Division and the Digital Language Divide: A Critical Perspective on Natural Language Processing Resources for the South and North Korean Languages

Benoit Berthelier

Korean Studies, Volume 47, 2023, pp. 243-273 (Article)

Published by University of Hawai'i Press
DOI: *https://doi.org/10.1353/ks.2023.a908624*

Division and the Digital Language Divide: A Critical Perspective on Natural Language Processing Resources for the South and North Korean Languages

Benoit Berthelier

The digital world is marked by large asymmetries in the volume of content available between different languages. As a direct corollary, this inequality also exists, amplified, in the number of resources (labeled and unlabeled datasets, pretrained models, academic research) available for the computational analysis of these languages or what is generally called natural language processing (NLP). NLP literature divides languages between high- and low-resource languages. Thanks to early private and public investment in the field, the Korean language is generally considered to be a high-resource language. Yet, the good fortunes of Korean in the age of machine learning obscure the divided state of the language, as recensions of available resources and research solely focus on the standard language of South Korea, thus making it the sole representant of an otherwise diverse linguistic family that includes the Northern standard language as well as regional and diasporic dialects. This paper shows that the resources developed for the South Korean language do not necessarily transfer to the North Korean language. However, it also argues that this does not make North Korean a low-resource language. On one hand,

Benoit Berthelier is a lecturer in the Department of Korean Studies at the University of Sydney, Australia (benoit.berthelier@sydney.edu.au).

South Korean resources can be augmented with North Korean data to achieve better performance. On the other, North Korean has more resources than commonly assumed. Retracing the long history of NLP research in North Korea, the paper shows that a large number of datasets and research exists for the North Korean language even if they are not easily available. The paper concludes by exploring the possibility of "unified" language models and underscoring the need for active NLP research collaboration across the Korean peninsula.

Keywords: large language models, artificial intelligence, natural language processing, North Korea, ideology

Introduction

In 1997, a whitepaper published by the South Korean Ministry of Culture and entitled *National Competitiveness and the Digitization of the Korean language* noted, with a sense of urgency, that "developed countries like America, Japan and, Europe have been steadily investing in the development of natural language processing technologies since the dawn of the computer age."[1] The computers of the future, surely, would understand English but would they understand Korean? The picture was dire: "It is not even that there is a dearth of valuable Korean language data on the Internet, there simply isn't any." Without any data to develop natural language processing (NLP) systems for its own language, South Korea would end up a digital and linguistic colony, "degraded to a nation that consumes NLP systems made in Japan or America."[2] To counter this trend, the Ministry of Culture launched a plan, the *Sejong Plan for the 21st Century*, that would invest in Korean-specific NLP resources such as digital dictionaries, corpora, and software.[3]

Twenty-five years later, these fears are largely assuaged. Korean stands on the safer side of the "digital language divide"—the inequal level of digital resources available for different languages.[4] While Chinese and English content accounts for 40% of the web, Korean nonetheless represents 1% of all online content.[5] The language also belongs to a small category of about a dozen "high-resource" languages—that is, languages endowed with a large number of NLP resources.[6] A recent survey on digital linguistic diversity classified Korean as belonging to a group of up-and-coming languages that possessed "dedicated NLP communities" and had "the potential to become winners and enjoy the fruits of 'digital superiority'."[7] While the South Korean government cannot be solely

credited with the reasons for this success, the Sejong Plan did produce some of the largest and most commonly used Korean language corpora, upon which a number of NLP software were developed in the aughts.[8] Other initiatives by universities and public institutions led to the release of freely available corpora such as the KAIST corpora or the National Information Society Agency's AI HUB.[9] More recently, the AI-research branches of private companies (Kakao Brain, Naver AI Lab, SKT Brain, etc.) have become increasingly important players, in particular in the production of large language models (LLMs) for the Korean language.[10]

However, categorizing Korean as a high-resource language does not take into account the diversity and divisions that exist within the language. Like all other official national idioms, Korean has a number of variants tied to context, class, and geography (both within and outside of the peninsula). But the English term "Korean" used without any further qualifiers conceals the more specific pluricentric[11] nature of the language: the existence of two separate underlying national languages, the Southern standard language (*p'yojunŏ*) and the Northern standard language (*munhwaŏ*). Between the two, publicly available research and resources are virtually exclusively dedicated to the former.

But is such a distinction meaningful when gauging the availability of NLP resources? After all, both South and North Korean languages share a common origin, deriving from earlier efforts—in the colonial period—to formalize a modern, standardized national language based on the Seoul dialect and under the influence of Japanese linguistics.[12] But, as the first part of this paper demonstrates, while NLP models trained on South Korean data can be used on North Korean texts, they can be expected to perform significantly worse than models trained on North Korean data specifically. This is due not only to morphological and syntactic differences. As I argue in the first part of the paper after demonstrating the extent of the semantic gap between contemporary South and North Korean language using a technique known as word embeddings, there are significant differences in *meaning* between the same words in the two languages, which can be attributed to the ideological and cultural gap between the two countries.

If these observations emphasize the need to rely on North Korean language data for NLP, it does not necessarily imply that North Korean is a low-resource language. Indeed, South Korean resources can still be used as a basis for the development of North Korean-specific NLP models, thus limiting the amount of data necessary to develop them. Furthermore, I argue that the development of LLMs with the ability to encode polysemy

into contextual representations[13] opens the possibility of producing unified language models trained on equally large amounts of South and North Korean data. I offer a benchmark of language models trained using different strategies on North and South Korean data to show the large performance increase that can be obtained on NLP tasks by combining resources developed for the South Korean language with North Korean language-specific datasets.

Compiling sufficient digitized North Korean textual data for such tasks may seem difficult in light of North Korea's relatively small presence in the global digital world. But, as the final part of this paper asserts, the North Korean language is in fact quite resource-rich—its resources merely suffer from a problem of accessibility. Measures of a language's NLP resources typically look at textual content availability on the Internet, publicly available datasets, and academic research. As internet access is extremely limited in the Democratic People's Republic of Korea (DPRK) and locally produced research is hardly ever consultable overseas, North Korea would clearly appear as a low-resource language. Yet looking at the history of digitization and the development of NLP and machine learning research in the country since the 1980s shows that the North Korean language is far less destitute than commonly assumed. The problem is therefore not about the production of resources but about the ability to access and share them.

Morphosyntactic and Lexical Differences and Their (Minor) Implications

North and South Korea have been divided since 1945 and over time a notable number of linguistic differences have emerged between the languages spoken on each side of the 38th parallel today. These differences have been attributed to a number of factors, from diverging language policies[14] and ideological causes[15] to lifestyle differences[16] and preexisting geolectal variations.[17] There exists, particularly in South Korea, a very large body of scholarship dedicated to cataloging and monitoring the diverging evolution of the two languages.[18] While a few authors contend that the two languages have diverged a great deal,[19] most agree that differences are limited to vocabulary and that contemporary North and South Korean speakers still have no difficulty understanding each other.[20] This section maps out these differences and explores their implication for natural language processing systems. To investigate lexical differences, I compiled

and compared lexical corpora of each language. For the North, I extracted the index of the reference dictionary in North Korea, the *Great Dictionary of Korean Language* from a digital application as well as entries from the *Korean Dictionary of Word Frequencies*.[21] For the South, I used a list of the most frequent words in the South Korean language compiled by the National Institute of Korean Language (NIKL) and the institute's *Dictionary of Standard Korean*.[22]

Even if North and South Korean are mutually intelligible for human speakers, differences that may be trivial to resolve for them can prove difficult to handle for a computer algorithm. Indeed, it is an often-observed phenomenon that the performance of NLP systems degrades when applied to language variations (of the type observed between pluricentric languages, but also even diachronic variations of the same language).[23] For instance, while minor changes such as the addition or omission of a silent letter in a word's spelling may not even be noticed by a human speaker, it may be enough for a dictionary-based algorithm to consider the alternative spelling as a completely unknown term. Therefore, while cataloging all the differences between the two languages is outside the scope of the present paper (not to mention already well-trodden ground), it is nonetheless important to single out which of these differences can affect the performance of NLP systems and how. Note that, as the focus of the study will be on textual inputs, phonetic and phonologic differences will be disregarded.

There are few grammatical differences between North and South and, as a result, common NLP tools such as morphosyntactic taggers and dependency parsers[24] trained on one language may still be used on the other. Most grammatical differences really are differences in the usage of a form rather than in the existence of language-specific forms: a certain clause structure or modifier will be used more commonly in the North than the South but nonetheless exist in both.[25] For instance, the form ~ŭl te ~ (으)ㄹ 데 is common in the North where in the South -nŭn kŏt ~는 것 would be used. In certain cases, the difference may actually be interpreted as semantic: the widespread use of the deferential style in conversation in the North where the polite style would be common in the South indicates that the meaning of each style has come to differ.

More serious (perhaps somewhat counterintuitively), are differences in spacing (*ttŭŏssŭgi*), with North Korean orthographic rules and practice resulting in lengthy compound words being commonplace. While mainly a matter of comfort for a human reader, the different ways in which words are separated within a text has tremendous influence on an NLP system's

performance. The first step of any NLP pipeline is tokenization—the breaking down of the text into smaller semantic units which will constitute the vocabulary of terms the system uses.[26] Tokenization often relies on space: either because it uses the space in the text to separate words or because it relies on machine learning models such as hidden Markov models or conditional random fields that were trained on text spaced in a certain way to separate words regardless of how they were originally segmented in the text. In the latter case, a tokenizer trained on Southern data would simply reproduce the more granular spacing used in the South when processing a Northern text. This may still be practical for certain tasks such as building a classifier or an index for a search engine. However, because differences in spacing also stem from differences between how North and South Korean grammarians define syntactic units, one must be aware that the results obtained from tools such as part-of-speech taggers and parsers will reflect a Southern theory of Korean grammar.[27] For a number of NLP systems, from language models to classifiers, it is, however, possible to bypass the issue of spacing altogether. Over the past few years character or morpheme-based tokenizers such as WordPiece,[28] which do not rely on word boundaries, have proven extremely effective to handle spacing and even spelling differences between language variations.

Divergences in lexicon are widely acknowledged to be the most differentiating factor between North and South, but the breadth and relevance of that gap may be overstated, both for human speakers and NLP systems. Globalization in the South has led to the introduction of a large number of English-based loanwords while in the North language policies to "purify" the language of foreign loanwords and Sino-Korean words have resulted in the creation of neologisms based on pure Korean words.[29] While loanwords account for less than 3% of the North Korean lexicon, they make up 5 to 10% of the South Korean one.[30] The two languages also employ different spelling rules, for instance for certain initial consonants of Sino-Korean words (initial ㄴ and ㄹ are replaced by ㅇ in the South) and certain final consonants in compound words.

To evaluate the scope of these lexical differences, we can look at a list of the most common words in one language and see if these exist in a general-purpose dictionary of the other. For instance, naïvely attempting to match the most common South Korean words to an entry in the North Korean dictionary yields a match for less than half of the words. While this result may seem dramatic, looking at the orphan words in more detail reveals much fewer differences. The aforementioned difference in spelling rules (for initial and final consonants) is responsible for 3% of the missing

cases, and can easily be addressed, in NLP systems, with a dictionary or a heuristic conversion algorithm; 17% are foreign loanwords, one third of which are mismatched due to spelling differences (*t'ellebijŏn* vs. *t'ellebijyon* for "television," *p'ŏsent'iji* vs. *p'osent'eji* for "percentage") that can likewise be addressed with an additional dictionary or morpheme-based tokenization. The remainder consists of compound Sino-Korean words that for the most part do also exist in North Korea (e.g., *sinsedae* for "new era," *pangsongsa* "broadcasting company") but are not in the dictionary due to its indexing methodology. The opposite operation of trying to match common North Korea words with entries in a South Korean dictionary does not yield as many orphan words (less than 20% of the words do not have a match in the South Korean dictionary).[31] And spelling differences (*kŏp'usi hada* vs. *kŏp'ususu hada*, *chinggŭl sŭrŏpta* vs. *chingkŭrŏpta*) again account for a large number of cases. Both languages thus remain morphologically similar with only minor and easily addressable differences.

Semantic Drift Between North and South

If contrasting grammars and lexica only reveals relatively trivial differences, the comparison does not account for semantic divergences between the two languages. This difference in meaning can manifest itself in two ways: a difference in what words mean, but also a difference in what the languages are used to talk about. Both have implications for the transferability of NLP resources from one language to the other, particularly for what has now become the "standard way to represent word meaning in NLP": word embeddings.[32] Word embeddings are abstract, numerical representations of the meaning of words learned from their distributions in a large corpus. Embeddings can be *static*, representing the meaning of word types (the dictionary definition(s) of a word), or *contextual*, representing the meaning of word tokens (the meaning of a word in a particular context).[33] Since in both cases meaning is derived from the training corpora, cultural and ideological differences between North and South can have a larger impact than linguistic differences on the performance of NLP models.

To illustrate and quantify semantic drift between the South and North Korean language, I start by using two sets of word embeddings trained on large corpora of texts from each language.[34] Word embeddings are sets of coordinates in a vector space and it is possible to measure the semantic similarity between two words by measuring the distance between their coordinates.[35] Likewise, we can list the synonyms of a word in order of

similarity by finding its *n* closest neighbors (*n* being the number of synonyms in the list). Because embeddings for each language are in a different vector space, the coordinates of a word in one set of embeddings cannot be directly compared with its set of coordinates in the other without further mathematical adjustments. However, one may use, for each word, the number of common synonyms among the word's *n* closest neighbors in each set of embeddings.[36] Boggust, Carter, and Satyanarayan thus introduce a method relying on the Jaccard similarity measure (intersection over union, i.e., the number of common synonyms over the total number of synonyms, or, in our case, the number of synonyms the same word has in both the South and North Korean embedding sets divided by the sum of the number of synonyms in each set), which I will reuse here.[37] Taking the intersection of both sets of embeddings and $n = 100$, I assign to each word a similarity score between 0 (low semantic similarity between North and South) and 100 (perfect semantic overlap).

Ordering the resulting lexicon by similarity lets us see what has remained semantically stable despite division. We find that musical words (*paiollin* "violin," *p'iano* "piano," *kayagŭm* "Korean zither"), toponyms (*kyŏngsang-bukto* "North Kyŏngsang province," *kyŏngju-si* "Kyŏngju city"), colors (*saek* "color," *hŭinsaek* "white," *p'urŭn saek* "blue/green"), kinship terms (*ttal* "daughter," *ŏmŏni* "mother," *samch'on* "uncle"), and penal vocabulary (*ch'ep'o* "arrest," *sŏn'go* "sentence," *sahyŏng* "death penalty") retain the largest amount of semantic similarity across the 38th parallel. These results are unsurprising: for instance, musical words are all likely to co-occur with the same words in both languages. More informative would be to rank the most dissimilar words but, as a very large amount of words do not share any synonyms, the results are less easily interpretable. What is possible, however, is to plot the distribution of similarity scores as in Figure 1 to get an overview of the overall degree of difference between the two languages.

The distribution of similarity scores is strongly skewed to the left indicating that most words in the corpus have low semantic similarity and that embeddings trained on a South Korean corpus will, for over 75% of the words learn a representation that is different from the representation that can be extracted from a North Korean corpus. It is worth noting, however, that we are talking here about semantic similarity as used by NLP algorithms. While it may offer a proxy for semantic differences experienced by a human speaker, it is not the same thing (i.e., the results do not mean that 75% of the vocable is not mutually intelligible to human speakers). To better gauge how one might interpret this distribution, we can compare it

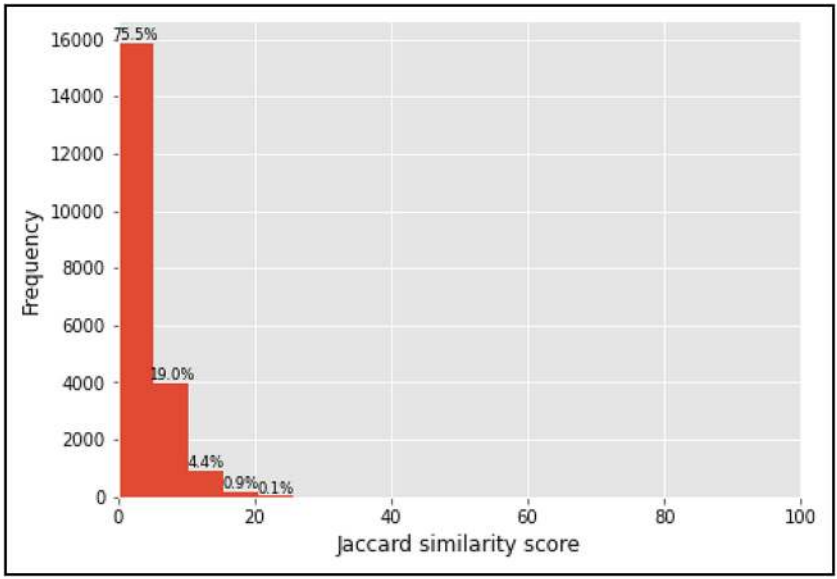

Fig. 1. Distribution of lexical similarity scores between North and South Korean, for each word in both sets of embeddings. Words with a high Jaccard similarity score retain the same meaning in both corpora, words with a low score have drifted semantically; 75.5% of the words in the corpus have a score between 0 and 5, indicating a high semantic drift. Only 0.1% of the words have a score above 20.

to a similar graph comparing a different form of variation. Figure 2 shows the result of the application of the same methodology to diachronic variations of English, with one set of embeddings trained on English language textual data from 1800 to 1810 and the other on data from 1990 to 2000.[38] While both figures exhibit an overall similar distribution, the distribution of Figure 1 is nonetheless more strongly skewed to the left, suggesting that the semantic difference between North and South Korean today are akin, yet slightly more marked, than the differences in the English language at two centuries of distance.

To better grasp what the semantic differences across the corpus might look like, we would need to be able to compare the South Korean and North Korean vectors of the same word. We cannot naively compare these as the embeddings are in different vector spaces, but several methods exist to learn a linear transformation from one set of embeddings to another given a set of common reference points.[39] This, in turn, allows one to project a word from one set of embedding into another and see what words it is similar to in this other space. For instance, in our case, we could take

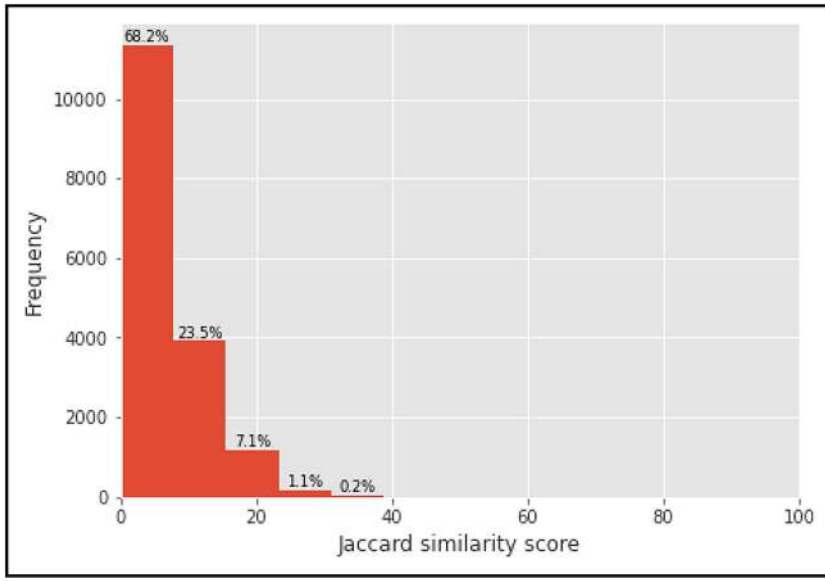

Fig. 2. Distribution of lexical similarity scores between early nineteenth century and late twentieth century English; 68.8% of the words in the corpus display a high semantic drift.

the coordinates representing a word in the North Korean embedding set, project them into the South Korean embedding's vector space and see what words the coordinates are the closest to (the reverse operation is likewise possible). More succinctly, this can be thought of as expressing the North Korean meaning of a word with a Southern vocabulary. We can then, using a dimension reduction technique, offer a simple 2D visualization of where a word stands in the other language's vector space. The following figures do exactly that for the word "socialism" (*sahoejuŭi*) after using the standard orthogonal Procrustes technique to align vector spaces.[40] Figure 3 represents the projection of the South Korean word vector for "socialism" and the North Korean word vector for "socialism," along with their respective nearest neighboring words, in the North Korean vector space. Figure 4 does the opposite, projecting the North Korean word in the South Korean vector space.

In Figure 3, we see that the term North Korean term "socialism" is associated with a number of grandiloquent qualifiers, representative of the way the ideology would be described in official discourse. But when what the South Korean language means by "socialism" is translated into North Korean terms, we find it associated with words such as "dictatorship,"

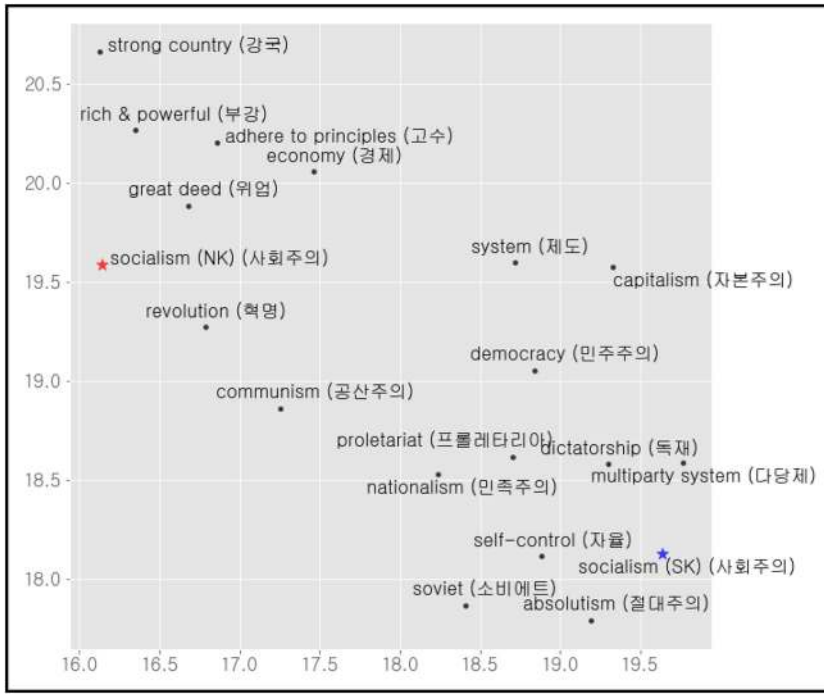

Fig. 3. Projection of the South Korean word vector for "socialism" (blue star) in a North Korean vector space. The word is surrounded by its nearest neighbors in the North Korean vector space, that is, words that are semantically close to it (or, more figuratively, words that might be used by a North Korean speaker to describe the meaning of "socialism" in the South). The North Korean word vector (red star) is also represented along with its nearest neighbors.

"absolutism," or "Soviet." The proximity to the term "multiparty system" (*tadangje*) in the North Korean vector space does not mean that the South Koreans associate the term with political pluralism. Much to the contrary, it further emphasizes the negative connotation of "socialism" in the South, since in North Korea, a one-party state, multiparty systems are a sign of class division (and therefore a "multiparty system" would be semantically close to "dictatorship" and "absolutism"). Conversely, in Figure 4, we see that South Korean embeddings unsurprisingly associate "socialism" with the left, and other ideologies such as "republicanism,"[41] "communism," or "conservatism." But the meaning of the North Korean term "socialism" is close to concepts such as "equality," "order," "economy," "globalization," and even "capitalism" (presumably in the general sense of an economic system).

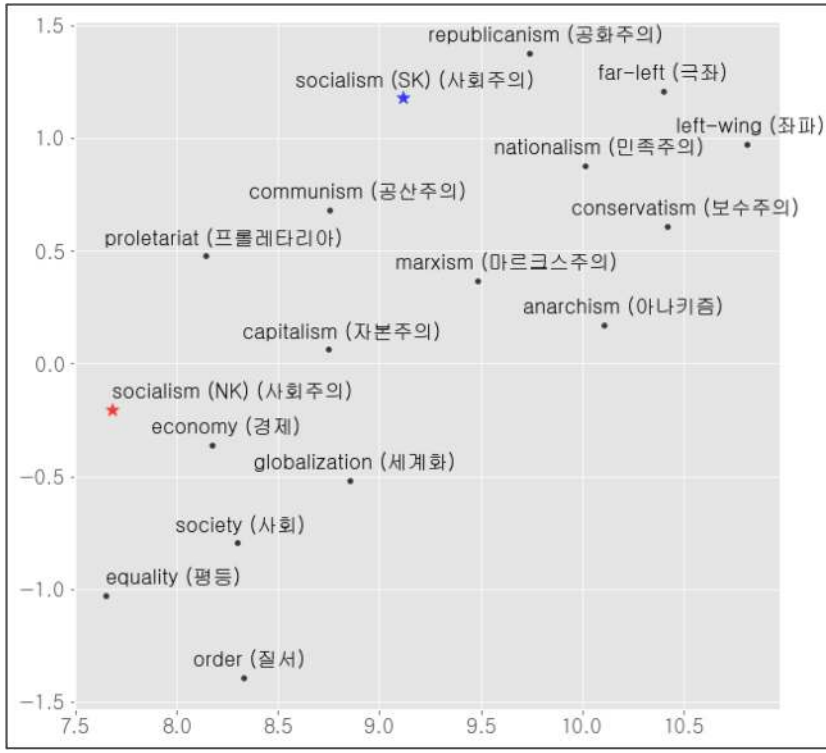

Fig. 4. Projection of the North Korean word vector for "socialism" (red star) in a South Korean vector space. The word is surrounded by its nearest neighbors in the South Korean vector space, that is, words that are semantically close to it (or, more figuratively, words that might be used by a South Korean speaker to describe the meaning of "socialism" in the North). The South Korean word vector (blue star) is also represented along with its nearest neighbors.

As one may object that the differences are marked because the term chosen is a polarizing one, it is worth pointing out that the Jaccard similarity score associated with the word puts it in the upper quartile of the lexicon, indicating that it is one of the terms whose meaning is most similar across the Northern and Southern embeddings, and, by proxy, in both languages. Using the same method or word vector arithmetic's ability to capture more complex semantic relations such as analogies,[42] it is easy to highlight how differences in social practices, ideologies or, definitions of gender can result in semantic divergences.[43]

This semantic drift between North and South affects embeddings but also any downstream NLP application that would rely on them. To

Table 1. Comparison of the Performance of Two Models Using the Same Architecture but Different Embeddings on a Classification Task with North Korean Textual Data

Embeddings Used	Precision	Recall	F1-Score
North Korean	0.86	0.86	0.86
South Korean	0.77	0.76	0.77

illustrate this, I create two similar barebone classifiers with an initial embedding layer, a few convolution layers and a softmax activation function. I compile a dataset with approximately 2000 sentences selected in equal proportion from three different North Korean sources.[44] Each sentence is associated with one of three categories depending on its source. I then train two separate models to identify the category of a given sentence, using each separate set of embeddings to encode the input sentences. Unsurprisingly, the model using North Korean embeddings performs significantly better (Table 1).

Building North Korean Language-Specific Models

The observations above show that while it is possible to use South Korean NLP resources on North Korean texts, they also highlight the significant benefits of using North Korean language-specific resources. Procuring such resources, however, is far from trivial. The lack of available training data—both annotated and unannotated—makes this a particularly challenging task. In the examples previously given, I created North Korean-specific embeddings to demonstrate a point, but the techniques used were far from the current state-of-the-art. The training data used was a corpus of 4.4 million sentences comprising 91 million tokens, which is, to my knowledge, the largest of such corpora available outside of the DPRK.[45] Such a corpus is enough to train static embeddings, but contextual embeddings which are used by more advanced models and perform better, especially on complex tasks,[46] require several orders of magnitudes more data (from 3.3 billion token for an LLM like BERT, 175 billion for GPT-3, and over 1 trillion for the most recent LLMs like Chinchilla or PaLM 2).[47]

For such situations where some, but still little, training data is available in a variation of a language and large amounts of data are available for another, transfer learning (training a machine learning model on one task

or domain and then applying the acquired knowledge to solve another new task) can prove to be a useful approach. Indeed, transfer learning has been successfully applied to other sets of language variations with asymmetrical resources. For instance, contextual embeddings trained on a large amount of data from a well-resourced language can be effectively fine-tuned with only a minimal amount of data from a resource-poor variation.[48] This strategy has been leveraged to develop taggers and parsers for resource-poor languages like African American Vernacular based on English data or the minority language Rusyn using a compilation of resources from several related Slavic languages with better resources.[49]

In the case of Korean, a base model trained on South Korean data would still learn linguistic aspects of the language that are shared by both Southern and Northern variations such as syntactic features, dependency relations, and common semantic information.[50] This South Korean base model can then be trained again, or "fine-tuned" on a smaller dataset of North Korean data and leverage its knowledge of South Korean to capture the specificities of the Northern variation. The resulting model would therefore perform better than a direct application of the base model to North Korean data or a model trained solely on a small amount of North Korean data. Table 2 offers an illustrative benchmark of the performance of four different BERT-based classifiers: one trained solely on South Korean data (**SK base**), one trained solely on North Korean data (**NK base**), one trained on a mix of South and North Korean data (**SK and NK base**), and one trained on South Korean data before being fine-tuned on North Korean data (**SK base with NK fine-tuning**)[51] on the same task of North Korean text classification.

The results clearly show the problem that arises from the scarcity of data characteristic of low-resource languages. The model trained from scratch on North Korean data performs worse than chance, due to the small size of the training data set. The model trained solely on South Korean data is markedly better but is itself outperformed by the model

Table 2. Comparison of the Performance of the Same Language Model (BERT) with Four Different Training Strategies

Training Strategy	Precision	Recall	F1-Score
SK base	0.76	0.72	0.74
NK base	0.16	0.40	0.23
SK and NK base	0.82	0.81	0.82
SK base with NK fine-tuning	*0.88*	*0.87*	*0.88*

trained on a mix of South and North Korean data and the fine-tuned model. The fine-tuned model offers the best results overall.

The benchmark above uses unannotated data for fine-tuning, but the transfer learning technique used to fine-tune the last model can also be used with small, annotated datasets for tasks such as POS tagging, dependency parsing or, machine translation.[52] In the case of machine translation specifically, in 2022, Kim et al. developed a neural machine translation (NMT) model for North Korean to English and Japanese by using transfer learning.[53] They first trained an NMT model on a large dataset of South Korean/English and South Korean/Japanese sentence pairs and fine-tuned it on small (1000 sentences) sets of North Korean/English and North Korean/Japanese sentence pairs. Unsurprisingly, the resulting model performs better than the base South Korean model on North Korean data than a model trained only on North Korean data.

Towards Unified Korean Models

The benchmarks for Kim et al.'s NMT paper also include a model trained on a combination of North and South Korean data. That is, rather than a first training pass on South Korean data followed by fine-tuning on North Korean data, the model was trained directly on a mixed corpus like the **SK and NK base** model above. The model does not perform better than the South Korean model when tested against unseen South Korean data and also underperforms the fine-tuned model on tests against North Korean test data. However, it still performs relatively well overall (whereas the South Korean model performs poorly on North Korean data and the fine-tuned model fares worse on South Korean data).

While fine-tuning can give us better domain-specific models, models that perform better on North Korean data only, training on mixed corpora can give us unified, variation-agnostic models capable of handling both North and South Korean data[54] with performance levels close to those of specific models. Developing such models, however, entails procuring similarly large amounts of data for both languages. Indeed, the benchmarks show that there is quite a difference in how much the mixed-corpora model underperforms the specific model for each language variation. On South Korean data, the degradation is only a few tenths of a percentage point compared to a South Korean model, but on North Korean data the degradation is on the order of several percentage points compared to a fine-tuned model. This differential in the amount of degradation, in turn,

can be explained by the unbalanced amount of training data for each language variation. With small amounts of North Korean data, one may still train a model for Korean, yet that model would be too unequal to be considered unified.

It would seem then, that we are back to the original problem of scant North Korean NLP resources. But what if North Korean's status as a low-resource language had more to do with the definitions of low-resource languages than with the actual amount of existing NLP resources? There are no commonly agreed-upon criteria and no official thresholds for what constitutes a low-resource language.[55] The designation is left at the discretion of the researchers. This pragmatic approach is not necessarily without merit: researchers typically know if their working language is a low-resource one because they can easily assess the gap between it and the state of the start in other well-resourced languages like English. Attempts to more formally define or evaluate languages' resources[56] look at a fairly consistent set of criteria such as the amount of unannotated data available (usually equal to online data), number of annotated datasets, amount of existing NLP systems, and amount of NLP research for the language.

Yet these approaches are ill-suited to the evaluation of a language like North Korean. While a researcher outside North Korea might find the number of available resources extremely limited, this may not be the experience of a North Korean researcher. On the contrary, while NLP papers published in the DPRK may emphasize the need to further "develop" corpus resources, few consider their language to be under-served.[57] This difference cannot simply be dismissed as the product of national pride or ignorance on the part of North Korean scholars. Scholars based outside of the DPRK have limited access to, and often sometimes even knowledge of the existence of, unannotated data since most of the DPRK's digitized textual content is kept on a local network inaccessible from overseas. North Korean NLP research papers and monographs are likewise also almost entirely unavailable online. While they can partially be accessed via South Korean institutions, datasets and NLP systems remain entirely unavailable outside of the DPRK.

That resources are not readily available, however, does not mean that they do not exist, nor that they aren't actively used by North Korean researchers. The DPRK has a long tradition of NLP scholarship, with research on machine translation beginning in the 1950s under the influence of Soviet research,[58] and the first efforts to develop large corpora for computational linguistics starting in the late 1980s.[59] While the country may not have the resources of online user-created content, most of its

intellectual production (novels, magazines, newspapers, etc.) since the 1950s has been digitized and made available on the country's intranet. This large amount of data is used by North Korean researchers who may create thematic or historical subsets depending on their use case.[60] What little North Korean data is available in the South usually stems directly from the North where the data was originally digitized.[61]

In addition to this raw, unannotated data, a look at published NLP research from DPRK also reveals the existence of a number of annotated corpora such as part-of-speech (POS) tagged corpora with over 1 million tagged sentences, several dependency corpora, including one with 40,000 annotated sentences, multiple bilingual corpora for machine translation, including a bilingual Chinese-North Korean corpus with 45,000 sentence pairs.[62] The corpora have been used to develop a number of NLP systems such as tokenizers, named entity recognition taggers, POS taggers, dependency parsers, automatic translators, and information retrieval systems.[63] The systems may not always be state-of-the-art and many use heuristic rules rather than more performant machine learning-based approaches, but, in combination with a large and diverse base of corpora they constitute a solid base and attest to the commitment of North Korean scholars and research institutions to the development of NLP.

More than low resources, the issue with North Korean NLP is one of availability of resources, and communication and collaboration between DPRK-based and non-DPRK-based scholars. From my interactions with a few North Korean NLP scholars prior to the onset of the pandemic, there is marked interest in collaboration and data-sharing from the DPRK. Whether sanctions, public health restrictions, and political and logistic hurdles will allow for the development of collaborative North-South corpora and NLP systems remains to be seen.

Notes

1. Chaewŏn Yu, *Kugŏ chŏngbohwa wa kukka kyŏngjaengnyŏk* [National Competitiveness and the Digitization of the Korean language] (Seoul: Ministry of Culture and Education, November 1997), 8.

2. Ibid., 12–3.

3. Natural Language Processing (NLP) refers to a discipline concerned with the analysis, understanding, and generation of human language through computers. As a practice, NLP encompasses a wide range of methods, from simple heuristic methods for segmenting sentences into words and formulas to compute the differences between words

and strings of characters to more advanced algorithms, often using artificial intelligence, to analyze the grammatical structure of a sentence or generate meaningful and syntactically correct text.

4. While in a limited sense, the expression simply refers to the amount of information available online for speakers of a given language, it can also extend to the availability of basic technologies such as keyboards or character encoding and to more complex resources like corpora and pre-trained NLP models. Laura Lu, "Digital Divide: Does the Internet Speak Your Language?," in *Proceedings of ED-MEDIA 2010—World Conference on Educational Multimedia, Hypermedia & Telecommunications*, ed. J. Herrington and C. Montgomerie (Toronto, Canada: Association for the Advancement of Computing in Education, 2010). Holly Young, "The Digital Language Divide," *British Academy* (2013), accessed April 20, 2022, http://labs.theguardian.com/digital-language-divide/. Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui, *Cross-Lingual Word Embeddings* (Toronto: Morgan & Claypool Publishers, 2019).

5. Observatory of Languages and Cultures in the Internet, "Indicators of Languages in the Internet" (2022), accessed April 2, 2022, https://funredes.org/lc2022/V3.2.htm.

6. These resources can include digital dictionaries, large corpora (both annotated and unannotated) for the training of artificial intelligence based models, software for semantic and syntactic analysis, academic research, etc.

7. Pratik Joshi et al., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), accessed September 25, 2020, https://arxiv.org/abs/2004.09095.

8. Hŭnggyu Kim, Kang Bŏmmo, and Hong Chŏngha, "21-segi sejong kyehoek hyŏndae kugŏ kich'o malmungch'i: sŏngkwa wa chŏnmang" [The Essential Modern Korean Language Corpora Sejong 21st century: Results and Prospects], in *Han'guk chŏngbo kwahakhoe ŏnŏ konghak yŏn'guhoe* [Proceedings of the Conference on Korean Information Sciences and Language Engineering] (2007), 311–6.

9. For a recent English-language overview of available datasets for the Korean language see: Ik Cho Won, Sangwhan Moon, and Youngsook Song, "Open Korean Corpora: A Practical Report," *Proceedings of Second Workshop for NLP Open Source Software* (2020), accessed April 28, 2022, https://aclanthology.org/2020.nlposs-1.12.pdf.

10. Language models are a class of artificial intelligence based models that learn to predict the word most likely to follow a sequence of words or appear within a certain context (i.e., to complete a sentence or to fill in the blanks in a sentence with missing pieces). The models are able to this because they learn the likelihood that a certain word will appear in a certain context from looking at a large number of sentences (training corpora). *Large* language models refer to a specific category of language models that use complex architectures and very large amount of training data to increase their predictive power.

11. Chin-Wu Kim, "Korean as Pluricentric Language," in *Pluricentric Languages: Differing Norms in Different Nations*, ed. Michael Clyne (Berlin/Boston: De Gruyter, Inc, 1991), 239–60.

12. Mitsui Takashi, *Singminji chosŏn ŭi ŏnŏ chibae kujo* [The Dominance of Language in Colonial Korea], trans. Im Kyŏnghwa and Ko Yŏngjin (Seoul: Somyŏng Ch'ulp'an, 2013).

13. Aina Garí Soler and Marianna Apidianaki, "Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses," *Transactions of the Association for Computational Linguistics* (2021): 825–44, https://doi.org/10.1162/tacl_a_00400.

14. Minch'ae Kim, "Nam-pukhan ŏnŏ ijirhwa yoin ŭrosŏ ŭi ŏnŏ chŏngch'aek" [Language Policy as an Influencing Factor of North-South Differences in the Korean Language], *Sahoe ŏnŏhak* [Sociolinguistics] 28, no. 1 (2020): 29–53.

15. Yunam Chŏng, "Nam-pukhan inyŏmjŏk ŏhwi wa ijirhwa chŏngdo" [A Study on the Degree of Differentiation Between North and South Korea Language and Its Ideological Vocabulary], *Journal of Korean Culture* 44 (2019): 37–76.

16. Ch'unggu Kwak, "Nam-pukhan ŏnŏ ijirhwa wa kŭ e kwallyŏn toen myŏt munje" [A Few Issues Related to the Differentiation of North and South Korean Languages], *Sae kugŏ saenghwal* 11, no. 1 (2001): 5–27.

17. Chinu Kim, "Han'guk mal kwa chosŏnmal punhwa ŭi paegyŏng silche mit yoin" [North and South Korean Languages: The Background Causes and Nature of the Differentiation], *Mal* 15 (1991): 35–47.

18. The academic study of North-South linguistic differences started in the 70s, peaked in the early years of the Sunshine policy, and continues today with mostly surveys of lexical differences. Yŏnsuk Hong, *Nam-pukhan ŏnŏ kaenyŏm ŭi ijirhwa yŏn'gu* [Research on the Differentiation of the Concept of Language in North and South Korea] (Seoul: Kukt'o t'ongirwŏn, 1977). Chin-wu Kim, "Linguistics and Language Policies in North Korea," *Korean Studies* 2 (1978): 159–75. Chaeho Chŏn, "Nam-pukhan ŏhwi ŭi hyŏngt'aeronjŏk pigyo punsŏk" [Comparative Morphological Analysis of North and South Korean Lexicon], *Pukhan* 166 (1985): 72–9. Ch'igŭn Ha, *Nam-pukhan munpŏp pigyo yŏn'gu* [Comparative Grammar of South and North Korean] (Seoul: Han'guk munhwasa, 1999). Chaesu Cho, "Nam-pukhan p'yojun mal ŭi ch'ai wa kongdong p'yojun mal kakkugi" [Differences in the Standard Languages of North and South Korea and Preparing a Common Standard Language], *Kyoyuk Han'gŭl* 13 (2000): 55–89. Muno Kim, *Nambuk kyogwasŏ haksul yongŏ pigyo yŏn'gu* [Comparative Study of Linguo in North and South Textbooks] (Seoul: Kungnip kugŏwŏn, 2007).

19. Yonggi Choi, "Nambuk ŭi ŏnŏ ch'ai wa tongjilsŏng hoebok pangan" [Linguistic Differences Between North and South and a Proposal to Restore Homogeneity], *Kukhak yŏn'gu* 10 (2007): 199–228.

20. Yunam Chŏng, "Nam-pukhan inyŏmjŏk ŏhwi," 38. Jaehoon Yeon, "On the Linguistic Divergence between North and South Korea," *Pigyo han'gukhak* 11, no. 1 (2003): 101–21.

21. *Chosŏnmal tae sajŏn* [Great Dictionary of Korean Language] (Pyongyang: Sahoe kwahak ch'ulp'ansa, 2006). Yongho Mun, *Chosŏnŏ pindosu sajŏn* [Korean Dictionary of Word Frequencies] (Pyongyang: Kwahak paekkwa sajŏn chonghap ch'ulp'ansa, 1993).

22. Hansem Kim, *Hyŏndae kugŏ sayong pindo chosa* [Survey of Usage Frequency in Contemporary Korean] (Seoul: National Institute of Korean Language, 2005), https://archive.ph/plzL1.

23. Marcos Zampieri, Preslav Nakov, and Yves Scherrer, "Natural Language Processing for Similar Languages, Varieties, and Dialects: A Survey," *Natural Language Engineering* 26 (2020): 595–612. Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. "Neural Machine Translation into Language Varieties," in *Proceedings of the Third Conference on Machine Translation: Research Papers* (Brussels: ACL, 2018), 156–164.

24. A morphosyntactic tagger or part-of-speech tagger (PoS tagger) assigns each word in a sentence their respective part of speech (noun, verb, article, adverb, etc.). A syntactic parser analyses the syntactic structure of a sentence and the relations between its constituents (A modifies B which is the subject of C).

25. Hyŏnsuk Shin, "Pukhan ŏnŏ silche punsŏk" [Analysis of the State of the North Korean Language], in *Pukhan ŭi mal kwa kŭl* [North Koreans Speech and Text], ed. Ko Yŏnggŭn (Seoul: Ŭryu munhwasa, 1989), 298–322.

26. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval* (Cambridge, Cambridge University Press: 2008), 19.

27. For an early attempt at building a morphological analyzer matching North Korean grammar, cf. Woon-ho Choi and Chŏng Hoi-Sun, "Pukhan munhwaŏ hyŏngt'aeso punsŏkki ŭi ŏjŏl kujo" [The Word Structure of the North Korean Morphological Analyzer], *Han'guk chŏngbo kwahakhoe ŏnŏ konghak yŏn'guhoe haksul palp'yo nonmunjip* (1998): 49–55.

28. WordPiece was originally developed specifically to handle the issue of tokenizing East Asian text such as Korean. Mike Schuster and Kaisuje Nakajima, "Japanese and Korean Voice Search," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing* (2012): 5149–52, https://doi.org/10.1109/ICASSP.2012.6289079. Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (Minneapolis: Association for Computational Linguistics, 2019), 4171–86.

29. These policies, however, were prescriptive and not all of their recommendations were widely followed. The word *ŏrŭm posungi*, introduced in the 1980s as a native Korean translation of the English "ice cream" disappeared from dictionaries shortly after as common usage favored the loanwords *esŭk'imo* and *aisŭk'ŭrim*.

30. Figures based on the *Great Dictionary of Korean Language* for North Korean and the NIKL's statistics on two of its dictionaries. "Sajŏn t'onggye" [Dictionary Statistics], NIKL, https://archive.ph/wip/fhexa (accessed May 1, 2022) and https://archive.ph/ipdzf (accessed May 1, 2022). For an overview of the treatment of loanwords in the North and

South and implications for the compilation of corpora, cf. Yunam Chŏng, "Chayŏn ŏnŏ ch'ŏri rŭl wihan nambuk oeraeŏ koch'al" [A Study on the Loanword of North-South Korean], *Han'guk sajŏnhak* 37 (2021): 421–30.

31. This may, however, be due to the fact that the North Korean frequency statistics were derived from a corpus of novels and newspapers containing less colloquial vocabulary than the South Korean corpus.

32. Daniel Jurafsky and James H. Martin, *Speech and Language Processing* (2021): 106, accessed May 20, 2022, https://archive.ph/wip/3uSNk.

33. Jurafsky and Martin, *Speech and Language Processing*, 252.

34. Both sets were generated using gensim (https://radimrehurek.com/gensim/). The South Korean embeddings are available at https://github.com/Kyubyong/wordvectors and the North Korean ones at https://github.com/digitalprk/north_korean_embeddings. The North Korean embeddings have about 65,000 words and the South Korean embeddings 30,000, for a total common vocabulary of 20,000. The words in the embeddings are the most frequent ones in the training corpus with a threshold to exclude low frequency words. All statistics are derived from the intersection of the two embeddings (i.e., the 20,000 words).

35. Peter D. Turney and Patrick Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research* 37 (2010): 141–88.

36. Hila Gonen et al., "Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (New York: Association for Computational Linguistics, 2020), 538–55. Because vector coordinates can change as a result of even minor alterations in the dataset or training parameters, the order of synonyms can change. This limitation can be mitigated by averaging over multiple bootstrap samples for more robust results. However as the method used here is largely independent from individual proximity measures (since we take all *n* first synonyms), such additional precaution is unnecessary. Cf. Maria Antoniak and David Mimno, "Evaluating the Stability of Embedding-based Word Similarities," *Transactions of the Association for Computational Linguistics* 6 (2018): 107–19, https://aclanthology.org/Q18-1008.pdf.

37. Angie Boggust, Brandon Carter, and Arvind Satyanarayan, "Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples," in *27th International Conference on Intelligent User Interfaces* (New York: Association for Computing Machinery, 2022), 746–66.

38. The datasets were obtained from the HistWords project at Stanford: https://nlp.stanford.edu/projects/histwords/.

39. The first attempts used a few anchor points from a bilingual dictionary to learn the mapping, while more recent methods work in an unsupervised manner. Tomas Mikolov, V. Le Quoc, and Ilya Sutskever, "Exploiting Similarities among Languages for Machine Translation," *ArXiv* (2013), https://arxiv.org/abs/1309.4168. Edouard Grave, Armand

Joulin, and Quentin Berthet, "Unsupervised Alignment of Embeddings with Wasserstein Procrustes," *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics* (Okinawa: PMLR, 2019), http://proceedings.mlr.press/v89/grave19a/grave19a.pdf.

40. I largely follow the methodology outlined in the seminal HistWords paper, including for the visualization where a centroid of the k-nearest neighbors is used to initialize the embedding of the target word. I use UMAP instead of t-SNE for dimension reduction. William L. Hamilton, Jure Leskovec, and Dan Jurafsky, "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change," *arXiv* (2016), https://arxiv.org/abs/1605.09096.

41. The proximity to "republicanism" (*konghwajuŭi*) can be explained by the fact that the term is generally used in opposition to liberalism and individualism and in relation to concepts like communitarianism and totalitarianism. Chongsŏp Shin, "2015 kaejŏng kyoyuk kwajŏng e nat'anan konghwajuŭi kaenyŏm honjae wa haegyŏl panghyang mosaek" [Finding the Solution Direction about Confusion of the Concept of Republicanism in the 2015 Revised Curriculum], *Todŏk yulli kwa kyoyuk* 63 (2019): 27–56.

42. Tomas Mikolov et al., "Distributed Representations of Words and Phrases and their Compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems* 2 (Red Hook, NY: Curran Associates Inc., 2013), 3111–9.

43. For some more examples, cf. "North and South Korea Through Word Embeddings," *Digital NK*, accessed May 21, 2022, https://archive.ph/wip/lbBgG.

44. The sources are the newspaper *Minju Chosŏn*, the Trade section of the website *Nae Nara*, and the news of the Kim Il Sung University's website. The sources were selected to roughly form separate thematic categories (general news items, trade information, scientific news) that could be used for a classification task. The corpus was tokenized and lemmatized prior to the training.

45. As a comparison, the part of the Sejong corpus dedicated to the North Korean language and other diasporic languages contains 11 million tokens, and the North Korean digital data available at the NIKL is estimated to comprise 48 million tokens. Kangch'un So, "Nambuk ŏnŏ charyo kuch'uk kwa chŏngbi pangan" [Compiling and Improving North-South Language Materials], in *Kwangbok 70-chunyŏn kinyŏm kyŏre mal t'onghap ŭl wihan kukche haksul hoeŭi* [Conference for the Unification of the National Language], ed. Lee Taesŏng (Seoul: NIKL, 2015), 126–76.

46. Simran Arora et al., "Contextual Embeddings: When Are They Worth It?," *Association for Computational Linguistics* (2020), https://aclanthology.org/2020.acl-main.236/.

47. BERT, or Bidirectional Encoder Representations from Transformers, is a language model based on the transformers architecture. Upon its public release in late 2018, the model achieved state-of-the-art results in a number of NLP tasks and went on to become an industry standard. More recent models, like PaLM 2 or Chinchilla, rely on even larger amounts of data. For ChatGPT/GPT-4, OpenAI has not released technical details about the training data, but the models are likewise likely to have been trained on several

trillions of tokens. Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre, "Training Compute-Optimal Large Language Models", *Deepmind* (2022), https://arxiv.org/pdf/2203.15556.pdf.

48. "Fine tuning" refers to the practice of using an existing language model trained on a large amount of data for a generic task and retraining it on a smaller dataset to improve its performance on a more specific task. The model leverages the knowledge learned during the training process on the large dataset when training on the smaller specific dataset. It is therefore able to perform significantly better than if it had only been trained on the smaller, specific dataset. For instance a model trained on a large amount of South Korean data can be retrained on a smaller corpus of North Korean data. Tal Schuster et al., "Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing," *Proceedings of NAACL-HLT* (2019), 1599–613, http://aclanthology.lst.uni-saarland.de/N19-1162.pdf.

49. Yves Scherrer and Achim Rabus, "Neural Morphosyntactic Tagging for Rusyn," *Natural Language Engineering* 25, no. 5 (2019): 633–50. https://doi.org/10.1017/S1351324919000287. Anna Jørgensen, Dirk Hovy, and Anders Søgaard, "Learning a POS Tagger for AAVE-like Language," NAACL (2016).

50. This could be the case for a specialized model or a generalist large language model such as BERT. Ganesh Jawahar, Benoît Sagot, and Djamé Seddah, "What Does BERT Learn about the Structure of Language?," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy: ACL, 2019), 3651–7.

51. The South Korean BERT model, KoBERT (https://github.com/SKTBrain/KoBERT/blob/0e4cae19f883fc15f22f260349da9eef27f002f2/README.md), was trained on a corpus compiled from Wikipedia and news articles made up of 25 million sentences comprising 324 million tokens, which is several times larger than the North Korean corpus used for the fine-tuned North Korean and trained from scratch versions of the model (4.4 million sentences, 91 million tokens). The mixed corpus used the North Korean corpus and data from Wikipedia and Namu wiki. All models used a similar tokenization technique and were trained for 40 epochs using a 1e–6 learning rate. Reported score is the highest obtained on the validation set during the training process.

52. For an overview cf. Marcos Zampieri, Preslav Nakov, and Yves Scherrer, "Natural Language Processing for Similar Languages."

53. Hwichan Kim et al., "Learning How to Translate North Korean through South Korean," *arXiv* (2022), accessed May 23, 2022, https://arxiv.org/abs/2201.11258.

54. As well as other regional variants as long as that data is included in the training dataset in sufficient amount.

55. Mika Hämäläinen, "Endangered Languages are not Low-Resourced!," *arXiv* (2021), accessed May 28, 2022, https://arxiv.org/abs/2103.09567.

56. Anil Kumar Singh, "Natural Language Processing for Less Privileged Languages: Where do we Come From? Where are we Going?," *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages* (2008). Christopher Cieri et al., "Selection Criteria for Low Resource Language Programs," *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (2016): 4543–9. Sabrina Mielke, "Language Diversity in ACL 2004–2016," September 22, 2016, accessed June 1, 2022, https://archive.ph/wip/BmxG9. Joshi et al., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," *ACL Conference 2020* (2020), accessed May 25, 2022, https://arxiv.org/abs/2004.09095. Barry Haddow et al., "Survey of Low-Resource Machine Translation," *arXiv* (2022), accessed June 1, 2022, https://arxiv.org/abs/2109.00486.

57. Kwanghŭi Pae, "Chŏngbo chawŏn kwa k'op'ŏsŭ" [Digital Resources and Corpora], *Kim ilsŏng chonghap taehak hakpo – ŏmunhak* 2 (2017): 53–7.

58. G.V. Chelenggebich'i, L.N. Kkot'ŭlepŭ, and S.N. Rajumobŭsŭggi "Chŏnja kyesangi bessŭm e wihan chadong pŏnyŏk ŭi sirhŏm" [Experiment on Automatic Translation with the BESM Electronic Computer], *Suhak kwa Mulli* 4 (1957): 55–9.

59. Yŏngho Mun, *Chosŏnŏ pindosu sajŏn* (1993).

60. *Kwangmyŏng paekkwa sajŏn* [Kwangmyŏng encyclopedia] (Pyongyang: paekkwa sajŏn ch'ulp'ansa, 2009), 262–3.

61. Kangch'un So, "Nambuk ŏnŏ charyo kuch'uk kwa chŏngbi pangan" (2015).

62. Kim Ch'ŏlho and Chu Kyŏngyŏp, "K'op'ŏsŭ ŭi k'ŭgi wa hwangnyŏllong kumun punsŏk" [Corpus Size and Probabilistic Parsing], *Chŏn'gi* 2 (2009): 24–5. Sangnam Ko, "Kil i chŏngbo rŭl riyonghan tu ŏnŏ k'op'ŏsŭ punhal pangpŏp taehan yŏn'gu" [Sentence Segmentation Based on Length of the Source and Target Language Sentence in the Bilingual Corpus], *Chŏngbo kwahak kwa kisul* 3 (2016): 23–4. Rimyang Chŏn, "Kyuch'ik e kich'o han yŏngjo kigye pŏnyŏk esŏ pyŏngnyŏl k'op'ŏsŭ rŭl iyong handae yŏgae maesŏng haeso ŭi han-kaji pangpŏp" [Disambiguation Method Using Parallel Corpora for English-Korean Machine Translation], *Chŏngbo kwahak kwa kisul* 5 (2019): 39–40. Namhyŏk Ri and Kim Kŭmchu, "K'op'ŏsŭ robut'ŏ kyŏkhŭreim sajŏn ŭi chadong kuch'uk e taehan yŏn'gu" [Study on Automatic Construction of Case Frame Dictionary from Corpus], *K'omp'yut'ŏ wa p'ŭrogŭram kisul* 5 (2015), 2–3.

63. Tongun Chi, "Chosŏnŏ k'op'ŏsŭ kuch'uk esŏ nasŏnŭn myŏt kaji munje" [Issues in the Compilation of Korean Corpora], *Kim ilsŏng chonghap taehak hakpo – chŏngbo kwahak* 4 (2014): 89–90. Namhyŏk Ri and Cho Sŏngyŏng, "Munsŏ k'op'ŏsŭ robut'ŏ chadong saengsŏng toen chilmun munjang e ŭihan kŏmsaek ch'egye chŏnghwaksŏng kaesŏn ŭi han-kaji pangpŏp" [A Method for Improving the Precision of Retrieval System using Question Automatically Generated from Document Corpus], *Kim ilsŏng chonghap taehak hakpo – chayŏn kwahak* 7 (2014): 35–8. Chŏngho Ch'oi, O Mihyang, and Sŏk Jun, "Pyŏnhyŏng kyuch'ik ŭl iyong han chosŏnŏ p'umsa p'yogi k'op'ŏsŭ ŭi sujŏng e taehan

yŏn'gu" [Correcting Korean Part-of-Speech Tagging Corpus Using Transformation Rule], *Kim ilsŏng chonghap taehak hakpo – chayŏn kwahak* 5 (2006): 35–8. Hyŏkch'ŏl Ri and Ri Kwangsik, "Tanŏ punhal k'op'ŏsŭ e kich'o han pokhabŏ punhal pangpŏp" [Tokenization Method for Compound Words using a Token Dictionary], *Chŏngbo Kwahak* 2 (2013): 33–4. Myŏngch'ŏl Park, "Yŏngjo kigye pŏnyŏk p'ŭro kŭram ryong namsan e toip han ryŏnŏ k'op'ŏsŭ ŭi ŏnŏhakchŏk t'ŭksŏng kwa kŭ kuch'uk" [Construction and Linguistic Specificities of a Parallel Corpus used for the Ryongnamsam Automatic English-Korean Translator], *Kim Il Sung University*, September 18, 2017, accessed June 1, 2022, https://archive.ph/wip/4Ny21.

References Cited

Antoniak, Maria, and David Mimno. "Evaluating the Stability of Embedding-based Word Similarities." *Transactions of the Association for Computational Linguistics* 6 (2018): 107–19. https://aclanthology.org/Q18-1008.pdf.

Arora, Simran, Avner May, Jian Zhang, and Christopher Re. "Contextual Embeddings: When Are They Worth It?" *Association for Computational Linguistics* (2020). https://aclanthology.org/2020.acl-main.236/.

Boggust, Angie, Brandon Carter, and Arvind Satyanarayan. "Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples." In *27th International Conference on Intelligent User Interfaces*, 746–66. New York: Association for Computing Machinery, 2022.

Chelenggebich'i, G.V., L.N. Kkot'ŭlepŭ, and S.N. Rajumobŭsŭggi. "Chŏnja kyesangi bessŭm e wihan chadong pŏnyŏk ŭi sirhŏm" [Experiment on Automatic Translation with the BESM Electronic Computer]. *Suhak kwa Mulli* 4 (1957): 55–9.

Chi, Tongun. "Chosŏnŏ k'op'ŏsŭ kuch'uk esŏ nasŏnŭn myŏt kaji munje" [Issues in the Compilation of Korean Corpora]. *Kim ilsŏng chonghap taehak hakpo – chŏngbo kwahak* 4 (2014): 89–90.

Cho, Chaesu. "Nam-pukhan p'yojun mal ŭi ch'ai wa kongdong p'yojun mal kakkugi" [Differences in the Standard Languages of North and South Korea and Preparing a Common Standard Language]. *Kyoyuk Han'gŭl* 13 (2000): 55–89.

Ch'oi, Chŏngho, O Mihyang, and Sŏk Jun. "Pyŏnhyŏng kyuch'ik ŭl iyong han chosŏnŏ p'umsa p'yogi k'op'ŏsŭ ŭi sujŏng e taehan yŏn'gu" [Correcting Korean Part-of-Speech Tagging Corpus Using Transformation Rule]. *Kim ilsŏng chonghap taehak hakpo – chayŏn kwahak* 5 (2006): 35–8.

Choi, Woon-ho, and Chŏng Hoi-Sun. "Pukhan munhwaŏ hyŏngt'aeso punsŏkki ŭi ŏjŏl kujo" [The Word Structure of the North Korean Morphological Analyzer]. *Han'guk chŏngbo kwahakhoe ŏnŏ konghak yŏn'guhoe haksul palp'yo nonmunjip* (1998): 49–55.

Choi, Yonggi. "Nambuk ŭi ŏnŏ ch'ai wa tongjilsŏng hoebok pangan" [Linguistic Differences Between North and South and a Proposal to Restore Homogeneity]. *Kukhak yŏn'gu* 10 (2007): 199–228.

Chŏn, Chaeho. "Nam-pukhan ŏhwi ŭi hyŏngt'aeronjŏk pigyo punsŏk" [Comparative Morphological Analysis of North and South Korean Lexicon]. *Pukhan* 166 (1985): 72–9.

Chŏn, Rimyang. "Kyuch'ik e kich'o han yŏngjo kigye pŏnyŏk esŏ pyŏngnyŏl k'op'ŏsŭ rŭl iyong handae yŏgae maesŏng haeso ŭi han-kaji pangpŏp" [Disambiguation Method Using Parallel Corpora for English-Korean Machine Translation]. *Chŏngbo kwahak kwa kisul* 5 (2019): 39–40.

Chŏng, Yunam. "Chayŏn ŏnŏ ch'ŏri rŭl wihan nambuk oeraeŏ koch'al" [A Study on the Loanword of North-South Korean]. *Han'guk sajŏnhak* 37 (2021): 421–30.

Chŏng, Yunam. "Nam-pukhan inyŏmjŏk ŏhwi wa ijirhwa chŏngdo" [A Study on the Degree of Differentiation Between North and South Korea Language and Its Ideological Vocabulary]. *Journal of Korean Culture* 44 (2019): 37–76.

Chosŏnmal tae sajŏn [Great Dictionary of Korean Language]. Pyongyang: Sahoe kwahak ch'ulp'ansa, 2006.

Cieri, Christopher, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. "Selection Criteria for Low Resource Language Programs." *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (2016): 4543–9.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–86. Minneapolis: Association for Computational Linguistics, 2019.

Gonen, Hila, Ganesh Jawahar, Djame Seddah, and Yoav Goldberg. "Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpra." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 538–55. New York: Association for Computational Linguistics, 2020.

Grave, Edouard Armand Joulin, and Quentin Berthet. "Unsupervised Alignment of Embeddings with Wasserstein Procrustes." In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*. Okinawa: PMLR, 2019. https://archive.ph/wip/OQikO.

Ha, Ch'igŭn. *Nam-pukhan munpŏp pigyo yŏn'gu* [Comparative Grammar of South and North Korean]. Seoul: Han'guk munhwasa, 1999.

Haddow, Barry, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. "Survey of Low-Resource Machine Translation." *arXiv* (2022). https://arxiv.org/abs/2109.00486.

Hämäläinen, Mika. "Endangered Languages are not Low-Resourced!" *arXiv* (2021). https://arxiv.org/abs/2103.09567.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." *ACL* (2016). https://arxiv.org/abs/1605.09096.

Hong, Yŏnsuk. *Nam-pukhan ŏnŏ kaenyŏm ŭi ijirhwa yŏn'gu* [Research on the Differentiation of the Concept of Language in North and South Korea]. Seoul: Kukt'o t'ongirwŏn, 1977.

Jawahar, Ganesh Benoît Sagot, and Djamé Seddah. "What Does BERT Learn about the Structure of Language?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–7. Florence, Italy: ACL, 2019.

Jørgensen, Anna, Dirk Hovy, and Anders Søgaard. "Learning a POS Tagger for AAVE-Like Language." *NAACL* (2016).

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020). https://arxiv.org/abs/2004.09095.

Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing* (2021): 106. Accessed May 20, 2022. https://archive.ph/wip/3uSNk.

Kim, Chinu. "Han'guk mal kwa chosŏnmal punhwa ŭi paegyŏng silche mit yoin" [North and South Korean Languages: The Background Causes and Nature of the Differentiation]. *Mal* 15 (1991): 35–47.

Kim, Chin-Wu. "Korean as Pluricentric language." In *Pluricentric Languages: Differing Norms in Different Nations*, edited by Michael Clyne. Berlin/Boston: De Gruyter, Inc, 1991.

Kim, Chin-wu. "Linguistics and Language Policies in North Korea." *Korean Studies* 2 (1978): 159–75.

Kim, Ch'ŏlho, and Chu Kyŏngyŏp. "K'op'ŏsŭ ŭi k'ŭgi wa hwangnyŏllong kumun punsŏk" [Corpus Size and Probabilistic Parsing]. *Chŏn'gi* 2 (2009): 24–5.

Kim, Hansem. *Hyŏndae kugŏ sayong pindo chosa* [Survey of Usage Frequency in Contemporary Korean]. Seoul: National Institute of Korean Language, 2005. https://archive.ph/plzL1.

Kim, Hŭnggyu, Kang Bŏmmo, and Hong Chŏngha. "21-segi sejong kyehoek hyŏndae kugŏ kich'o malmungch'i: sŏngkwa wa chŏnmang" [The Essential Modern Korean Language Corpora Sejong 21st Century: Results and Prospects]. In *Han'guk chŏngbo kwahakhoe ŏnŏ konghak yŏn'guhoe* [Proceedings of the Conference on Korean Information Sciences and Language Engineering]. 2007.

Kim, Hwichan Sangwhan Moon, Naoaki Okazaki, and Mamoru Komachi. "Learning How to Translate North Korean through South Korean." *arXiv* (2022). https://arxiv.org/abs/2201.11258.

Kim, Minch'ae. "Nam-pukhan ŏnŏ ijirhwa yoin ŭrosŏ ŭi ŏnŏ chŏngch'aek" [Language Policy as an Influencing Factor of North-South Differences in the Korean Language]. *Sahoe ŏnŏhak* [Sociolinguistics] 28, no. 1 (2020): 29–53.

Kim, Muno. *Nambuk kyogwasŏ haksul yongŏ pigyo yŏn'gu* [Comparative Study of Linguo in North and South Textbooks]. Seoul: Kungnip kugŏwŏn, 2007.

Ko, Sangnam. "Kil i chŏngbo rŭl riyonghan tu ŏnŏ k'op'ŏsŭ punhal pangpŏp taehan yŏn'gu" [Sentence Segmentation Based on Length of the Source and Target Language Sentence in the Bilingual Corpus]. *Chŏngbo kwahak kwa kisul* 3 (2016): 23–4.

Kwak, Ch'unggu. "Nam-pukhan ŏnŏ ijirhwa wa kŭ e kwallyŏn toen myŏt munje" [A Few Issues Related to the Differentiation of North and South Korean Languages]. *Sae kugŏ saenghwal* 11, no. 1 (2001): 5–27.

Kwangmyŏng paekkwa sajŏn [Kwangmyŏng Encyclopedia]. Pyongyang: paekkwa sajŏn ch'ulp'ansa, 2009.

Lakew, Surafel Melaku Aliia Erofeeva, and Marcello Federico. "Neural Machine Translation into Language Varieties." In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 156–64. Brussels: ACL, 2018.

Lu, Laura. "Digital Divide: Does the Internet Speak Your Language?" In *Proceedings of ED-MEDIA 2010 – World Conference on Educational Multimedia, Hypermedia & Telecommunications*, edited by J. Herrington and C. Montgomerie. Toronto, Canada: Association for the Advancement of Computing in Education, 2010.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

Mielke, Sabrina. "Language Diversity in ACL 2004–2016." Published September 22, 2016. Accessed June 1, 2022. https://archive.ph/wip/BmxG9.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. "Exploiting Similarities among Languages for Machine Translation." *ArXiv* (2013).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and Their Compositionality." In *Proceedings of the 26th International Conference on Neural Information Processing Systems* 2, 3111–9. Red Hook, NY: Curran Associates Inc., 2013.

Mun, Yongho. *Chosŏnŏ pindosu sajŏn* [Korean Dictionary of Word Frequencies]. Pyongyang: Kwahak paekkwa sajŏn chonghap ch'ulp'ansa, 1993.

NIKL. "Sajŏn t'onggye" [Dictionary Statistics]. *NIKL*. https://archive.ph/wip/fhexa (accessed May 1, 2022) and https://archive.ph/ipdzf (accessed May 1, 2022).

Observatory of Languages and Cultures in the Internet. "Indicators of Languages in the Internet." Last updated in March 2022. Accessed April 2, 2022. https://funredes.org/lc2022/V3.2.htm.

Pae, Kwanghŭi. "Chŏngbo chawŏn kwa k'op'ŏsŭ" [Digital Resources and Corpora]. *Kim ilsŏng chonghap taehak hakpo – ŏmunhak* 2 (2017): 53–7.

Park, Myŏngch'ŏl. "Yŏngjo kigye pŏnyŏk p'ŭro kŭram ryong namsan e toip han ryŏnŏ k'op'ŏsŭ ŭi ŏnŏhakchŏk t'ŭksŏng kwa kŭ kuch'uk" [Construction and Linguistic Specificities of a Parallel Corpus Used for the Ryongnamsam Automatic English-Korean Translator]. *Kim Il Sung University*, September 18, 2017. Accessed June 1, 2022. https://archive.ph/wip/4Ny21.

Ri, Hyŏkch'ŏl, and Ri Kwangsik. "Tanŏ punhal k'op'ŏsŭ e kich'o han pokhabŏ punhal pangpŏp" [Tokenization Method for Compound Words Using a Token Dictionary]. *Chŏngbo Kwahak* 2 (2013): 33–4.

Ri, Namhyŏk, and Cho Sŏngyŏng. "Munsŏ k'op'ŏsŭ robut'ŏ chadong saengsŏng toen chilmun munjang e ŭihan kŏmsaek ch'egye chŏnghwaksŏng kaesŏn ŭi han-kaji pangpŏp" [A Method for Improving the Precision of Retrieval System using Question Automatically Generated from Document Corpus]. *Kim ilsŏng chonghap taehak hakpo – chayŏn kwahak* 7 (2014): 35–8.

Ri, Namhyŏk, and Kim Kŭmchu. "K'op'ŏsŭ robut'ŏ kyŏkhŭreim sajŏn ŭi chadong kuch'uk e taehan yŏn'gu" [Study on Automatic Construction of Case Frame Dictionary from Corpus]. *K'omp'yut'ŏ wa p'ŭrogŭram kisul* 5 (2015): 2–3.

Scherrer, Yves, and Achim Rabus. "Neural Morphosyntactic Tagging for Rusyn." *Natural Language Engineering* 25, no. 5 (2019): 633–50. https://doi.org/10.1017/S1351324919000287.

Schuster, Mike, and Kaisuje Nakajima. "Japanese and Korean Voice Search." *2012 IEEE International Conference on Acoustics, Speech and Signal Processing* (2012): 5149–52. https://doi.org/10.1109/ICASSP.2012.6289079.

Schuster, Tal, Ori Ram, Regina Barzilay, and Amir Globerson. "Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing." In *Proceedings of NAACL-HLT* (2019): 1599–613. http://aclanthology.lst.uni-saarland.de/N19-1162.pdf.

Shin, Chongsŏp. "2015 kaejŏng kyoyuk kwajŏng e nat'anan konghwajuŭi kaenyŏm honjae wa haegyŏl panghyang mosaek" [Finding the Solution Direction about Confusion of the Concept of Republicanism in the 2015 Revised Curriculum]. *Todŏk yulli kwa kyoyuk* 63 (2019): 27–56.

Shin, Hyŏnsuk. "Pukhan ŏnŏ silche punsŏk" [Analysis of the State of the North Korean Language]. In *Pukhan ŭi mal kwa kŭl* [North Koreans Speech and Text], edited by Ko Yŏnggŭn, 298–322. Seoul: Ŭryu munhwasa, 1989.

Singh, Anil Kumar. "Natural Language Processing for Less Privileged Languages: Where do we Come From? Where are we Going?" *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages* (2008).

So, Kangch'un. "Nambuk ŏnŏ charyo kuch'uk kwa chŏngbi pangan" [Compiling and Improving North-South Language Materials]. In *Kwangbok 70-chunyŏn kinyŏm kyŏre mal t'onghap ŭl wihan kukche haksul hoeŭi* [Conference for the Unification of the National Language], edited by Lee Taesŏng, 126–76. Seoul: NIKL, 2015.

Søgaard, Anders Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. *Cross-Lingual Word Embeddings*. Toronto: Morgan & Claypool Publishers, 2019.

Soler, Aina Garí, and Marianna Apidianaki. "Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses." *Transactions of the Association for Computational Linguistics* (2021). https://doi.org/10.1162/tacl_a_00400.

Takashi, Mitsui. *Singminji chosŏn ŭi ŏnŏ chibae kujo* [The Dominance of Language in Colonial Korea], translated by Im Kyŏnghwa and Ko Yŏngjin. Seoul: Somyŏng Ch'ulp'an, 2013.

Turney, Peter D., and Patrick Pantel. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37 (2010): 141–88.

Won, Ik Cho, Sangwhan Moon, and Youngsook Song. "Open Korean Corpora: A Practical Report." *Proceedings of Second Workshop for NLP Open Source Software* (2020). https://aclanthology.org/2020.nlposs-1.12.pdf.

Yoon, Jaehoon. "On the Linguistic Divergence between North and South Korea." *Pigyo han'gukhak* 11, no. 1 (2003): 101–21.

Young, Holly. "The Digital Language Divide." *British Academy* (2013). Accessed April 20, 2022. http://labs.theguardian.com/digital-language-divide/.

Yu, Chaewŏn. *Kugŏ chŏngbohwa wa kukka kyŏngjaengnyŏk* [National Competitiveness and the Digitization of the Korean language]. Seoul: Ministry of Culture and Education, November 1997.

Zampieri, Marcos, Preslav Nakov, and Yves Scherrer. "Natural Language Processing for Similar Languages, Varieties, and Dialects: A Survey." *Natural Language Engineering* 26, no. 6 (2020): 595–612.