# Revisiting Korean Corpus Studies through Technological Advances

**Won Ik Cho**
Seoul National University*
Seoul, Korea
tsatsuki@snu.ac.kr

**Sangwhan Moon**
Tokyo Institute of Technology
Tokyo, Japan
sangwhan@iki.fi

**Youngsook Song**
Sionic AI Inc.
Seoul, Korea
song@sionic.ai

## Abstract

The Korean language has been largely studied recently owing to the active development of Korean-specific language models and the disclosure of natural language processing (NLP) benchmarks that followed. Such alignment of technological advances and the proposal of challenging datasets is common in the progress of artificial intelligence research; each affects one another, driving new approaches from the other side, driving the trend. In this paper, we remind how recent achievements in Korean NLP relate to corpus studies so far. Along with a comprehensive diachronic overview, we see how downstream tasks correspond with the advent of modern NLP techniques, at the same time discussing the change of trend in volume, task type, and topic.

## 1 Introduction

The importance of data and scale in modern natural language processing (NLP) has become an essential issue. Advances in computing devices and engineering methodologies have guaranteed sufficient scalability in NLP systems, especially those that use machine learning (ML)-based methods.

However, the complexity and difficulty of tasks for model evaluation are also aligned with such developments. Accordingly, the struggle regarding data construction to reproduce performance under experimental conditions or in the real world has become more visible than before. The direction of corpus building has also been expanded from unannotated text (Francis and Kucera, 1967) and treebanks (Marcus et al., 1993) – that are mainly used in structural or syntactic analysis of human language, to comparatively purpose-specified or semantic-level tasks such as sentiment analysis (Maas et al., 2011) and question answering (Rajpurkar et al., 2016). It also includes transitioning from the rule and dictionary-based early-stage NLP to those with more learning-based methodologies (Manning and Schutze, 1999).

This alignment of technological advance and evaluation schemes has been mainly observed in English, the lingua franca – for instance the advent of comprehensive benchmarks such as general language understanding evaluation (GLUE) (Wang et al., 2018) and architectures like BERT (Devlin et al., 2019) and ULMFiT (Howard and Ruder, 2018), and occurred more accessibly in other Indo-European languages that are relatively easy to extend evolving methodologies. From that point of view, even though Korean is a solitary language concerning linguistic typology, modern techniques have been quickly applied to it (Yang, 2021), and corresponding benchmarks were actively suggested as in Park et al. (2021) and Jang et al. (2022).

This trend has been significant in recent years, and we look through the data construction trend and ongoing studies in Korean NLP. In addition, we examine which factors have influenced this trend, discuss what this phenomenon in the Korean language means, and how the case study can be expanded to other languages or domains.

## 2 Korean NLP Studies

Background include surveys of Korean corpora, Korean NLP model development, and all related works by academia, industry, or government.

### 2.1 Korean Corpora

Research on Korean corpora includes survey studies such as Park et al. (2016), Cho et al. (2020), several blog articles[1], and Github reports[2]. These materials each has their own purpose (curation, classification, recommendation, etc.) and allows researchers to recognize and access datasets of diverse topics. Among them, we try to refer to our

* Work done after graduation.

[1] https://littlefoxdiary.tistory.com/42
[2] https://github.com/datanada/Awesome-Korean-NLP

previous work (Cho et al., 2020) covering 62 documented open corpora, and especially more than 50 written corpora among them.

### 2.2 Korean NLP Paradigm

The development of Korean NLP largely follows approaches from classical computational linguistics. An example that a similar approach was borrowed in Korean as it was motivated is Treebank (Choi et al., 1994; Han et al., 2001), which was first established in English (Marcus et al., 1993). Along with this, baseline models were also adopted, and the direction of research has been aligned at the implementation level (Han et al., 2002). However, as in any other multilingual study, Korean is significantly distinguishable from English studies in several areas.

First of all, the spacing and agglutinativeness of scriptio continua are combined, and the tokenization scheme has a great impact on the task. As there are research papers on this topic (Park et al., 2020b) and modeling papers suggest performance changes accordingly (Park et al., 2021), the unit of word, morpheme, and token in Korean is a highly controversial and frequently studied topic. In addition, modern written Korean is recorded with unique symbols (Hangul Jamo) used only in the Koreanic language (Lee, 2009), which do not include symbols that appear comprehensively in worldwide articles such as the Latin alphabet or CJK ideograph (except some use cases in code-switching manner). This adds challenges to applying transfer learning methods when jointly pretrained with or transferred from any neighboring languages, which can be seen as one of the factors that motivated the independent development of Korean.

### 2.3 Past and Ongoing Projects

Despite various limitations, multiple institutes have invested their resources in research on Korean NLP. In terms of corpus construction, Sejong Corpus (Kim, 2006) and ModuCorpus[3] led by the government (especially National Institute of Korean Language[4], NIKL (2020)) and ExoBrain[5] project led by Electronics and Telecommunications Research Institute[6] (ETRI) are typical examples, and corpora of overseas institutes such as Linguistic Data Consortium[7] (LDC) were also actively utilized (Cieri et al., 2022). Recently, attempts to activate the AI ecosystem through dataset construction and distribution have increased, and NLP datasets of various topics have been proposed by National Information Society Agency[8] (NIA) and others. In addition, recently, corpora in small or medium-scale volume are being disclosed with their building schemes transparently published, possibly as an academic contribution at the individual or organization level (sometimes led by industry) (Cho et al., 2020).

Technological advance, represented by ML models, is largely aligned with the advent of aforementioned datasets (though not necessarily causal). The establishment of pretrained language models (PLMs) such as Word2Vec (Mikolov et al., 2013) or BERT (Devlin et al., 2019) has been quickly applied to Korean[9 10] (Al-Rfou et al., 2013; Lee, 2020; Park, 2020; Kim et al., 2021b), and recent foundation models difficult for individual researchers to handle, such as GPT-3 (Brown et al., 2020), are provided as Korean-targeted, like HyperCLOVA (Kim et al., 2021a) or Polyglot-Ko (Ko et al., 2023), so that researchers can access them in the form of APIs or as a checkpoint.

## 3 Diachronic Overview

We skim over various Korean works introduced earlier from a diachronic viewpoint. Before and after the appearance of Sejong corpus (Kim, 2006), which was the first large-scale corpus available to the public, visible changes occurred in terms of task type and data volume. These breakthroughs tend to originate in the advance in embedding/encoding methods such as Word2Vec (Mikolov et al., 2013) or Transformer (Vaswani et al., 2017), and changes in the volume of pretrained knowledge such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018) that mostly comes from breakthroughs in training methodology (self-supervised learning, SSL) and scalability (Kaplan et al., 2020).

### 3.1 Early Stages

Before the appearance of Sejong Corpus, a large-scale government-driven digital corpus annotation, Korean corpus construction was mainly driven by researchers who handled computational linguistics and classical NLP pipelines (Tenney et al., 2019).

[3] https://corpus.korean.go.kr/
[4] https://www.korean.go.kr/
[5] http://exobrain.kr/pages/ko/
[6] https://www.etri.re.kr/intro.html
[7] https://www.ldc.upenn.edu/
[8] https://www.nia.or.kr/
[9] https://word2vec.kr/search/
[10] https://github.com/Kyubyong/wordvectors

Most of the corpora for various morphological properties were disclosed at LDC or established by Korea Advanced Institute of Science and Technology (KAIST) (Choi et al., 1994) and these corpora could be purchased through a catalog or provided by submitting an application form.

### 3.2 Statistical Models and Word2Vec

Classical NLP pipeline studies such as tagging, parsing, chunking, and relation linking based on the Sejong Corpus became popular (Kim et al., 2010; Lee and Kim, 2013; Park et al., 2014), and correspondingly, NLP techniques based on statistical models were also widely exploited. Studies including semantic and pragmatic level ones appeared more frequently after the advent of Word2Vec, and this led to further investigation of new datasets and benchmarks such as sentiment analysis, e.g., Naver Sentiment Movie Corpus (NSMC[11]) which has long been a representative Korean sentence classification benchmark.

Word2Vec drove the practical application of distributional semantics to NLP communities, helping conventional ML models reach the desired downstream task performance with less effort. However, it did not necessarily bring significant achievement in performance by parameter training neural network architectures of similar sizes. That is, despite the major change, the insertion of distributional semantic knowledge in word embedding, the need for challenging benchmarks was not observed without fundamental modification in the training scheme and architecture of ML models.

### 3.3 Advent of Transformer and Pretrained Language Models

What has redefined the direction of natural language processing since Word2Vec is undoubtedly the emergence of attention mechanism (Bahdanau et al., 2014), self attention and Transformers (Vaswani et al., 2017), and the subsequent development of various Transformer-based PLMs (Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2019). SSL and scaling laws (either architecture or training data) arose as fundamental keys of LM pretraining and Transformer was a timely architecture. The resulting models acquire (linguistic) knowledge to some extent so that even with a relatively small amount of training data, one may obtain downstream performance comparable to training a vanilla ML model from scratch. As a result, model evaluation regarding linguistics, domain, or commonsense displayed distinguished aspect compared to previous ones. With the advent of large language models (LLMs), some benchmarks only accommodate evaluation as a feature[12].

## 4 Discussion

### 4.1 Trends in Volume: Large-scale, raw text to small, specific, annotated text

The datasets in the early stages of corpus construction were used for the purpose of analyzing the corpus itself or, furthermore, investigating the trend of language use. Brown corpus (Francis and Kucera, 1967) or Corpus of Contemporary American English (COCA) (Davies, 2009) in English is representative, and in Korean, some datasets released for this purpose were in the LDC catalog, but were not fully publicly available. In addition, the datasets released by KAIST or NIKL were the results of annotating raw text to grammatical or functional components or properties; semantic or discourse annotations beyond syntactic properties (e.g. sentence/document level) could be applied in some modifications, but most approaches were not published and done only in-house not to violate the license of the original text.

However, in order to overcome the bias of dataset research, recent corpora tend to choose to annotate a limited amount of raw corpus or to add annotations to already published corpus with open and redistributable license. The topics covered are not being limited to general domain or colloquial text, and increasingly reflecting specified regions (See Section 4.3).

### 4.2 Trends in Task Type: Token-annotated to document-annotated, classification to span/generation

How the annotated corpus is utilized is mostly up to the dataset user or the service provider of the further product.However, in some cases it is necessary to clarify the nature of the benchmark through the intended use, so an evaluation metric is usually presented together with a baseline model, which will inevitably determine the in-out style of the data (Wang et al., 2018; Park et al., 2021).

In the early stages, the task was dominated by token-tagging annotation, which mainly deals with

[11] https://github.com/e9t/nsmc

[12] As in a recent LLM benchmark: https://huggingface.co/spaces/upstage/open-ko-llm-leaderboard

syntactic properties such as part-of-speech tagging (Han and Han, 2001), dependency and constituency parsing (Choi et al., 1994; Han et al., 2001; Park and Kwon, 2008), that sentence-level analysis was not easily observed or dataset undisclosed. NSMC, which covers one of the most popular tasks, sentiment classification, has been the representative, publicly available sentence-level classification dataset. It was created for Korean after the advent of Word2Vec, following Maas et al. (2011).

Despite that sentence-level classification is a fundamental NLP task, not enough datasets that suit as benchmark have been disclosed. 3i4K (Cho and Kim, 2022) for single utterance level speech acts, BEEP! (Moon et al., 2020) for toxic speech, and YNAT, KLUE benchmark's topic classification task (Park et al., 2021), are representative, but in general, sentence classification datasets that are built for a specific purpose are often not disclosed to the public. One assumption is that, though they are the most accessible and useful type of corpus to build in both academia and industry, annotated corpora usually follow the original license of the source corpus, which may not permit redistribution (Moon et al., 2022), or sometimes their disclosure is prohibited for the security or the interest of the organization. Also, unlike the current trend where all the datasets and models used for the experiment should be reported transparently, studies in the past were often not asked to submit the relevant materials. These would have prevented the disclosure of NLP datasets used in studies before the post-Word2Vec era and let researchers rely on few publicly available annotated corpora. Fortunately, institutes like NIA are struggling to enlighten the NLP ecosystem by helping create various topics of classification benchmarks. However, since those dataset are usually not open global and the construction process is not peer-reviewed, benchmarks suitable for the academic purpose is still limited.

Document-level tasks have been less frequently constructed because processing long-length inputs seemed infeasible until the development of ML techniques. In addition, document-level annotation usually requires a larger construction budget compared to sentence-level ones, and the difficulty of building and publishing such data at the individual level contributed to its scarcity.

However, as passage-based inference tasks such as question answering (Yang et al., 2015; Rajpurkar et al., 2016) have gained popularity, more document-level tasks have been published than before (Lim et al., 2019; Kim et al., 2019) as well in Korean (usually driven by industry). In addition, span tagging, a classic method of question answering, is theoretically a token classification, but it can also be seen as an answer generation process. Therefore, the development of generative models (Radford et al., 2018; Brown et al., 2020) has driven the development of decoding strategies, the development of open domain question answering (Karpukhin et al., 2020), and the development of evaluation of generation tasks (Gehrmann et al., 2021), which is still evolving in Korean but being recognized as a new direction for future NLP research, for instance in NIKL competition[13]. Translation and transliteration have been regarded as typical examples of such tasks[14][15] (Park et al., 2016), but datasets for paraphrasing (Yang et al., 2019; Cho et al., 2022; Kim, 2022) or summarization [16], as well as conversation datasets (Lee et al., 2022) have also been created and published to boost the studies on rephrasing and generation.

### 4.3 Trends in Topics: Written or spoken (web) text to texts in various areas

The last significant change in trend is the increase in the diversity of corpus topics. The advent of PLM has brought enhancement of general language understanding the performance of machine learning models, and it accordingly, brought demand for model-based solutions for tasks in domains that were previously considered unfeasible, e.g. law (Hwang et al., 2022), cultural heritage (Kim et al., 2022), non-Seoul Korean dialect (e.g., Jejueo (Park et al., 2020a)).

The advent of new tasks in some sense means that the society and community require a new direction of research that is either necessary or timely, but also implies that previously addressed topics or domains are sufficiently handled by state-of-the-art models. We interpret this phenomenon as having led to the expansion of benchmark construction, as seen in the motivation of the development of KLUE (Park et al., 2021) or KoBEST (Jang et al.,

[13] https://corpus.korean.go.kr/taskOrdtm/useTaskOrdtmList.do

[14] http://semanticweb.kaist.ac.kr/home/index.php/Evaluateset2

[15] http://semanticweb.kaist.ac.kr/home/index.php/Corpus9

[16] https://github.com/machinereading/K2NLG-Dataset

2022) that follows the case of GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019).

## 5 Conclusion

Through this paper, we skimmed discussions that provide a diachronic view of the development of Korean corpora in view of technological development. We addressed that there were significant changes in overall corpus characteristics in terms of corpus volume, annotation type, and topic. Also, we qualitatively checked that such changes are also associated with up-to-date natural language processing in lingua franca such as English and the application of the technologies to Korean.

We have not anticipated that the emergence of LLM would drive a revitalization of the AI ecosystem, and thus boost the importance of challenging and evaluation-oriented benchmarks. Similarly, we cannot ensure the future direction of corpus construction, or even whether the corpus building process itself will be meaningful. However, the high quality corpus has inevitably been accompanied by the development of fine-grained guidelines, and building such a scheme is essential even in the contemporary LLM era where the emphasis is on prompt engineering. In addition, language model safety related tasks such as detection of hate speech or bias (Lee et al., 2023b), acceptability (Lee et al., 2023a), or checking AI reasoning ability (Dziri et al., 2023), tend to attract attention in recent periods.

A core limitation of our study is the lack of quantification of the findings. This includes a comprehensive organization and arrangement of existing works, which were not all covered in this research. In the future, we plan to develop this research to visualize changes in corpus statistics and figure out the trends[17], where such information can play important role in the analytics and assuming the next direction.

## Acknowledgements

We thank all contributors of Korean NLP and language dataset studies for providing the motivation of this research.

[17]To be updated in the following project page: https://github.com/ko-nlp/Open-korean-corpora

## References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Won Ik Cho and Nam Soo Kim. 2022. Text implicates prosodic ambiguity: A corpus for intention identification of the korean spoken language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(1).

Won Ik Cho, Sangwhan Moon, Jongin Kim, Seokmin Kim, and Nam Soo Kim. 2022. StyleKQC: A style-variant paraphrase corpus for Korean questions and commands. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7122–7128, Marseille, France. European Language Resources Association.

Won Ik Cho, Sangwhan Moon, and Youngsook Song. 2020. Open korean corpora: A practical report. *arXiv preprint arXiv:2012.15621*.

Key-Sun Choi, Young S Han, Young G Han, and Oh W Kwon. 1994. KAIST tree bank project for Korean: Present and future development. In *Proceedings of the International Workshop on Sharable Natural Language Resources*, pages 7–14. Citeseer.

Christopher Cieri, Mark Liberman, Sunghye Cho, Stephanie Strassel, James Fiumara, and Jonathan Wright. 2022. Reflections on 30 years of language resource development and sharing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 543–550, Marseille, France. European Language Resources Association.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. 2023. Faith and fate: Limits of transformers on compositionality. *arXiv preprint arXiv:2305.18654*.

W Nelson Francis and Henry Kucera. 1967. Computational analysis of present-day american english. *Providence, RI: Brown University Press. Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, AB (2014). Emotion and language: Valence and arousal affect word recognition. Journal of Experimental Psychology: General*, 143:1065–1081.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Chung-hye Han and Na-Rae Han. 2001. Part of speech tagging guidelines for penn korean treebank.

Chung-hye Han, Na-Rae Han, Eon-Suk Ko, and Martha Palmer. 2002. Development and evaluation of a Korean treebank and its application to NLP. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Chung-hye Han, Na-Rae Han, Eon-Suk Ko, Martha Palmer, and Heejong Yi. 2001. Penn Korean Treebank: Development and evaluation. In *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*, pages 69–78.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *arXiv preprint arXiv:2206.05224*.

Myeongjun Jang, Dohyung Kim, Deuk Sin Kwon, and Eric Davis. 2022. Kobest: Korean balanced evaluation of significant tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3697–3708.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. 2021a. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424.

Gyeongmin Kim, Jinsung Kim, Junyoung Son, and Heuiseok Lim. 2022. KoCHET: A Korean cultural heritage corpus for entity-related tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3496–3505, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hansaem Kim. 2006. Korean national corpus in the 21st century Sejong project. In *Proceedings of the 13th NIJL International Symposium*, pages 49–54. National Institute for Japanese Language Tokyo.

Ildoo Kim, Gunsoo Han, Jiyeon Ham, and Woonhyuk Baek. 2021b. Kogpt: Kakaobrain korean(hangul) generative pre-trained transformer. `https://github.com/kakaobrain/kogpt`.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2010. A cross-lingual annotation projection approach for relation detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 564–571, Beijing, China. Coling 2010 Organizing Committee.

Seonghyun Kim. 2022. Smilestyle: Parallel style-variant corpus for korean multi-turn chat text dataset. `https://github.com/smilegate-ai/korean_smile_style_dataset`.

Youngmin Kim, Seungyoung Lim, Hyunjeong Lee, Soyoon Park, and Myungji Kim. 2019. Korquad 2.0: Korean qa dataset for web document machine comprehension. In *Annual Conference on Human and Language Technology*, pages 97–102. Human and Language Technology.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Sungho Park, et al. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *arXiv preprint arXiv:2306.02254*.

Changki Lee and Hyunki Kim. 2013. Automatic korean word spacing using pegasos algorithm. *Information processing & management*, 49(1):370–379.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoungpil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park, and Jung-Woo Ha. 2023a. SQuARe: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6692–6712, Toronto, Canada. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023b. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Junbum Lee. 2020. Kcbert: Korean comments bert. In *Annual Conference on Human and Language Technology*, pages 437–440. Human and Language Technology.

Sang-Oak Lee. 2009. The korean alphabet: an optimal featural system with graphical ingenuity. *Written Language & Literacy*, 12(2):202–212.

Yoon Kyung Lee, Won Ik Cho, Seoyeon Bae, Hyunwoo Choi, Jisang Park, Nam Soo Kim, and Sowon Hahn. 2022. "feels like i've known you forever": empathy and self-awareness in human open-domain dialogs.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD 1.0: Korean QA dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Sangwhan Moon, Won Ik Cho, Hye Joo Han, Naoaki Okazaki, and Nam Soo Kim. 2022. OpenKorPOS: Democratizing Korean tokenization with voting-based open corpus annotation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4975–4983, Marseille, France. European Language Resources Association.

NIKL. 2020. Nikl corpora 2020 (v.1.0).

Cheon-Eum Park, Kyoung-Ho Choi, and Changki Lee. 2014. Korean coreference resolution using the multi-pass sieve. *Journal of KIISE*, 41(11):992–1005.

Jangwon Park. 2020. Koelectra: Pretrained electra model for korean. `https://github.com/monologg/KoELECTRA`.

Jungyeul Park, Jeen-Pyo Hong, and Jeong-Won Cha. 2016. Korean language resources for everyone. In *Proceedings of the 30th Pacific Asia conference on language, information and computation: Oral Papers*, pages 49–58.

Kyubyong Park, Yo Joong Choe, and Jiyeon Ham. 2020a. Jejueo datasets for machine translation and speech synthesis. In *In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

Kyubyong Park, Joohong Lee, Seongbo Jang, and Dawoon Jung. 2020b. An empirical study of tokenization strategies for various Korean NLP tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 133–142, Suzhou, China. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Yong-uk Park and Hyuk-chul Kwon. 2008. Korean syntactic analysis using dependency rules and segmentation. In *2008 International Conference on Advanced Language Processing and Web Information Technology*, pages 59–63. IEEE.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Kichang Yang. 2021. Transformer-based korean pretrained language models: A survey on three years of progress. *arXiv preprint arXiv:2112.03014*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692.