of Korean]. Seoul: Thayhaksa.

Lee, I and R.S. Ramsey. 2000. *The Korean Language*. Albany, NY: State University of New York Press.

Lee, K-M and R.S. Ramsey. 2011. *A History of the Korean Language*. Cambridge: Cambridge University Press.

Lee, S-K. 2003. *Kwukepangenhak* [Korean Dialectology]. Seoul: Hwayensa.

Levenshtein, V.I. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10(8): 707–10.

Nerbonne, J. 2009. Data Driven Dialectology. *Language and Linguistics Compass* 3(1): 175–98.

Nerbonne, J. 2010. Measuring the Diffusion of Linguistic Change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1559): 3821–8.

Nerbonne, J., C. Rinke, C. Gooskens, P. Kleiweg and T. Leinonen. 2011. Gabmap – A Web Application for Dialectometry. *Dialectologia* Special Issue 2: 1–23.

Nerbonne, J., J. Prokić, M. Wieling and C. Gooskens. 2010. Some further dialectometrical steps. *Tools for Linguistic Variation, Supplements of the Anuario de Filologia Vasca "Julio Urquijo", XIII* eds. G. Aurrekoetxea and J.L. Ormaetxea, 41-56. Bilbao: University of the Basque Country.

Nerbonne, J. and W.A. Kretzshmar Jr. 2013. Dialectometry ++. *Literary and Linguistic Computing* 28(1): 2–12.

Ogura, S. 1940. *The Outline of the Korean Dialects*. Tokyo: Tokyo Bunko. Reprinted in 2009. *Cosenepangensacen* [Korean Dialect Dictionary]. Paju: Thayhaksa.

Prokić, J., Ç. Çöltekin and J. Nerbonne. 2012. Detecting Shibboleths. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*: 72–82.

Sampson, G. 1985. *Writing Systems: A Linguistic Introduction*. London: Hutchinson.

Séguy, J. 1973. La Dialectométrie dans l'Atlas Linguistique de la Gascogne [Dialectometry in the Linguistic Atlas of Gascogne]. *Revue de Linguistique Romane* 37:1–24.

Prokić, J. and J. Nerbonne. 2008. Recognising groups among dialects. *International journal of humanities and arts computing* 2(1-2): 153–72.

Wieling, M. and J. Nerbonne. 2011. Measuring Linguistic Variation Commensurably. *Dialectologia* Special Issue 2: 141–62.

Yeon, J. and L. Brown. 2015. Varieties of Contemporary Korean. *The Handbook of Korean Linguistics*, eds. Yeon Jaehoon and L. Brown, 459–76. Hoboken, NJ: Wiley Blackwell.