Full title: A Digital Humanities Approach to Inter-Korean Linguistic Divergence: Stylometric Analysis of ROK and DPRK Journalistic Texts

Abstract: Linguistic divergence between standard varieties of Korean has been much studied, however, it has largely concerned itself with fine-grained analyses of single points of divergence, for example vocabulary, and the language policy behind such divergence. In contrast, this paper examines general trends of language in use in the ROK and DPRK in a specific genre of writing.

We first briefly review prior research on the linguistic divergence which the standard varieties of these countries have undergone to contextualize our argument that a digital humanities approach could provide new insights for the field. This includes taking advantage of internet mediated data collection and quantitative analyses applied to relatively large amounts of data.

In order to demonstrate the potential of this approach more fully, we present a small-scale stylometric analysis of ROK and DPRK journalistic texts. This pilot study suggests that national origin determines the stylistic characteristics of these texts to a greater extent than the topic and allows us to tentatively propose general characterizing features of ROK and DPRK journalistic style. We conclude with a prospectus for the incorporation of such methods into the study of ROK/DPRK linguistic divergence

Keywords: Linguistic divergence, *munhwaŏ, pyojunŏ,* Stylometry, Digital Humanities

A Digital Humanities Approach to Inter-Korean Linguistic Divergence: Stylometric Analysis of ROK and DPRK Journalistic Texts

1. Introduction[1].

Diversity has been present in the languages of the Korean Peninsula from their very earliest attestations (Lee and Ramsey 2011). In contemporary scholarship a great deal of linguistic diversity is still found conditioned by geographical (e.g. Jeong 2013) or social factors (e.g. Pak 2001). While variation is found and studied universally in natural, human languages, one area of research unique to the Korean language is the examination of how the division of the Korean Peninsula into two states has affected the language used in both countries. Not only is mobility and contact between people residing in both Koreas restricted, each country pursues different language policies from one another and subscribes to different definitions for the standard Korean language.

This paper provides a brief overview of research carried out to date on linguistic divergence which focuses on the context of the divided Korean Peninsula and identifies methods and an area of research which would supplement our existing understanding of this phenomenon, namely, the computer mediated methods of digital humanities and the much-overlooked field of linguistic style. We go on present arguments for the implementation of these approaches and present a pilot study on the stylistic divergence of linguistic style in journalistic texts in the Republic of Korea (ROK) and Democratic People's Republic of Korea (DPRK) in the digital humanities framework. We conclude by putting forward, along with the tentative results of this study, suggestions for taking this research forward to further investigate, and perhaps even mitigate the effects of, the on-going divergence of the Korean language in South and North Korea.

[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)
I would also like to extend my gratitude to the two anonymous reviewers whose thoughtful comments and corrections have improved the final version of this article in addition to being immensely encouraging.

2. **Prior Research on S/N Korean Linguistic Divergence**

Here we provide a brief overview of the research which has been carried out with a focus on the divergence between the standard varieties of Korean used in the ROK and DPRK. We do not address the long-studied geographical variation of non-standard varieties of Korean over the peninsula (e.g. Lee 1932) except where it is relevant to the development of the standard varieties.

A unified standard for the contemporary Korean language was first proposed at the beginning of the 20th century, but following the division of the Korean Peninsula, the standard languages of the ROK and the DPRK came to be defined differently. Separate standards and contradictory language policy are key factors underlying the linguistic divergence between the ROK and DPRK. The current definition of the standard languages of the ROK (*pyojunŏ*) is presented below. In slight contrast to Lee and Ramsey's assertion (2000: 309) that Pyeongyang speech was adopted in 1966 as the standard language of the DPRK (*munhwaŏ*), we contest that a single, concise, official definition of *munhwaŏ* which identifies Pyeongyang speech as its basis does not appear explicitly in *Chosŏnmalgyubŏmchip* or even in more recent works and guidelines on language standardisation produced in the DPRK. Rather, in terms of the theory of language standardisation, precisely what constitutes *munhwaŏ* must be inferred from the somewhat more general directions proposed by Kim Il-sung and principles underlying the standardization of individual aspects of the language, such as spelling and pronunciation (see National Language Committee 1988; Kim 2005; Choi and Kim 2005). Nevertheless, in practice it must be accepted that the language of Pyeongyang is influential at least in the conception of the DPRK standard. Consequently, we take the dictionary definition of *munhwaŏ* provided in translation by Yeon (2000: 148) and present it alongside the official definition of *pyojunŏ*:

> "*Pyojunŏ* is in principle the language used by refined people in contemporary Seoul."
>
> (NLA 2017)

> "*Munhwaŏ* is the language that is cultivated and refined to fit the feelings of the working class centring around the revolutionary capital city [of Pyongyang] under the great leadership of the labour [sic] Party that assumed sovereignty during the reconstruction period of socialism; all [North] Korean people regard it as their standard speech; our *munhwaŏ* was developed by our Party and by the autonomous view of linguistic ideas of our people's revered leader Kim Il Sung and was furthered by our Party's proper language policy after Korean liberation [from Japanese rule in 1945]; it is based on Pyongyang speech, which is an independently promoted beautiful Korean, cast in a nationalistic spirit."
>
> (Sahoe kwahagwŏn ŏnŏhak yŏn'guso 1981: 1007)

The extent of the influence of these different definitions on the actual form of the language is debatable, however, it is clear that the linguistic divergence between the ROK and DPRK, with a special focus on divergent vocabulary, has long been a research pre-occupation (representative examples include Choi and Jeon 1994; Kim 2002, 144-163; Yeonhap 2002; Yeon 2006, 33-36; Cho 2007; Kim 2012a; Ministry of Unification 2016 etc.). A common theme of prior research is a focus on particular semantic fields where it is thought that divergence is greatest, for example technology or plant nomenclature. Thus, the global extent of lexical (dis)similarity is challenging to assess and the degree to which the vocabulary other semantic fields have diverged in the ROK and DPRK has not been specifically investigated. Nevertheless, Pyeongan dialect words that have been incorporated into *munhwaŏ* as the standard form in contrast to the Central dialect words of *pyojunŏ* may be identified. For example, *buru* and *sangch'u* are such words, and denote 'lettuce' in *munhwaŏ* and *pyojunŏ*, respectively.

The underlying ideologies of the two countries also condition the choice between words which may be available in both standard languages. The influence of ideology on language policy has received a comparatively great deal of attention in (see Kumatani 1990 and Song 2012 for summaries). Special emphasis has been placed upon Kim Il-sung's writings on the subject (Kim

1964; Kim 1966), which are thought to have set the course of DPRK language policy, particularly with regard to so-called *maltatumgi* (language purification/refinement) over the latter half of the twentieth century. A specific outcome of this purification is the replacement of a large number of Sino-Korean words which are common in *pyojunŏ* with Native Korean equivalents in *munhwaŏ* exemplified here by the words *gwanjŏl* (關節) and *ppyŏmadi*, both meaning 'joint', but in *pyojunŏ* and *munhwaŏ*, respectively.

There are also less consciously motivated linguistic effects of the different social realities of the ROK and DPRK. This is particularly evident in terms referring to either the political or modern historical sphere, but its influence is felt in such phenomena as the different sources from which loanwords are borrowed (for example *pyojunŏ t'ŭraekt'ŏ* borrowed from English and *munhwaŏ ttŭrakttorŭ* borrowed from Russian, both meaning 'tractor') and divergent specialist terminology, for example that of the field of linguistics itself (Kwon 2006). Specific examples of this include the titles of government officials and heads of state (for example *munhwaŏ chusŏk* – 'premier/president' and *pyojunŏ taet'ongryŏng*) and it even problematizes such fundamental concepts as the name of the language (*chosŏnmal/han'gugŏ*). As a result of the divergence revealed in the research discussed above, a considerable body of work devoted to linguistic re-unification also exists, largely carried out in the more prescriptive fields of lexicography (e.g. Hong 2007), and language education (e.g. Kang et al. 2016).

The research briefly summarized in the foregoing section has granted us many insights into the on-going development of the standard languages of the ROK and DPRK, however, it may be considered to fall short in its examination of language in use. Perhaps in connection with this, it is also somewhat restricted in terms of its levels of linguistic analysis, i.e. it is mostly restricted to the lexical and phonological (although see Kim 2012b for a rare example of a research focused on grammatical divergence). Of course, given the difficulties attendant upon carrying out research within the borders of the DPRK for international researchers, it should come as no surprise that

only few studies concerning spoken language are available, while no shortage of attempts to compensate for the lack access to consultants have been made in other ways. For example, Park (2003; 2004) relies on standardized guidelines rather than primary data for a contrastive analysis of pronunciation and narrative speech. Written language originating in the DPRK, however, has become far less difficult to come by, especially media texts disseminated on-line. This allows us to take ask broader, novel questions about possible linguistic divergence between these two standardized varieties of Korean with the application of novel methods. In the next section, we provide some background to the digital humanities paradigm which enables the analysis of these texts, before moving on to outline the specific methods used for the data gathering and stylometric analysis used in the pilot study put forward in this paper as an example of the potential of these methods.

3. A Role for Digital Humanities and Korean Linguistic Divergence

Broader access to increasingly powerful computers has radically changed the scope and potential of humanities research over the late twentieth and early twenty first centuries. Not only have technological developments enabled the handling of larger quantities data, leading to the emergence of so-called 'big data', they have also had direct and indirect consequences on the whole process of research, from data gathering and analysis to the dissemination of that research. While the problem of finding a precise definition beyond the somewhat circular "intersection of the humanities and the digital" is well known in the discipline (Gardiner and Musto 2015: 1-13), the changes described above when taken together may in a very general an impressionistic way be identified as the major contributing factors to the creation of Digital Humanities.

While the precise scope of this emerging discipline or approach is still the subject of an on-going negotiation, linguistics has unquestionably been part of the digital humanities movement since its very inception, as demonstrated by the following excerpt taken from the first issue of one of the field's first journals:

> "We define humanities as broadly as possible. Our interests include literature of all times and countries, music, the visual arts, folklore, **the non-mathematical aspects of linguistics**, and all the phases of the social sciences that stress the humane"
>
> (Prospect 1966, 1 emphasis added)

This sub-field of digital humanities, often referred to as 'computational linguistics' may be broadly defined as the "field of science that deals with computational processing of a natural language" (Hajič 2004). In the same overview of the field, Hajič goes on to make it manifestly clear that the role of computers in diverse fields of linguistic research has only grown over the latter half of the twentieth century. He further contends that computer-aided research is of relevance for long-standing questions in theoretical as well as applied linguistics. These developments have not been passed over by Korean linguistics; localized versions of tools for natural language processing such KoNLPy (Park and Cho 2014) or pieces of software dedicated to the linguistic analysis of Korean such as *kkokkoma* (Lee et al. 2010) have been developed. It is not only the production of individual works of research or research tools which is enabled by the incorporation of digital humanities into linguistics, but whole fields such as corpus linguistics, contemporary dialectometry, and machine translation.

Turning to the matter at hand, we note that the question of linguistic (dis)similarity, especially as it is observed over time and space is a pre-occupation of multiple linguistic sub-disciplines. Computational techniques which allow its examination may also be considered an active field of research (see Lebart and Rajman 2000 for an overview). To date, the application of such methods to variation in Korean, be it national, geographical, or social, has been somewhat limited. We provide a summary of this research below, before concluding this section with a summary of arguments for allowing digital humanities thinking to play a greater role in this field. We then move on to the pilot study mentioned in the introduction, presenting techniques and a method drawn from a specific sub-field of computational linguistics along with our findings.

As noted above, Korean linguistics has not been immune to digital humanities approaches, although they are rarely explicitly identified as such. With regard to our narrow focus on the examination of linguistic variation in Korean, it seems that only a relatively small amount of research that uses methods of analysis which fall within the scope of digital humanities, even broadly defined, has been carried out. Dialectology has largely continued to pursue traditional methods although more data-driven work or work advocating the incorporation of more computational techniques may exceptionally be found, such as Kang's quantitative dialect taxonomy (2014). Similarly, exceptions may be found in historical linguistics, for example Kim et al.'s examination of the replacement of Native Korean and Sino-Korean vocabulary by loanwords (2017), although that specific field as a whole tends towards the philological. Given the quantitative descriptive and analytic tendencies of sociolinguistics since its earliest days (e.g. Labov 1966; Cedergren and Sankoff 1974) it is not surprising that we find studies into socially conditioned linguistic variation in the ROK which fall within this paradigm (for a recent example see Jang 2015). It must be noted, though, that such studies make up only a very small proportion of the research carried out on the Korean language.

The fields mentioned above are all of relevance to carrying out research on linguistic divergence between the ROK and the DPRK and studies which draw upon the methodologies and techniques of each of them may be identified. However, in contrast with the above fields, the influence of the digital humanities on national linguistic divergence on the Korean peninsula appears to extend only into research dissemination and, arguably, lexicography (e.g. Ministry of Unification 2017).

There are a number of significant advantages to adopting digital humanities methodologies for the purposes of examining this question. Chief among them is the fact that internet mediated data gathering grants us access to a wealth of data and aggregate quantitative analytical tools enable us to gain somewhat objective insights into the general picture of the on-going divergence, at least in comparison to the impressionistic, fine-grained analyses of the earlier work reviewed above.

It must be borne in mind, however, that simply because these measures are quantitative in nature, they do not provide an entirely objective perspective on linguistic divergence. Rather, they simply move the locus of researcher subjectivity. Rather than enumerate the various quantitative approaches to linguistic variation which have been developed to date, we instead narrow our focus to the specific sub-discipline upon which we draw in the small-scale study presented in this paper.

3.1 Stylometry

Linguistic style has been broadly defined as "situationally distinctive uses of language" (Crystal 2009: 460). It has been an object of interest to linguists since at least the mid-nineteenth century and has been pursued within an explicitly quantitative research paradigm since the early 1960s (McEnery and Oakes 2000). Similar to the digital humanities, it is a young field of uncertain scope and with no consensus as yet as to what constitutes canonical stylometry. For the purposes of this paper we take our definition of the discipline as follows:

> "Stylistic analysis is open-ended and exploratory. It aims to bring to light patterns in style which influence readers' perceptions and relate to the disciplinary concerns of literary and linguistic interpretation. Authorship studies aim at "yes or no" resolutions to existing problems, and avoid perceptible features if possible, working at the base strata of language where imitation or deliberate variation can be ruled out."
>
> Craig (2004)

The distinction drawn above between "stylistic analysis" and "authorship studies" is highly relevant to this paper. The discipline of stylometry encompasses a wide range of quantitative techniques which are most commonly used for the purposes of authorship attribution, especially in cases of disputed authorship. Despite its traditional focus on answering very specific questions about the language of individuals, the methods and tools of this linguistic sub-discipline have also been recognized as a possible inductive approach to linguistic analysis which may be applied to more sociolinguistic questions, such as examining the stylistic features of language produced by

people of different genders (Rayson et al. 1997). For the study presented in this paper, we extend this logic using stylometric tools and techniques to determine whether significant stylistic' differences may be found between texts produced in North and South Korea.

3.2 Method

In line with the digital humanities paradigm outlined above, data was gathered from internet mediated sources of texts. This involved the compilation of relatively small textual databases of South and North Korean media language. The collection of texts was composed of roughly fifty thousand words of contemporary journalistic language, drawn in equal parts from the ROK and DPRK and from five genres of journalistic writing (politics, economics, international affairs, culture, and sports[2]) published over the same timeframe, that is, June-July 2016. These texts were furthermore drawn from two North Korean and four South Korean on-line journalistic publications. The restriction to journalistic texts and to particular topics, as detailed above, serves to minimize the influence of a particularly strong extra-linguistic factor: genre difference. To a lesser extent, by limiting the time period within which the data was gathered and by taking data from multiple publications, the influence of topic and publication-specific editorial policy on the language used are also reduced. The primary factor determining the choice of June-July 2016 as the data collection period was that texts were drawn from the same, contemporary period, but it must be acknowledged that this timeframe was otherwise arbitrarily chosen. In terms of its duration, only the requirement that the period of data collection was long enough to provide sufficient material was influential.

The texts gathered underwent minimal pre-processing. They were converted into a machine readable format (UTF-8 Unicode encoding) without any mark-up. Due to issues with the Korean language text processing capabilities of the analytic software to be used, most especially with regard

[2] DPRK news was gathered from *chosŏnjunangt'ongsin* and *rodong sinmun* and ROK news was gathered from KBS *nyŭsŭ, yŏnhap nyŭsŭ, chosŏn ilbo* and *tonga ilbo.* Each source was represented by a roughly equal amount of text within its sub-corpus. The categorisation of texts largely followed the categories set up by the content producers themselves except in the case of *rodong sinmun*, the texts of which were sorted manually to fit into the broad themes of the corpora.

to the mixed script ROK texts in which Hanja also appeared, the texts were automatically transliterated into the Roman alphabet using the Yale Romanisation. This Romanisation was chosen since it is nearer to a true transliteration of written Korean than alternative Romanisation systems, such as the McCune-Reischauer system or the Revised Romanisation of 2000, but since all analyses are to be carried out on the lexical rather than graphemic level the choice of Romanisation system should not influence our findings and may be regarded as arbitrary in this case.

Texts which shared both topic and national origin were stored together in single files of each of which had total sizes of roughly 5,000 words. In this case, "words" is used to denote orthographic words, that is, they are defined by being surrounded by white space. While it would be possible to, for example, extract the nouns from the texts or divide the texts into their constituent morphs using one of the available natural language processing tools, the decision was taken not to do so for the purposes of this small-scale, exploratory pilot study. Since such pre-processing is not regularly carried out on languages with more extensive nominal and verbal morphology than English (Rybicki and Eder 2011) it does not appear to be an essential requirement for analysis. Furthermore, its effects would be challenging to predict or explain, and conclusively determining the correct procedure for pre-processing raw language data before analysis falls outside the scope of this paper. While the decision to examine only orthographic words was taken in this case, we do not advocate it exclusively over and above investigating parsed corpora. We acknowledge, however, that comparing the stylometric analysis of more highly processed Korean texts could potentially provide a fruitful direction for future research.

In the first instance, the choice was made not to restrict the list of frequently occurring words (conventionally abbreviated to MFWs) upon which the stylometric analyses were to be carried out, since we follow the norms of the discipline in which the most commonly used markers of textual

identity are the very frequently occurring function words, since these are thought to be less sensitive to the subject matter of the texts (Craig 2004).

While there is some discussion within the discipline of stylometry as to which programs and mathematical methods have the highest discriminatory power, it falls outside the scope of this paper to act as arbiter on such matters and we use tests and software which are widely accepted. Thus, using the stylo package (Eder et al. 2016) for the R statistical computing environment (R Core Team 2017) the processed collections of texts were analyzed inductively using the following statistical tests: agglomerative clustering, bootstrap clustering, and a quantitative contrastive analysis. It must be borne in mind that these are all 'unsupervised' statistical techniques, consequently the interpretation of their results is somewhat subjective. It is the responsibility of the researcher to avoid not only overlooking patterns which may not be manifestly obvious in the results, but also to refrain from drawing spurious conclusions which support a pre-determined position.

Carrying out the tests outlined above on even the relatively small amount of data collected for the purposes of this pilot study should allow us to establish whether there is a quantifiable difference in style between texts of South and North Korean origin, and if a difference is found what patterns of language use contribute to it. In the next section we go on to detail the results of these analyses.

3.2 Findings

The clustering analyses listed above, which categorize and then and visualize texts on the basis of stylistic similarity in terms of frequently used words or n-grams, that is, sequences of words or individual characters, (see Burrows 2002; Burrows 2007; Eder et al. 2015) were carried out. These were followed by tests which interrogated their reliability (Craig and McKinney 2009). All tests were performed using the stylo package (Eder et al. 2013) for R (R Core Team 2016). We present the results of these tests in graphical form along with some further analysis below.

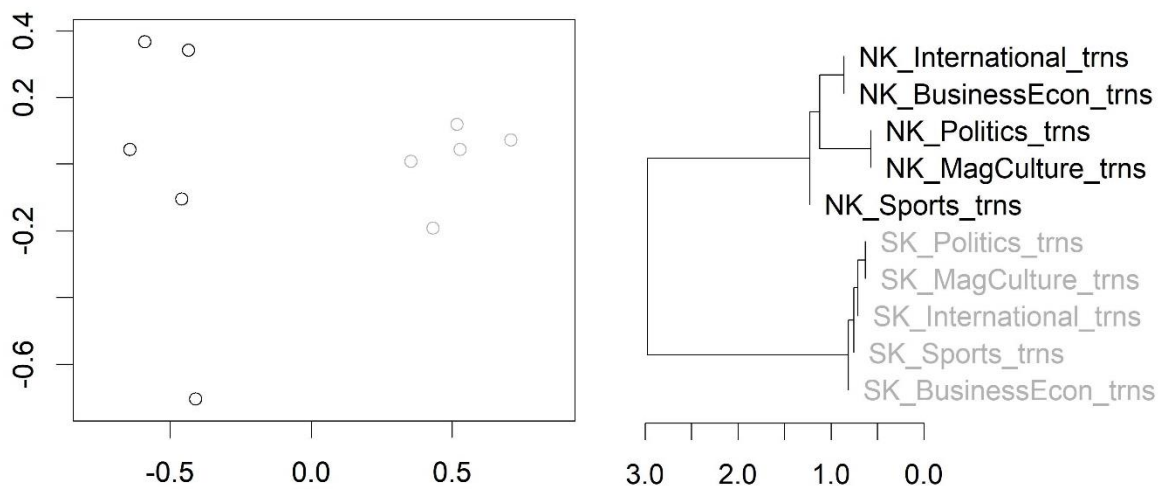

Figure 1:MDS Plot and Clustering Dendrogram of ROK and DPRK Journalistic Texts. Burrow's Classic Delta Distance, 100 MFW Culled at 0%

These two visualizations show the categorization of the thematically grouped ROK and DPRK texts in a multi-dimensional scaling (MDS) plot and a cluster analysis dendrogram, both based on the distribution of the one hundred most common words in the corpus across the thematically grouped sub-corpora of texts. These stylistic distinctions between journalistic texts of the DPRK (darker) and ROK (lighter) and the relative homogeneity between the style of the ROK journalistic texts across the thematic sub-groups in comparison to the DPRK texts are clearly noticeable (i.e. the lighter points are tightly grouped in the MDS plot and the ROK nodes marked with 'SK' show quite compressed branching in the dendrogram). A specific point worth emphasizing is the fact that the national origin of the thematic groupings appears to be more influential in determining textual similarity than topic. This is demonstrated most clearly in the dendrogram, in which we see the ROK and DPRK texts separated by two long branches, which represents a high degree of dissimilarity. The grouping of the thematic sub-groups at the terminal nodes of the dendrogram seems to imply that the relationship between the writing styles of different topics also varies between the two sub-corpora. Given the relatively small size of this sample, though, we would advocate caution in drawing this conclusion.

Turning to the MDS plot, it may be observed that the points representing collections of ROK texts are clustered a good deal more tightly than those representing the DPRK texts. Again, it is advisable to be wary of drawing conclusions which are too far-reaching, but these results do provide some indication that there is a more uniform or homogenous journalistic style of writing in the ROK to which writers adhere regardless of the topic, whereas the topic of a piece of journalistic writing will have a stronger influence on the choice of words used and their relative frequency for the production of a specific text in the DPRK.

Additional tests were then run to ascertain the reliability of the results found above. As mentioned above, there is not a complete consensus over the most authoritative methods in stylistic analysis. This is of particular relevance to a language as agglutinative as Korean. Rybicki and Eder (2011: 319-320) found that while the so-called "classic" Delta distance measure was "the most successful method of authorship attribution based on word frequencies, its success is not independent of the language of the texts studied". They further found that this measure was less reliable for Indo-European languages more highly inflected than English, such as Latin and Polish, but, contrary to this pattern, highly successful when applied to Hungarian data, i.e. a very highly inflected language. We do no propose to determine the applicability or not of the "classic" Delta distance measure to Korean data in this paper, but we do seek to verify our findings by making comparisons with clustering runs carried out using alternative distance measures. Specifically, these were Eder's Delta, Argamon's Delta, and the Canberra distance (which is recommended for more highly inflected languages). While we do not reproduce the visualizations here, the clustering runs using these diverse distance measures all re-iterated the underlying finding that texts originating in the DPRK clearly and unambiguously clustered together as did those originating in the ROK[3]. We note that the precise relationships between sub-corpora, that is the order of the agglomerative

[3] A large number of visualisations mentioned in the text are not reproduced here for reasons of space and readability. A larger selection of full colour visualisations will be made available through my personal website at the following address: https://thehackjar.com/category/publication-resources/

clustering schedule, varied between distance measures. Therefore, we cannot comment on degree of stylistic similarity between particular sub-corpora and, by extension, journalistic writing on particular topics.

A further point of contention is the length of the strings which may be taken as strongest proof of stylistic (dis)similarity. The above analyses were carried out on the basis of the most frequently occurring single words in the entire corpus, however, it has been convincingly demonstrated by Hoover (2002: 158) "cluster analysis of very frequent words often fails to produce completely accurate authorship attribution when it is performed on groups of texts by known authors", albeit with the caveat that frequent words may be preferred for particularly large corpora drawn from particularly diverse sources, which our collection of journalistic texts is not. To resolve this issue, Hoover suggests examining frequently occurring word sequences rather than frequently occurring single words. Thus, clustering runs using frequently occurring sequences of words and multiple different distance measures were carried out. Representative results (the clustering of bigrams using the "classic" Delta) are presented below:

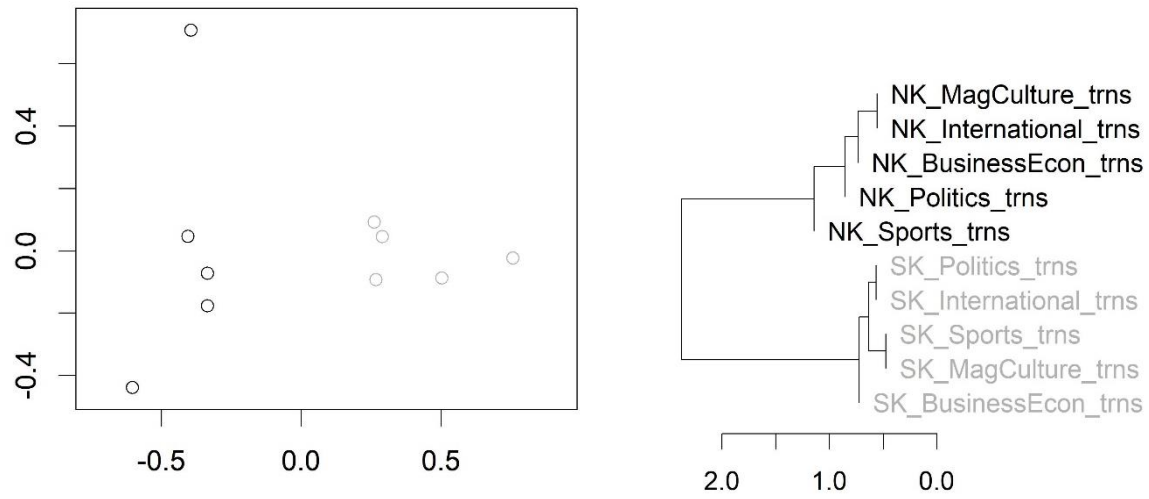

Figure 2: MDS Plot and Clustering Dendrogram of ROK and DPRK Journalistic Texts. Burrow's Classic Delta Distance, 100 MFW bigrams Culled at 0%

Clustering runs using the "classic" Delta distance were repeated for strings of two and three words (bigrams and trigrams). While it may be argued that analyzing still longer sequences of words could reveal furtherer patterns in the data terms of the use of collocations, the very low frequency with which such sequences of words appear in general (e.g. Hoover 2002: 162), but especially in a textual database as small as the one used here precludes such analysis. In addition, this clustering of bigrams and trigrams using the "classic" Delta distance were repeated using the three additional distance measures identified above used to verify the robustness of the single word clustering. All of the results for bigrams demonstrated a clustering schedule as above, forming two distinct clusters according to national origin. Clustering carried out on distances derived from the frequencies of sequences longer than three words were not so unambiguous. A notable outlier, the DPRK sports sub-corpus, often emerged as less similar to all other sub-corpora of DPRK journalistic writing than the combined cluster of ROK journalistic writing. The dendrogram below illustrates this situation:

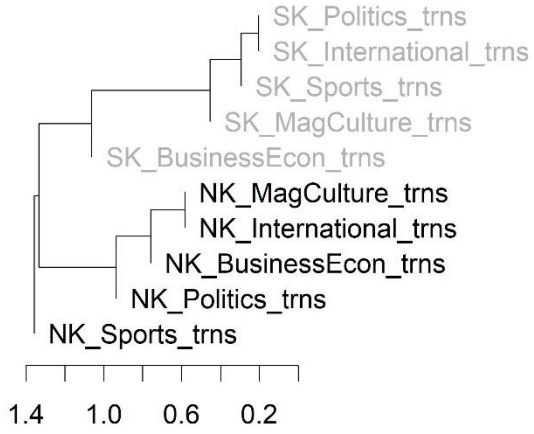

Figure 3: Clustering Dendrogram of ROK and DPRK Journalistic Texts. Burrow's Classic Delta Distance, 100 MFW trigrams Culled at 0%

Given the size of this textual-database, we are reluctant to ascribe significance to these results, though they may be regarded as an interesting point of departure for future work.

Finally, with regard to the clustering, we acknowledge that it is possible that the sorting of articles into thematically unified sub-corpora of roughly 5,000 words each may have played a role, since 5,000 words is appreciably longer than each journalistic article from which these sub-corpora are constructed. Consequently, clustering was carried out on sampled slices of these texts. These included clustering runs which divided each sub-corpus into ten sequential slices of five hundred words and which took ten random samples of five hundred words from each sub-corpus, again using the "classic" Delta distance as applied to the observed frequencies of single words and bigrams. Both the sequentially and randomly sampled smaller text slices also regularly clustered into two major sub-groups depending on their origin in the ROK or DPRK when distances were derived based on the frequency of single words.

The sampled clusters of bigrams, however, appeared to produce more nuanced results in that a cluster of both DPRK and ROK text slices emerged in the clustering run which used sequential sampling and an outlying cluster of DPRK sports texts emerged in the clustering run which used random sampling. Interpreting these results is challenging since they run contrary to the pattern which may be observed in the clustering runs so far. Tentatively, we suggest that the overall dissimilarity between ROK and DPRK journalistic texts is reinforced by these results. Stylometric methods are considered to increase in reliability as text length increases, therefore short, five hundred word samples of different national origin drawn from our textual database may cluster together for no more reason than chance. Despite this, the clustering does not appear to be completely random and there is an impressionistically noticeable tendency for large clusters of predominantly DPRK and ROK texts to form.

In order to provide a less impressionistic overview of these clustering runs, we performed bootstrap clustering. This produces a visualization known as a 'consensus tree' which is described as "a statistically justified 'compromise' [sic] between a number of virtual cluster analyses for a variety of MFW and Culling parameter settings" (Eder et al. 2017: 15). In other

words, an even more general picture of the relative (dis)similarity of the sub-corpora may be derived; one which simultaneously takes into account a smaller or larger number of features which appear more or less consistently in all the texts included in the database.

The bootstrap consensus trees below combine the results of 186 clustering runs performed on wordlists composed of between ten and three hundred words (increasing in tens) which appear in the whole corpus, then in at least ten, twenty, thirty, forty, and fifty percent of the texts. The structure of the different trees gives us some insight into which sub-corpora cluster together most frequently. Where the consensus strength is set to 0.9, direct linkages are made between sub-corpora which form such linkages in at least ninety percent of the clustering runs. In the other bootstrap consensus tree the consensus strength is set to 0.5, therefore direct linkages are made between sub-corpora which form such linkages in only fifty percent of the clustering runs.

Figure 4: Bootstrap Consensus Trees of Clustering Runs Performed on the 10, 20, 30, … 300 MFWs with Culled at 0%, 10%, 20%...50%. Consensus Strength 0.9 (above left) and 0.5 (above right).

The lack of visible structure for the journalistic texts produced in the ROK in the bootstrap consensus tree with the consensus set to a higher value (i.e. the fact that they are all linked at a single node) may be taken to indicate a relatively high level of internal stylistic homogeneity, since sub-clusters do not form between particular thematic sub-corpora with such near total regularity as for the journalistic texts produced in the DPRK. In the above left visualization we see clearly that specific topics or genres of DPRK journalistic texts regularly cluster together, specifically, the sub-corpora comprised of political journalism and cultural journalism cluster together, as do the sub-corpora of international journalism and business journalism. When the consensus strength is lowered, as in the visualization above right, we see some internal structure appear for the ROK journalistic texts. This may be interpreted as meaning that the sub-grouping of the sub-corpora of ROK sports and business journalism may be somewhat stylistically distinct from the sub-grouping composed of the sub-corpora of international, cultural, and political journalism, since the former two do not cluster together with any other specific sub-corpus in more than fifty percent of the clustering runs.

Thus, we feel confident in concluding that, despite our relatively small textual database, journalistic texts produced in the ROK and the DPRK may be distinguished on the basis of frequently used word and sequences of words. Put simply, they are written in different styles.

We now move on to the question of which words played the greatest roles in the disambiguation of journalistic texts by national origin demonstrated above and what they may suggest about the stylistic divergence between these countries' journalistic language. We take this opportunity, though, to emphasize that the textual database upon which these findings are based is relatively very small and the lexical discriminators identified in this instance between ROK and DPRK journalistic texts should not be generalized.

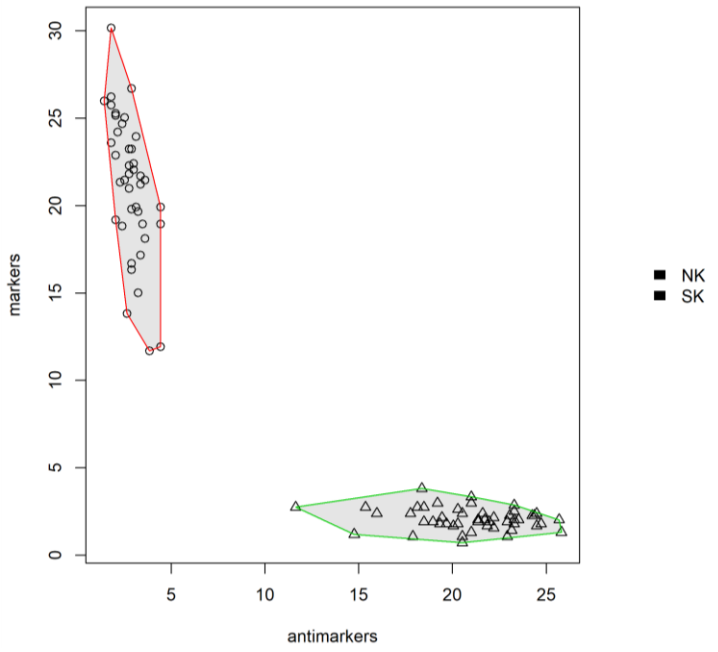

Figure 5: Visualisation of Discriminators used in 1,000 word slices of ROK and DPRK Journalistic Texts. Craig's Zeta

The above graph analyses the (dis)similarity of the language included in our database in terms of words preferred and avoided in journalistic texts produced in the DPRK. The whole corpus was analyzed for frequently occurring words, then the words most significantly preferred and avoided in DPRK journalistic writing were identified (in this case around 70 words each). Each sub-corpus was then sliced into consecutive 1,000 word sections with an overlap of 500 words. The plot above shows the number of instances of appearance of this smaller set of consistently preferred words (markers) against the number of instances of appearance of the small set of consistently avoided words (antimarkers) in the 1,000 word textual slices of both ROK and DPRK journalistic texts.

This visualization emphasizes the findings of the cluster analysis, namely, that the national origin of Korean journalistic texts may be clearly identified from their stylistic features. Further, we see that the slices of the DPRK texts show greater heterogeneity of style, at least in terms of the frequency of use of the discriminators identified as marker of DPRK texts. Both samples

of texts slices, however, are consistent in their avoidance of words which characterize the texts in the other set. That is, all the ROK text slices very rarely use words which are characteristic of DPRK texts and vice versa.

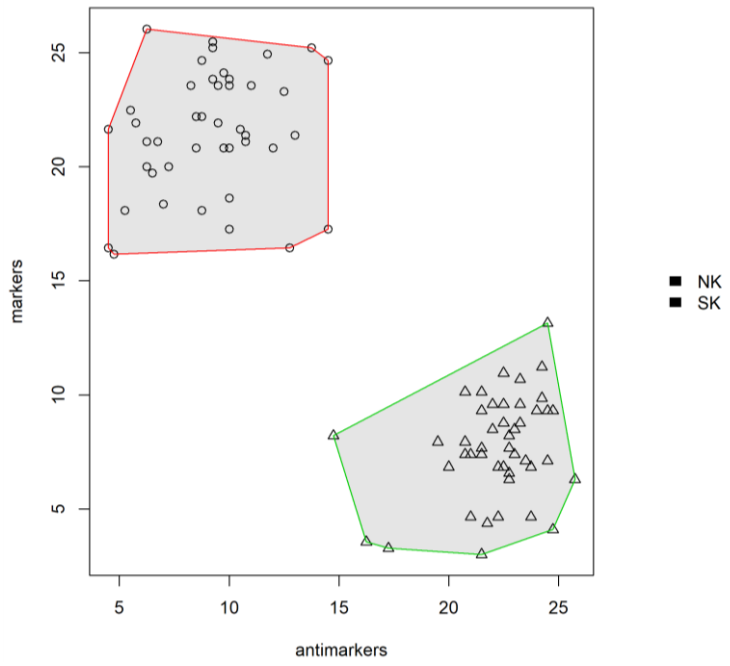

Figure 6: Visualisation of Discriminators used in 1,000 word slices of ROK and DPRK Journalistic Texts. Eder's Zeta.

The above visualization presents a less clear-cut distinction between the DPRK and ROK text slices, but does still clearly allow us to distinguish between them on the basis of the discriminators automatically identified. This may be attributed to the difference in distance measure used in these tests, with Eder's Zeta being based on the Canberra distance, and its resultant sensitivity to rare word occurrences (Eder et al. 2016: 26).

Inspection of the lists of words identified as discriminators between ROK and DPRK journalistic style allows us to make impressionistic observations. Chief among these is that the journalistic texts produced in the ROK appear to be less evaluative than those produced in the DPRK. Of the twenty words preferred most significantly in the DPRK journalistic texts, five of them (*jŏkkŭk* – 'active', *ch'ŏljŏhi* – 'thoroughly', *hyŏnmyŏnghan* – 'wise', *jungyohan* – 'important', *matke* – 'correctly') arguably connote a positive evaluation. There is another

feature of the style of these DPRK texts which may be inferred from the list of preferred words that is both less pronounced and harder to articulate. The words *uri* – 'us/our', *chosŏn* - 'Korea', and *nara* – 'country' were found to appear preferentially in texts of DPRK origin, even though the topics of the sub-corpora drawn from both ROK and DPRK sources were similar and included international news. While this may fall more in the realms of editorial guidelines than linguistic style *per se*, it may be inferred from this that a feature of DPRK journalistic writing is to write from a perspective which relates what is reported directly back to the nation and its people.

Turning to the ROK texts, we find a very different set of words, as implied by the visualizations above. We may characterize these texts as factual reportage (or at least adopting a style which aspires to it), again in respect to five of the most significantly preferred words (*haessta* – 'did/said', *ko* – 'marker for direct quotation', *irago* – 'marker for direct quotation', *ttarŭmyŏn* – 'according to', *nat'anassta* – literally 'appeared' or 'presented itself', but used in the common sentence ending *...gŏsŭro nat'anassta* 'to appear to be the case' or 'to be shown to be the case' when reporting information). We further note that words concerned with temporal sequence and location in space were preferred in ROK texts (*ihu* – 'thereafter', *jŏn* – 'before', *jiyŏk* – 'region', *si* – 'city'), which further implies that striving to provide precise factual information is a characteristic of this style.

At the close of this section, it must be reiterated that these findings are highly tentative and provisional in nature. Only the exploration of more extensive corpora will be able to determine the extent to which these observations represent general tendencies in ROK and DPRK journalistic writing. Nevertheless, while their stylistic characterization remains an open question, the finding that journalistic texts produced in the ROK a DPRK may be distinguished on the using quantitative stylometric techniques may be made with greater confidence. Finally, although this pilot study has only examined published language and no data were presented for

broadcast language, we consider it highly likely that a similarly large difference in style would be found in both registers.

4. **Prospectus**

While the results of the pilot study presented in this paper are highly suggestive of extensive differences in style between texts of nominally the same genre produced in the DPRK and ROK, it must also be placed in the wider context of what could be achieved using either methods similar to those outlined above or which likewise draw on the digital humanities paradigm for either data gathering or analysis. The explanatory power of the above results and the extent to which they may be generalized to the language of the DPRK are limited when compared to what might be achieved with larger corpora compiled with different or less specific aims in mind. Furthermore, there is currently no consensus on precisely which quantitative techniques are most applicable to problems such as the divergence of the Korean language and other fundamental questions, such as how texts should be prepared for such analyses, are only just beginning to be examined.

In lieu of a conclusion, then, we present a prospectus for the integration of digital humanities methods into this particular area of Korean linguistics. In the first instance, the data upon which studies of the linguistic divergence between the DPRK and ROK could be dramatically expanded. Sources outside of prescriptive works on language or limited face-to-face interaction could be collated from on-line sources or digitized versions of printed works. These sources need not be limited to written language appearing in journalistic or literary works, but could also incorporate the increasing amount of spoken language appearing in audio and video recordings disseminated online. With appropriate archiving and curation, this would allow both a wider range and a larger volume of data to be examined when pursuing research on this topic. Access to this data would then enable the investigation of wholly novel research questions, such as that posed in the pilot study above and provide insights into more general trends of language use than we are currently able to discern with more the more intuitive yet impressionistic qualitative techniques currently at

our disposal. The broad trends in linguistic divergence which could be identified need not supplant the excellent, fine-grained research into the linguistic features which are taken as representative of linguistic divergence but would rather complement it and lend it empirical weight, leading to a more comprehensive, understanding of the phenomenon which incorporates usage along with phonology, morphology, and vocabulary.

To conclude, we propose that many insights which could potentially be gained from research carried out with reference to the framework laid out above would not only be of purely academic interest, but could play a role in practical linguistic re-unification.

5. **References**

Burrows, J. "Delta: A Measure of Stylistic Difference and Guide to Likely Authorship." *Literary and Linguistic Computing* 17,3 (2002): 267-287

- "All the Way Though: Testing for Authorship in Different Frequency Strata." *Literary and Linguistic Computing* 22, 1 (2007): 27-48

Cedergren, H.J. and Sankoff, D. "Variable Rules: Performance as a Statistical Reflection of Competence". *Language* 50, 2 (1974): 333-355

Cho Chae-su. *Nambukhanmal pigyosajŏn: nambukhan, chungguk, chungangasia esŏ 3man ŏhoe rŭl karyŏ moun kyŏre mal sajŏn* [A comparative dictionary of North and South Korean language: a dictionary of 30,000 words of the common folk language of North and South Korea, China and Central Asia]. Seoul: Hankyŏre ch'ulp'an, 2007.

Choi Jeong-ho and Kim Seong-gun. *Chosŏnŏgyubŏmch'ŏnsa* [The History of the Development of Korean Language Standardisation]. Pyeongyang: Sahoegwahakch'ulp'ansa, 2005.

Choi Yun-gap and Jeon Hak-seok. *Chungguk chosŏn han'guk chosŏnŏ ch'ai yŏn'gu* [Research on the differences between the Korean of China, North Korea and South Korea]. Seoul: Munhwasa, 1994.

Craig, H. "Stylistic Analysis and Authorship Studies". In *A Companion to Digital Humanities*, edited by Schreibman, Susan, Siemens, Ray, and Unsworth, John. Oxford: Blackwell, 2004

Eder, M., Kestemont, M. and Rybicki, J. (2013). "Stylometry with R: A Suite of Tools." In Digital Humanities Conference Abstracts, University of Nebraska-Lincoln, NE. (2013): 487-489

- (2015). 'Stylo': A Package for Stylometric Analysis

Eder, M., Rybicki, J. and Kestemont, M. "Stylometry with R: a package for computational text analysis." *R Journal* 8, 1 (2016): 107–121

Gardiner, E. and Musto, R.G. *The Digital Humanities: A Primer for Students and Scholars*. New York, NY: Cambridge University Press, 2015

Hajič, J. "Linguistics Meets Exact Sciences". In *A Companion to Digital Humanities*, edited by Schreibman, Susan, Siemens, Ray, and Unsworth, John. Oxford: Blackwell, 2004

Hong Yun-pyo. "'Kyŏremalk'ŭnsajŏnŭi p'yŏnch'an panghyang [The Direction of Dictionary Compulation for 'The Unabridged Unified Korean Dictionary]". *Hanguk sajŏnhak* [Korean Lexicography] 9 (2007): 23-52

Hoover, D.L. "Frequent Word Sequences and Statistical Stylistics". *Literary and Linguistic Computing* 17, 2 (2002): 157-180

Jang Chung-deok. "Ch'ŏngjujiyŏkŏ ch'eŏn ŏganmal jaŭm gyoch'eŭi sahoecŏk byŏni" [Social Variation of the Replacement of Word Final Consonants in the Substantives of Cheonju Regional Speech]. *Bangŏnhak* [Dialectology] 22 (2015): 253-278

Jeong Seung-chol. *Han'gukŭi pangŏn gwa pangŏnhak* [Korean Dialects and Dialectology]. Paju: T'aehaksa, 2013.

Kang Bo-seon, Kim Jin-suk, and Park Su-ryeon. "Pukhanŭi 2013 kaejŏng kugŏgwa kyoyukkwajŏnguy t'ŭkching" [The Characteristics of the Revised Korean Language Curriculum of North Korean in 2013]. *Kugŏgyoyugyŏn'gu* [National Language Education Research] 62 (2016): 1-34

Kang Kyoungwon. "Namhanŭi pangŏnjiyŏk kubun" [A Classification of South Korea's Dialect Areas]. *Munhwayŏksajiri* [Cultural and Historical Geography] 26,1 (2014): 34-49

Kim Dong-chan. *Chosŏnŏgyubŏmniron* [The Theory of Korean Language Standardisation]. Pyeongyang: Sahoegwahakch'ulp'ansa, 2005.

Kim Il-hwan, Lee Do-gil, Lee Sang-hyeok, Moon Han-byeol. "Sigyeyŏl kongiŏ net'ŭwŏk'ŭ punsŏgŭl iyonghan yuuyŏ punsŏk" [Longitudinal Network Analysis of Synonym Collocates]. Paper presented at the 13th Conference of the International Society for Korean Studies, Auckland, New Zealand, August 3-4, 2017.

Kim Il-sung. "Chosŏnŏrŭl paljŏn sik'igi wihan myŏt kaji munje." [Several Problems in the Development of the Korean Language]. In *Kimilsŏng chŏjakcip 18kwan* [Kim Il-sung's collected Works Vol. 18], 14-27. Pyeongyang: Cosŏnnodongdang ch'ulp'ansa, 1964 [1979]

- "Chosŏnŏŭi minjokcŏk t'ŭksengŭl olhke sallye nagalde te hayŏ: ŏnŏhakchadŭlgwa tamhwa" [On the Correct Preservation of the Ethnic Characteristics of the Korean Lanuage: A Dialogue with Linguists]. In *Kimilsŏng chŏjakcip* 20*kwan* [Kim Il-sung's Collected Works Vol. 20], 335-352. Pyeongyang: Cosŏnnodongdang ch'ulp'ansa, 1966 [1979]

Kim Kwang-su. "Yŏnbyŏnesŏ pon nambukŭi ŏhoe: chŏnmunyongŏrŭl chungsimŭro [North Korean and South Korean Vocabulary in Yanbian: Focus on Technical Terms]. In *Chosŏnŏ koch'algwa yŏngu* [Perspectives and Research on the Korean Language], edited by Kim Kwang-su, 437–447. Yanji, PRC: Yŏnbyŏninminch'ulp'ansa, 2012a.

- "Pungnam mich' chungguk cosŏnŏ munbŏbyongŏ sayongŭi ch'ai wa t'ongilŭi pangan." [The Difference between North Korean, South Korean and Chinese-Korean Usage of Grammatical Terms and a Framework for their Unification]. In *Chosŏnŏ koch'algwa yŏngu* [Perspectives and Research on the Korean Language], edited by Kim Kwang-su, 416–436. Yanji, PRC: Yŏnbyŏninminch'ulp'ansa, 2012b.

Kim Sang-jun. *Nambwukhan podo pangsongŏnŏ yŏn'gu: uri ŏnŏŭi t'ongjilsŏng hoebogŭn kanŭnghan'ga?* [Research into the broadcast language of North and South Korean reporting: is the recovery of our language's homogeneity possible?]. Seoul: K'ŏmyunik'esyŏn puksu, 2002.

Kumatani Akiyasu. "Language Policies in North Korea". *International Journal of the Sociology of Language* 82 (1990): 87-108

Kwon Jae-il. "Nambukhanuy ŏnŏhak chŏnmunyongŏ p'yojunhwa pangan yŏngu [Research on the Standardisation of the Linguistic Terminology of South and North Korea]". *Hangul* 274 (2006): 231-266

Labov, W. *The Social Stratification of English in New York City*. Washington D.C.: Center for Applied Linguistics, 1966

Lebart, L. and Rajman, M. "Computing Similarity". In *Handbook of Natural Language Processing*, edited by Robert Dale, Hermann Moisl, and Harold Somers, 493-524. New York, NY; Basel: Marcel Dekker Inc., 2000

Lee Dong-ju, Yeon Jong-hum, Hwang In-beom, Lee Sang-gu. "Kkokkoma: kwan'gyehyŏng teit'ŏbeisŭrŭl hwalyonghan sejong malmungch'i togu" [Kkokkoma: A Relational Database Tool for the Sejong Corpus]. *Chŏngbogwahakhoenonmunji: k'ŏmo'yut'inguy slice mit let'ŏ* [Journal of KIISE: Computing Practices and Letters] 16, 11 (2011): 1046-1050

Lee Guk-lo. "Chosŏnmalŭi sat'uri" [The Dialects of Korean]. *Tonggwangsa* 29 (1932): 9-12

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

Lee Ki-moon and Ramsey, S. Robert. *A History of the Korean Language*. Cambridge: Cambridge University Press, 2011.

McEnery, T. and Oakes, M. "Authorship Idenification and Computational Stylometry". In *Handbook of Natural Language Processing*, edited by Robert Dale, Hermann Moisl, and Harold Somers, 563-580. New York, NY; Basel: Marcel Dekker Inc., 2000

Ministry of Unification. "Nambukhan ŏnŏbigyo" [Linguistic comparison of North and South Korean]. Accessed November 14, 2017.
http://nkinfo.unikorea.go.kr/nkp/term/skNkLangCompare.do

National Language Association "Che 1pu pyojunŏ sajŏng wŏnjik." [Section One of the Principles of the Standard Language] Accessed November 13, 2017.
http://www.korean.go.kr/front/page/pageView.do?page_id=P000085&mn_id=94

National Language Committee. *Chosŏnmalgyubŏmchip* [A Collection of Standardised Korean]. Pyeongyang: Sahoegwahagwonch'ulp'ansa, 1988.

Pak Yeong-sun. *Han'gugŏŭi sahoe ŏnŏhak* [Korean Sociolinguistics]. Seoul: Han'guk Munhwasa, 2001

Park, E.L. and Cho Sungzoon. "KoNLPy: Korean Natural Language Processing in Python". Paper presented at the 26th Annual Conference on Human and Cognitive Language Technology, Chuncheon, Republic of Korea, October 2014

Park See-Gyoon. "Nambukhan ŏnŏe taehan pigyo yŏn'gu: parŭmgwa ŏhwirŭl chungsimŭro [A Contrastive Study on the South and North Korean Languages: Focused on Pronunciation and Vocabulary]. *Kukŏmunhak* [National Language and Literary Studies] 38 (2003): 29-54

- "Nambukhan ŏnŏe taehan pigyo yŏn'gu 2: hwasulgwa hwabŏp, pangsonghwabŏpŭl chungsimŭro [A Contrastive Study on the South and North Korean Languages 2: Focused on Narrative Speech and Broadcast Speech]. *Kukŏmunhak* [National Language and Literary Studies] 39 (2004): 117-141

Prospect. *Computers and Humanities* 1 (1966): 1–2

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, 2017

Rayson, P., Leech, G. and Hodges, M. "Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus". *International Journal of Corpus Linguistics* 2,1 (1997): 133-152

Rybicki, J. and Eder, M. "Deeper Delta Across Genres and Languages: De We Really Need the Most Frequent Words?" *Literary and Linguistic Computing* 26, 3 (2011): 315-321

Sahoe kwahagwŏn ŏnŏhak yŏn'guso. *Hyŏndae Chosŏnmal sajŏn* [Contemporary Korean Dictionary]. Pyeongyang: Kwahak, Paekkwasajŏn ch'ulp'ansa, 1981.

Song Jae-jung. "South Korea: Language Policies and Planning in the Making." *Current Issues in Language Planning* 13 (2012): 1–68

Yeon Jaehoon. "Standard Language' and 'Cultured Language". In *Korean Language in Culture and Society*, edited by Sohn Ho-min, 31–43). Honolulu, HI: University of Hawai'i Press, 2006.

Yeonhap. *Pukhan ŏhoe sajŏn* [North Korean Glossary]. Seoul: Yeonhap Nyusŭ, 2002.