CantoNLU: A benchmark for Cantonese natural language understanding

Junghyun Min[†] **York Hay Ng**[‡] **Sophia Chan**[§]
Helena Shunhua Zhao[‡] **En-Shiun Annie Lee**[‡&]

[†]Georgetown University [‡]University of Toronto
[&]Ontario Tech University [§]Independent Researcher
jm3743@georgetown.edu

Abstract

Cantonese, although spoken by millions, remains under-resourced due to policy and diglossia. To address this scarcity of evaluation frameworks for Cantonese, we introduce **CantoNLU**, a benchmark for Cantonese natural language understanding (NLU). This novel benchmark spans seven tasks covering syntax and semantics, including word sense disambiguation, linguistic acceptability judgment, language detection, natural language inference, sentiment analysis, part-of-speech tagging, and dependency parsing. In addition to the benchmark, we provide model baseline performance across a set of models: a Mandarin model without Cantonese training, two Cantonese-adapted models obtained by continual pre-training a Mandarin model on Cantonese text, and a monolingual Cantonese model trained from scratch. Results show that Cantonese-adapted models perform best overall, while monolingual models perform better on syntactic tasks. Mandarin models remain competitive in certain settings, indicating that direct transfer may be sufficient when Cantonese domain data is scarce. We release all datasets, code, and model weights to facilitate future research in Cantonese NLP.

Keywords: Cantonese, natural language understanding, transfer learning, benchmark

1. Introduction

Mandarin Chinese is considered a high-resource language, abundant with pre-trained language model (PLM) support (Devlin et al., 2019; Liu et al., 2020; Conneau et al., 2020), corpora, and evaluation benchmarks (Xu et al., 2020; Xiang et al., 2021), and most recently commercial large language models (Bai et al., 2023; Liu et al., 2024). However, the same cannot be said about other variants of the Sinitic language family, including Cantonese. Mandarin is the official state language and the prestige language in media, business, and academia. Owing to this status, Cantonese, along with other Sinitic languages, remains a primarily vernacular language without written standardization (Snow, 2004; Li, 2006).

Such lack of written standardization, prestige, and official status results in a shortage of language resources in Cantonese. It is frequently described as a low-resource language (e.g. Liu, 2022; Xiang et al., 2022; Jiang et al., 2025b) despite having millions of speakers (Eberhard et al., 2023). As a result, Cantonese language processing systems often rely on datasets, models, and corpora adapted from Mandarin (Xiang et al., 2024), despite a lack of mutual intelligibility between the two languages (Norman, 1988; Tang and van Heuven, 2007).

Cross-lingual transfer from Mandarin for Cantonese language processing has proven effective, with empirical success across a range of tasks, including language modeling and reading comprehension (Jiang et al., 2025b), translation (Liu, 2022; Suen et al., 2024), and speech recognition (Li et al., 2019; Luo et al., 2021). However, despite empirical success, there remains a gap in formal investigations into the best practices for Cantonese natural language understanding (NLU). This is attributable to a lack of a centralized evaluation framework for Cantonese language processing or understanding.

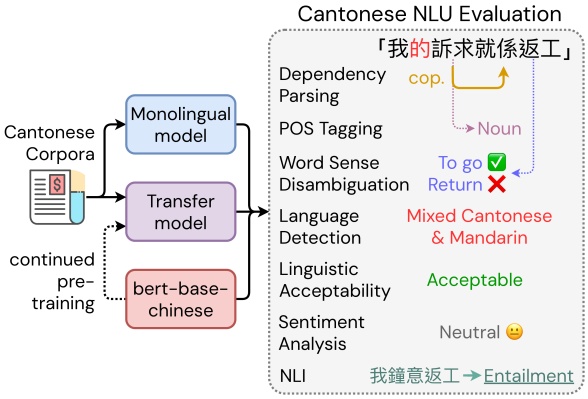

Figure 1: An overview of the tasks in CantoNLU, and our framework for investigating when and how cross-lingual transfer learning from Mandarin is effective for natural language understanding in Cantonese.

In this paper, we introduce **CantoNLU**, a GLUE-like (General Language Understanding Evaluation; Wang et al., 2018) natural language understanding benchmark in Cantonese. **CantoNLU** provides

Task	Requires	Size	Source
Sentence-level tasks			
Acceptability	L, F, M	1.6k	MT error dataset (Liu et al., 2025) adapted
Lang Detection	L, F	47k	Parallel corpus (Dai et al., 2025) aligned and purturbed
NLI	M	570k	Machine-translated English NLI (Chung Shing, 2025)
Sentiment	M	12k	Hong Kong restaurant reviews (Zhang et al., 2011)
Word-level tasks			
WSD	L, M	109	Manual compilation
POS, Dep parsing	F	14k	Universal Dependencies dataset Wong et al. (2017)

Table 1: An overview of the 7 NLU tasks that make up **CantoNLU**, what the task requires, their size and source. L, F, M stand for lexicon, form (syntax), and meaning (semantics), respectively.

an in-depth evaluation of syntax, lexicon and semantic understanding, comprising 7 tasks: word sense disambiguation (WSD), linguistic acceptability judgment (LAJ), language detection (LD), natural language inference (NLI), sentiment analysis (SA), part-of-speech tagging (POS), and dependency parsing (DEPS). In particular, the WSD dataset is entirely novel, providing the first resource for sense-level lexical understanding in Cantonese. LAJ and LD represent novel adaptations of existing datasets, while NLI, SA, POS, and DEPS datasets are direct adoptions of existing datasets (Chung Shing, 2025; Zhang et al., 2011; Wong et al., 2017, respectively). We describe each task and underlying datasets in Section 4, whose overview is outlined in Table 1.

Using CantoNLU, we evaluate three types of models – a Mandarin model[1] without explicit training on Cantonese, Cantonese-adapted transfer models with additional Cantonese training on a Mandarin model, and a monolingual Cantonese model, as outlined in Section 5. In Section 6, we discuss how monolingual Cantonese, Cantonese-adapted, and Mandarin models compare across various aspects in Cantonese NLU.

In spite of limited training data in Cantonese, we demonstrate that our monolingual Cantonese model excels in syntactic tasks such as POS and DEPS. On the other hand, Cantonese-adapted models excel in semantic tasks such as NLI, LD, WSD, and SA. Simultaneously, mandarin models offer a competitive alternative to additional or monolingual training on Cantonese data, excelling in NLI and LAJ.

Based on our results, we recommend monolingual Cantonese models for syntactic tasks, and Cantonese-adapted models for semantic tasks. In domains where Cantonese corpora are scarce, Mandarin models without Cantonese adaptation may be sufficient. In addition to the benchmark and analysis, we publicly release our code and model weights at `github.com/aatlantise/sinitic-nlu`.

[1] `bert-base-chinese`

2. Background

Cantonese, also known as Yue (Eberhard et al., 2023), is the second most widely used Sinitic language after Mandarin, spoken by an estimated 85 million people worldwide. However, it remains primarily a spoken language (Norman, 1988), with most speakers defaulting to written Standard Chinese (Xiang et al., 2024). This diglossic situation and its short history as a written language (Snow, 2004) has contributed to the scarcity of high-quality textual resources for Cantonese NLP, in stark contrast to Mandarin Chinese, a related language rich in resource across corpora, models (Devlin et al., 2019; Liu et al., 2020; Conneau et al., 2020; Bai et al., 2023; Liu et al., 2024), and evaluation resources (e.g. Xu et al., 2020; Xiang et al., 2021). Thus, as highlighted in Section 1, prior work have used varying degrees of transfer from models with Mandarin knowledge to perform downstream tasks in Cantonese (Li et al., 2019; Chung Shing, 2024; Jiang et al., 2025b).

Cross-lingual transfer. Transfer learning is a common strategy in deep learning (DL) to address data sparsity, where features or signal learned from a resource-rich domain is used to augment a low-resource domain, task, or dataset (Bengio, 2012). In the context of NLP, a common application of transfer learning is in cross-lingual transfer for low-resource languages (Ruder et al., 2019; Conneau et al., 2020). Two types of cross-lingual transfer exists: model transfer and data transfer, as described in García-Ferrero et al. (2022). Data transfer involves translating a dataset or corpus from a high-resource language to a lesser-resourced target language, as performed in `yue-all-nli` (Chung Shing, 2025), the MMLU (Hendrycks et al., 2021) portion of the HKCanto-Eval benchmark (Cheng et al., 2025), and Yue benchmarks from Jiang et al. (2025a).

On the other hand, model transfer involves adapting models trained on a high-resource source language to a lesser-resourced target language, such

as Mandarin to Cantonese and Danish to Faroese (Snæbjarnarson et al., 2023). Model transfer relies on the lexical or typological similarity between the two languages, thereby enabling the transfer of linguistic knowledge (Khan et al., 2025). While some implementations of cross-lingual transfer explicitly designate a source language for cross-lingual transfer (Snæbjarnarson et al., 2023; Thangaraj et al., 2024), others rely on multilingual models. Rather than transferring from a specific source language, such models are thought to capture general cross-linguistic patterns that extend beyond typological or lexical similarity, allowing them to perform reasonably well even on unseen or low-resource languages (Conneau et al., 2020; Scivetti et al., 2025). Hybrid approaches combine both paradigms: they begin with a multilingual model, then continue pre-training or fine-tuning on a specific target language before applying the resulting model to downstream tasks (Protasov et al., 2024; Manafi and Krishnaswamy, 2024).

Due to the richness of Mandarin language resources and the strong performance of Mandarin PLMs and LLMs, most cross-lingual transfer to Cantonese use Mandarin as the source language (Li et al., 2019; Liu, 2022; Luo et al., 2021; Suen et al., 2024; Chung Shing, 2024), with exceptions using a multilingual model (e.g. Jiang et al., 2025b).

However, there is emerging work suggesting limitations in models' ability to capture Cantonese idiosyncrasies. Factors previously identified include substantial dissimilarities in lexicon, syntax and writing systems (Suen et al., 2024), along with the prevalence of colloquial phrases and code-switching in more recent Cantonese corpora (Jiang et al., 2025a). These issues hinder the ability of Mandarin-trained models in Cantonese (Xiang et al., 2024) due to over-reliance on Mandarin linguistic knowledge (Cheng et al., 2025).

Cantonese grammar. Cantonese diverges from Mandarin in word order, particle and grammatical word inventory, and morphology, as highlighted in theoretical linguistics work (Matthews, 2006; Yap and Chor, 2011; Matthews and Yip, 2011) as well as prior work in Cantonese-Mandarin machine translation and corpus linguistics (Zhang, 1998; Lam, 2020; Liu et al., 2025). For example, in double object constructions, Mandarin takes the direct-indirect order as seen in (1), while Cantonese takes the indirect-direct order as seen in (2) (Matthews and Yip, 2011).

(1) *gei3 ni3 qian2*
give you money
'(I) give you money'
(Mandarin)

(2) *bei2 cin2 nei5*
give money you
'(I) give you money'
(Cantonese)

Cantonese also features a substantially larger and more diverse inventory of particles and aspect markers than Mandarin, enabling speakers to encode subtle distinctions of tense, stance, and speaker attitude (Yap and Chor, 2011; Matthews and Yip, 2011). Cantonese morphology is more flexible, with more frequent verb serialization (Matthews, 2006) and reduplication (Lee, 2020; Lam, 2020) than in Mandarin. This allows sequences such as (3), sourced from O'Melia (1965), where three verbs combine to describe one scene and (4), where reduplicating the classifier (counting noun) yields an *every* quantifier. Given such unique properties of Cantonese grammar, availability of high-quality Cantonese data is critical to performing well on Cantonese linguistic tasks.

(3) *keoi5 jap6 heoi3 co5*
3SG enter come sit
'He went in and sat down'

(4) *zek3 zek3 gau2*
CL CL dog
'every dog'

3. Related Work

Cantonese language resources. Although widely considered low-resource, recent efforts have begun to address scarcity in Cantonese language resources. Machine-translations of non-Cantonese resources may offer potential utility as data transfer, as provided by Chung Shing (2025) for Cantonese natural language inference (NLI) and Cheng et al. (2025); Jiang et al. (2025a) for LLM knowledge evaluation. Foundational tools such as PyCantonese (Lee et al., 2022) provide `NLTK`-like (Bird and Loper, 2004) essential language processing utilities, while a Cantonese Universal Dependencies dataset (Wong et al., 2017) offers a small yet significant syntactically annotated treebank. Multilingual resources such as Wikipedia (Foundation, 2023), SIB-200 (Adelani et al., 2024), and NLLB (NLLB Team et al., 2024) include Cantonese portions, which may be useful for representation learning or evaluation in topic classification and multilingual translation. Larger-scale corpora such as YueData (Jiang et al., 2025b) have also emerged, though they often contain substantial portions of non-Cantonese text.

Beyond corpora, work has extended to specific applications, including sentiment analysis (Zhang et al., 2011), automatic speech recognition (Li et al.,

2019), machine translation (Liu, 2022; Suen et al., 2024; Dai et al., 2025; Liu et al., 2025). Most recently, Cheng et al. (2025) introduced HKCanto-Eval, a comprehensive benchmark for evaluating linguistic and cultural understanding in Cantonese. Jiang et al. (2025a) presents another benchmark for evaluating LLM reasoning, knowledge, and logic in Cantonese.

On the modeling side, prior work has explored both general-purpose and Cantonese-specific pre-trained models. While commercial multilingual models such as Qwen (Bai et al., 2023) and DeepSeek (Liu et al., 2024) provide Cantonese support, their Cantonese proficiency lags behind that in English or Mandarin (Jiang et al., 2025a). Jiang et al. (2025b) use their `YueData` corpus to train `YueTung-7b`, a continually pre-trained model based on a Qwen-7B model (Bai et al., 2023), which exhibit improved Cantonese performance compared to other open-source and commercial LLMs. On the smaller end in terms of the number of parameters, Chung Shing (2024) represents the only encoder-only Cantonese model to our knowledge. It is continually pre-trained `bert-base-chinese` on Cantonese news articles, social media posts, and web pages on. Implementation details, training recipes, and corpora selection for `bert-base-chinese` and `bert-base-cantonese` are both obscure as neither model provides a description paper or technical report. In particular, Devlin et al. (2019) describes the BERT architecture in general but not the Chinese model.

CantoNLU and other Cantonese benchmarks. Although we have described Cantonese benchmarks HKCanto-Eval (Cheng et al., 2025) and Yue-Benchmark (Jiang et al., 2025a) as related work, we wish to clarify our contributions to Cantonese benchmarking in comparison to these two benchmarks. While HKCanto-Eval includes OpenRice (Zhang et al., 2011) as a SA dataset, which overlaps with our CantoNLU, along with minor datasets in Cantonese phonology and orthography, HKCanto-Eval and Yue-Benchmark primarily target generative LLMs on general world knowledge and reasoning, in a comparable way to MMLU (Hendrycks et al., 2021). We highlight CantoNLU's focus on discriminative NLU tasks, explicitly evaluating a model's ability to understand the Cantonese lexicon (WSD), syntax (POS, DEPS), semantics (NLI, SA), and overall well-formedness with respect to both syntax and semantics (LAJ). The focus is similar to those of GLUE (Wang et al., 2018) and its derivatives–in Korean (Park et al., 2021), Chinese (Xu et al., 2020), Vietnamese (Do et al., 2024), and Indonesian (Wilie et al., 2020).

4. Building CantoNLU

We introduce a benchmark of seven Cantonese NLU tasks, encompassing word sense disambiguation (WSD), linguistic acceptability judgment (LAJ), language detection (LD), natural language inference (NLI), sentiment analysis (SA), part-of-speech tagging (POS), and dependency parsing (DEPS). While all tasks require Cantonese proficiency, tasks such as LD may reward Mandarin knowledge, whereas tasks such as DEPS and WSD may penalize Mandarin knowledge. WSD, LAJ, and LD datasets are novel contributions to Cantonese NLU.

Novel WSD dataset via manual compilation. First, we manually compile Cantonese words with more than one attested meaning to create the first Cantonese word sense disambiguation dataset. We collect the words' multiples senses and two example sentences for each sense, resulting in 41 multi-sense words with a total of 109 senses. For each sense, there are least 2 example sentences containing the word. The dataset does not require fine-tuning for evaluation—model predictions are obtained by masking the target word in each sentence and comparing cosine similarities between the hidden representations at the mask position. For each target word w with two contexts s_i and s_j, we obtain hidden representations for each at the masked position from the model $\mathbf{h}_i$ and $\mathbf{h}_j$.

Using cosine similarity, we define same-sense score s_{same} and different-sense score s_{diff} as :

$$s_{same} = \frac{1}{|P_{\text{same}}|} \sum_{(i,j)\in P_{\text{same}}} \cos(\mathbf{h}_i, \mathbf{h}_j),$$

$$s_{\text{diff}} = \frac{1}{|P_{\text{diff}}|} \sum_{(i,j)\in P_{\text{diff}}} \cos(\mathbf{h}_i, \mathbf{h}_j),$$

where P_{same} and P_{diff} are the sets of sentence pairs containing the same and different senses of w, respectively. A model prediction is considered correct if $s_{\text{same}} > s_{\text{diff}}$.

Novel LAJ dataset adapted from error span dataset. LAJ is a classification task, where the model predicts whether the given sequence is linguistically acceptable. This often aligns with grammatical judgment acceptability, but may also include judgment on semantic plausibility or pragmatic felicity. We compile the first Cantonese LAJ dataset by adapting the Cantonese portion of SiniticMTError (Liu et al., 2025), a dataset of Sinitic translation error span annotations, where each datapoint consists of a well-formed reference sentence `ref`, a machine translated sentence `mt`, and annotations of errors in the `mt` sentence. We consider error-free `ref` as acceptable, `mt` with error annotations not acceptable, to create pairs with one

acceptable and one not acceptable versions of the same sentence. Unlike previous LAJ datasets such as CoLA (Warstadt et al., 2019), which asks for a binary acceptable-not acceptable judgment, we implement a more robust setup of providing two versions of the same sentence and asking for a more acceptable version as is preferred in psycholinguistics and cognitive science (Mahowald et al., 2016; Linzen and Oseki, 2018). The dataset consists of 1.6k pairs.

Novel LD dataset from word-aligning and perturbing parallel corpus. Language Detection (LD) is a three-label classification task that identifies whether a given sentence is written in Cantonese, Mandarin, or mixed. We construct the novel dataset from the parallel translation corpus of Dai et al. (2025), selecting the first 10,000 sentence pairs. To create mixed-language examples, we randomly replace tokens in Cantonese and Mandarin sentences with their counterparts from the other language with a set probability. In a given mixed sentence, 15%, 33%, 50% of the sentence may be from the other language, thus producing up to 6 mixed sentences for each pair. Word-level alignments are obtained using SimAlign (Jalili Sabet et al., 2020), and all text is converted to traditional orthography using HanziConv (Yue and Gallant, 2014) to prevent script-based shallow heuristics. The resulting dataset contains 47,578 sentences, of which 27,578 are mixed. We reserve 5% each for the validation and testing splits. We note that this task requires proficiency in both Mandarin and Cantonese to perform well.

Machine translated English NLI datasets. NLI is a classification task where the model is asked to predict whether the premise entails, or implies the truth of, the hypothesis. The NLI portion of our benchmark is the `yue-nli-all` dataset (Chung Shing, 2025), which is a machine translation of English NLI datasets SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) into Cantonese. The dataset comprises of 557k train examples, 6.6k development examples, and 6.6k test examples. Each example includes a reference premise and two hypotheses, one entailing and one contradicting, from which we create two examples with each label. This results in a balanced, two-label classification dataset.

Sentiment analysis dataset from restaurant reviews. Sentiment analysis is a sentence-level classification task of predicting the sentiment of a given sentence. We use the OpenRice dataset (Zhang et al., 2011) compiled from restaurant reviews in Hong Kong, with the 3-way label space of positive, neutral, and negative. The dataset is balanced, with 10k datapoints in its train split, 1k in its development split, and 1k in its test split.

POS and DEPS from Cantonese UD. Finally, POS and DEPS are both token-level classification tasks. A POS model predicts the part-of-speech tag (e.g. noun, verb, etc.) of a given word, while a DEPS model predicts the dependency head (answering, which word is the syntactic head of this word?) and dependency type (answering, what is the relationship between this word and its head?) of the given word. For POS and DEPS, we use the Cantonese-HK Universal Dependency dataset (Wong et al., 2017), which comprises of 1k sentences and 14k tokens. We split the dataset 9:1 for training and testing, respectively. The dataset contains 15 POS tags and 48 dependency relations, of which 17 are specific to Cantonese. We report both unlabeled (UAS) and labeled attachment scores (LAS) to measure DEPS performance.

5. Experimental Benchmark Setup

Using the compiled Cantonese NLU benchmark, we evaluate three types of models–Cantonese monolingual, Cantonese-adapted from Mandarin, and Mandarin.

Pre-training corpora. For those requiring training or adaptation into Cantonese, we use two corpora: the Cantonese Wikipedia (Foundation, 2023) and a list of 30 million Cantonese sentences compiled by Kwok (2024). The two corpora are the publicly available and open-source subset of YueData (Jiang et al., 2025b). The Wikipedia dump includes empty parentheses from their pre-processing stage that removes text in other languages. For example, (5) becomes (6), whose process often yields parentheses that are either empty or only contain punctuation marks. We remove them. Then, the corpora are divided into excerpts of maximum length 128 for pre-training. The Cantonese Wikipedia contains 137k articles, totaling 40M characters, while Kwok (2024) contains 30M sentences, totaling 660M characters. In total, the pre-training corpora consists of 700M characters.

(5) 影（英：movie/ film），特点是运动／移动的画面（英：motion/ moving picture)[2]

(6) 电影（/ ），特点是运动／移动的画面（）

Cantonese-adapted model. The Cantonese-adapted model represents the most common text processing effort in Cantonese, taking an existing

[2]The illustrated example is from Mandarin, due to LaTeX compilers' difficulty with Cantonese text.

model with Mandarin support and performing continued pre-training on Cantonese before applying the adapted text or speech model to downstream tasks (Luo et al., 2021; Chung Shing, 2024; Jiang et al., 2025b). In our implementation, we take the off-the-shelf `bert-base-chinese` model (Devlin et al., 2019)[3] and continually pre-train on Cantonese text described above. We do not make any changes to the model's tokenizer.

In addition, we offer a comparison to a concurrent BERT-based Cantonese work in `bert-base-cantonese` (Chung Shing, 2024), also a continually pre-trained model, but on a close-source corpus of news articles, social media posts, and web content.

Mandarin model without Cantonese adaptation. Direct transfer from Mandarin without conditioning on Cantonese also represents a small subset of Cantonese NLP efforts (Liu, 2022; Li, 2024), including evaluating non-Cantonese trained multi-lingual LLMs such as Llama models on Cantonese knowledge benchmarks (Cheng et al., 2025). We employ `bert-base-chinese` (Devlin et al., 2019) as a model representing direct transfer from Mandarin, and fine-tune the model on the downstream tasks without additional conditioning on Cantonese text. We do not make any changes to the model's tokenizer.

Monolingual Cantonese model. We train a monolingual Cantonese model from scratch using the BERT architecture (Devlin et al., 2019). While previous work in Cantonese language modeling have incorporated additional datasets (Jiang et al., 2025b), many of them are proprietary, may include non-Cantonese text, or may not be freely used as pointed out by Xiang et al. (2024). Thus, similar to the Cantonese-adapted model, we train a monolingual Cantonese model using the publicly available Cantonese Wikipedia dump (Foundation, 2023) and the `cantonese-sentences` corpus (Kwok, 2024). The amount of data, totaling 700M characters, is an order of magnitude smaller than what was used to train `bert-base-uncased` at 3.3B words (Devlin et al., 2019). We are unable to make an exact comparison to `bert-base-chinese` as training details of the model are not publicly available, although we suspect the Mandarin Wikipedia dump was used for a part of its training, which consists of around 1 million articles (cf. 137k articles for Cantonese) at the time of `bert-base-chinese` training.

For the monolingual model, we train a sentencepiece byte-pair encoding tokenizer on the same data to obtain a Cantonese-only tokenizer. Unlike the tokenizer from `bert-base-chinese` which only includes 8k tokens of character length 1, the Cantonese tokenizer captures subword structure by including around 32k tokens, of which 27k are multi-character tokens. Moreover, a high overlap in lexicons is maintained, with 4.8k characters being represented in both `bert-base-chinese` and our Cantonese tokenizer. While we expected greater coverage of Cantonese-only characters from Hong Kong Supplementary Character Set (HKSCS), this is not the case as `bert-base-chinese`'s tokenizer boasts a greater coverage with 603 characters, while our tokenizer covers only 233 characters in HKSCS. This result reflects the relative scarcity of Cantonese-specific data compared to the abundant Mandarin data available.

Each model is fine-tuned on the training split of the downstream NLU task, then evaluated on the test split of the same task with the exception of LAJ and WSD which use model surprisal without fine-tuning. We report accuracy metrics for NLI, LAJ, WSD; F1 metrics for POS, LD and SA; and UAS and LAS for DEPS. Following NLU convention (Devlin et al., 2019), we do not freeze model weights and allow them to be updated during fine-tuning. We report task-specific hyperparameter choice during fine-tuning in Table 2.

Task	LR	Batch	Epochs
NLI	2e-5	16	3
POS	2e-5	32	3
DEP	3e-5	16	20
LD	2e-6	16	3
SA	2e-5	16	3
LAJ	No fine-tuning performed		
WSD	No fine-tuning performed		

Table 2: Hyperparameters used for fine-tuning across tasks. LR = learning rate. Patience refers to the number of epochs before early stopping.

6. Results and Discussion

Our results across the 7 Cantonese NLU tasks are shown in Table 3, where our Cantonese monolingual model demonstrates the highest performance for POS and DEPS; Cantonese-adapted for NLI, LD, and WSD; Mandarin model without Cantonese adaptation for NLI and LAJ. Our transfer model's performances slightly tail that of `bert-base-cantonese`. We highlight three main findings. First, Cantonese monolingual models excel in syntactic tasks, while Cantonese-adapted Mandarin models are currently the most effective approach for Cantonese semantic tasks. Second,

[3]The citation is attributable to the `bert` language model as a whole; it does not include implementation details on `bert-base-chinese`.

Model	WSD Acc.	LAJ F1	LD F1	NLI Acc.	SA F1	POS F1	DEPS UAS	DEPS LAS	Avg.
No Cantonese adaptation									
`bert-base-chinese`	78.9	**91.7**	76.4	**93.2**	70.2	74.6	29.1	25.9	67.5
Cantonese-adapted									
`bert-base-cantonese`	**92.7**	89.2	**78.7**	**93.2**	**71.9**	72.2	30.2	26.8	**69.4**
Our transfer model	85.3	89.6	78.4	87.5	71.3	74.8	30.0	26.8	68.0
Monolingual Cantonese									
Our monolingual model	70.6	85.7	73.3	82.6	70.1	**78.2**	**32.4**	**27.9**	65.1

Table 3: Performance of Mandarin, Cantonese-adapted, and monolingual models across Cantonese NLU tasks. We offer performances of open-weight but closed-source models `bert-base-chinese` (Devlin et al., 2019) and `bert-base-cantonese` Chung Shing (2024) as comparison.

Mandarin-only models can still perform competitively on some tasks. Finally, despite the relative success of monolingual and transfer models, there remains substantial room for improvement in representing Cantonese lexicon, syntax, and semantics.

Transfer from Mandarin is the most effective for Cantonese NLU. Supporting the empirical success of Mandarin-to-Cantonese transfer seen in contemporary Cantonese NLP, Cantonese-adapted Mandarin models offer the strongest performance with a task-averaged score of 69.4 (`bert-base-cantonese`) and 68.0 (our open-source transfer model). While the monolingual model excels in syntactic tasks POS and DEPS, its average score of 65.1 lags behind those the Cantonese-adapted and Mandarin models. We attribute this primarily to the small size of our Cantonese pre-training corpus (roughly 700M characters). As discussed in Section 5, this is an order of magnitude smaller than the datasets used to train comparable English or Mandarin models, and therefore insufficient for learning robust linguistic representations from scratch.

A well-trained Mandarin model may be sufficient for some Cantonese NLP Interestingly, `bert-base-chinese` achieves comparable, or in some cases, superior performance to its Cantonese-adapted and Cantonese-monolingual counterparts with an average score of 68.1. When scholars discuss mutual intelligibility among Sinitic languages, they typically refer to the spoken form (Tang and van Heuven, 2007; Gooskens and Van Heuven, 2021). However, written Sinitic languages are far more mutually intelligible, given the historical influence of Mandarin as the dominant language of education and literacy across Chinese-speaking regions (Snow, 2004; Xiang et al., 2024). This suggests that effective written Cantonese understanding can emerge from training on Mandarin text without explicit exposure to Cantonese text, as also observed in Jiang et al. (2025b), where LLMs without explicit Cantonese support perform relatively well on Cantonese and Hong Kong-related knowledge benchmarks.

Despite effective transfer, Cantonese representations remain limited. While models achieve respectable performance on NLI, POS tagging, and lexical disambiguation, dependency parsing remains notably weak—likely due to both the small fine-tuning dataset of around 1k sentences and the inherent difficulty of the task. Nonetheless, dependency parsing for other low-resource languages such as Buryat (Badmaeva and Tyers, 2017) and Old English (Levine et al., 2025) has reached higher performance with smaller datasets, suggesting that current Cantonese representations still lack sufficient syntactic and semantic grounding.

Taken together, these results point to the promise of continued pre-training for Cantonese NLP and to the need for richer, higher-quality Cantonese corpora to close the representational gap.

7. Limitations

While we proposes a novel evaluation framework and recommend insights into Cantonese representation learning, several limitations remain. First, the size and coverage of available Cantonese corpora significantly constrain our results. Our monolingual Cantonese model was trained on roughly 700M characters, an order of magnitude smaller than corpora typically used to pre-train large language models in other major languages (Devlin et al., 2019). This data sparsity likely limits the model's ability to capture the full lexical and semantic diversity of written Cantonese, and may explain the comparatively weaker performance of the monolingual model in semantic tasks. Future work would benefit from expanding and diversifying Cantonese text resources, especially in informal and user-generated domains that reflect contemporary usage.

Second, our evaluation benchmark, though designed to cover multiple aspects of Cantonese NLU,

is itself constrained by data availability. Some tasks, such as dependency parsing (DEPS), rely on small fine-tuning datasets (Wong et al., 2017), making the results more susceptible to statistical noise and limiting generalization. In addition, the benchmark focuses on written Cantonese and does not address spoken or colloquial aspects of the language such as code-switching (Yim and Bialystok, 2012) or informal register in-depth; these aspects are integral to Cantonese as it is a primarily spoken language (Snow, 2004).

As a result, while our findings suggest that Mandarin-to-Cantonese transfer is effective for semantic tasks and Mandarin models perform sufficiently well, they should be interpreted as a reflection of current data and resource disparities, rather than representative features of Mandarin and Cantonese.

8. Conclusion

In this paper, we describe a benchmark of Cantonese NLU tasks, and evaluate a monolingual Cantonese model, two transfer models from Mandarin, and a Mandarin model on said Cantonese benchmark to investigate contexts where each type of model is most effective. Our results indicate that both Cantonese monolingual models and Cantonese-adapted models with cross-lingual transfer from Mandarin both have merit for Cantonese NLP in today's landscape of language resources. In addition, direct transfer from a Mandarin model without Cantonese representation learning may suffice for some tasks. Our findings also suggest that the existing open-source Cantonese corpora are insufficient to train a reliable representation of Cantonese lexicon, syntax, and semantics.

Our benchmark and analyses provide the first systematic investigation into and evidence of whether and how transfer from Mandarin is effective for performing Cantonese linguistic tasks. By establishing a framework and pipeline for training monolingual or transfer models and evaluating them, we hope to catalyze broader progress in the space of Cantonese NLP.

Transfer may light the way, but data will pave the road.

Responsible Research Statement

While we do not collect human annotations for our work, we acknowledge that we make use of datasets that were annotated by humans or otherwise adapted from a human source.

We make use of a variety of Cantonese NLP and language resources. In addition to our discussion of them in Sections 3, 4, and 5, we acknowledge their use and organize them below.

- `bert-base-chinese` `huggingface.co/google-bert/bert-base-chinese` (Devlin et al., 2019)
- `Cantonese Wikipedia` `huggingface.co/datasets/wikimedia/wikipedia/viewer/20231101.zh-yue` (Foundation, 2023)
- `Cantonese-HK UD Corpus` `github.com/UniversalDependencies/UD_Cantonese-HK` (Wong et al., 2017)
- `Cantonese Sentences` `huggingface.co/datasets/raptorkwok/cantonese_sentences` (Kwok, 2024)
- `Yue-All-NLI` `huggingface.co/datasets/hon9kon9ize/yue-all-nli` (Chung Shing, 2025)
- SiniticMTError (Liu et al., 2025)
- OpenRice `/www.openrice.com/en/hongkong`, collected by Zhang et al. (2011)
- Parallel corpus from Dai et al. (2025)
- `cantonese-chinese-parallel-corpus` `huggingface.co/datasets/HKAllen/cantonese-chinese-parallel-corpus`
- `Hong Kong Cantonese Corpus` `github.com/fcbond/hkcancor` (Luke and Wong, 2015)
- `Words.HK` `words.hk` (Lau et al., 2022)
- `CC-Canto` `cantonese.org`
- `CEDict` `cedict.org`
- `KaiFang CiDian` `kaifangcidian.com`
- `Tatoeba` `tatoeba.org`
- Hong Kong Supplementary Character Set (HKLSCS) `www.ccli.gov.hk/en/hkscs/what_is_hkscs.html`

Finally, we disclose that we use ChatGPT[4] and PyCharm's AI[5] as a writing and coding assistant during the project's implementation and the paper's writeup.

[4] `https://chat.openai.com`

[5] `https://www.jetbrains.com/pycharm/features/ai/`

Bibliographical References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Elena Badmaeva and Francis M. Tyers. 2017. Dependency treebank for buryat. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*, pages 1–12.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, volume 27 of *Proceedings of Machine Learning Research*, pages 17–36, Bellevue, Washington, USA. PMLR.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tsz Chung Cheng, Chung Shing Cheng, Chaakming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.

Cheng Chung Shing. 2024. hon9kon9ize/bert-base-cantonese.

Cheng Chung Shing. 2025. yue-all-nli: Cantonese triplet nli dataset for sentence embedding.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yuqian Dai, Chun Fai Chan, Ying Ki Wong, and Tsz Ho Pun. 2025. Next-level Cantonese-to-Mandarin translation: Fine-tuning and post-processing with LLMs. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 427–436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Phong Nguyen-Thuan Do, Son Quoc Tran, Phu Gia Hoang, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. VLUE: A new benchmark and multi-task knowledge transfer learning for Vietnamese natural language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 211–222, Mexico City, Mexico. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas.

Wikimedia Foundation. 2023. Wikimedia downloads.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Charlotte Gooskens and Vincent J Van Heuven. 2021. Mutual intelligibility. *Similar languages, varieties, and dialects: A computational perspective*, pages 51–95.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob

Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1627–1643, Online. Association for Computational Linguistics.

Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2025a. How well do LLMs handle Cantonese? benchmarking Cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4464–4505, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiyue Jiang, Alfred Kar Yin Truong, Yanyu Chen, Qinghang Bao, Sheng Wang, Pengan Chen, Jiuming Wang, Lingpeng Kong, Yu Li, and Chuan Wu. 2025b. Developing and utilizing a large-scale cantonese dataset for multitasking in large language models. *arXiv preprint arXiv:2503.03702*.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Raptor Kwok. 2024. Cantonese sentences.

Charles Lam. 2020. Forms and meanings of lexical reduplications in Cantonese: a corpus study. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 562–567, Hanoi, Vietnam. Association for Computational Linguistics.

Chaak-ming Lau, Grace Wing-yan Chan, Raymond Ka-wai Tse, and Lilian Suet-ying Chan. 2022. Words. hk: a comprehensive cantonese dictionary dataset with definitions, translations and transliterated examples. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 53–62.

Jackson Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. PyCantonese: Cantonese linguistics and NLP in python. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6607–6611, Marseille, France. European Language Resources Association.

Peppina Po-lun Lee. 2020. On the semantics of classifier reduplication in cantonese. *Journal of Linguistics*, 56(4):701–743.

Lauren Levine, Junghyun Min, and Amir Zeldes. 2025. Building UD cairo for Old English in the classroom. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 97–104, Ljubljana, Slovenia. Association for Computational Linguistics.

Bryan Li, Xinyue Wang, and Homayoon Beigi. 2019. Cantonese automatic speech recognition using transfer learning from mandarin. *arXiv preprint arXiv:1911.09271*.

David C. S. Li. 2006. Chinese as a lingua franca in greater china. *Annual Review of Applied Linguistics*, 26:149–176.

Qing Li. 2024. Fine-tuning guangdong cantonese based on wav2vec 2.0 xlrs model pretrained on mandarin chinese to improve asr performance. Master's thesis, University of Groningen, Campus Fryslân, July. Supervised by Dr. Shekhar Nayak; second reader: Associate Professor Matt Coler.

Tal Linzen and Yohei Oseki. 2018. The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics*, 3(1).

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Hannah Liu, Junghyun Min, Ethan Yue Heng Cheung, Shou-Yi Hung, Syed Mekael Wasti, Runtong Liang, Shiyao Qian, Shizhao Zheng, Elsie Chan, Ka Ieng Charlotte Lo, et al. 2025. Siniticmterror: A machine translation dataset with error annotations for sinitic languages. *arXiv preprint arXiv:2509.20557*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual

denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kang-Kwong Luke and May LY Wong. 2015. The hong kong cantonese corpus: design and uses. *Journal of Chinese Linguistics*, pages 309–330.

Jian Luo, Jianzong Wang, Ning Cheng, Edward Xiao, Jing Xiao, Georg Kucsko, Patrick O'Neill, Jagadeesh Balam, Slyne Deng, Adriana Flores, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, and Jason Li. 2021. Cross-language transfer learning and domain adaptation for end-to-end automatic speech recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Kyle Mahowald, Peter Graff, Jeremy Hartman, and Edward Gibson. 2016. Snap judgments: A small n acceptability paradigm (snap) for linguistic acceptability judgments. *Language*, 92(3):619–635.

Shadi Manafi and Nikhil Krishnaswamy. 2024. Cross-lingual transfer robustness to lower-resource languages on adversarial datasets. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4174–4184, Torino, Italia. ELRA and ICCL.

Stephen Matthews. 2006. On serial verb constructions in cantonese. *Serial verb constructions: A cross-linguistic typology*, 2.

Stephen Matthews and Virginia Yip. 2011. *Cantonese: A comprehensive grammar*, 2nd edition. Routledge.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

T.A. O'Melia. 1965. *First Year Cantonese*. Number v. 1 in First Year Cantonese. Catholic Truth Society.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Vitaly Protasov, Elisei Stakovskii, Ekaterina Voloshina, Tatiana Shavrina, and Alexander Panchenko. 2024. Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 94–108, Bangkok, Thailand. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Wesley Scivetti, Lauren Levine, and Nathan Schneider. 2025. Multilingual supervision improves semantic disambiguation of adpositions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3655–3669, Abu Dhabi, UAE. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Don Snow. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*, volume 1. Hong Kong University Press.

King Yiu Suen, Rudolf Chow, and Albert Y.S. Lam. 2024. Leveraging Mandarin as a pivot language

for low-resource machine translation between Cantonese and English. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 74–84, Bangkok, Thailand. Association for Computational Linguistics.

Chaoju Tang and Vincent J van Heuven. 2007. Mutual intelligibility and similarity of chinese dialects: Predicting judgments from objective measures. *Linguistics in the Netherlands*, 24(1):223–234.

Harish Thangaraj, Ananya Chenat, Jaskaran Singh Walia, and Vukosi Marivate. 2024. Cross-lingual transfer of multilingual models on low resource african languages.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. 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 843–857, Suzhou, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John Lee. 2017. Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 266–275, Pisa, Italy. Linköping University Electronic Press.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Rong Xiang, Ming Liao, and Jing Li. 2024. Cantonese natural language processing in the transformers era. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 69–79, Bangkok, Thailand. Association for Computational Linguistics.

Rong Xiang, Hanzhuo Tan, Jing Li, Mingyu Wan, and Kam-Fai Wong. 2022. When Cantonese NLP meets pre-training: Progress and challenges. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 16–21, Taipei. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Foong Ha Yap and Winnie Chor. 2011. Asymmetry in grammaticalization – the case of directional particles in cantonese. 2011 The 5th Conference on Language, Discourse and Cognition (CLDC-5) ; Conference date: 29-04-2011 Through 01-05-2011.

Odilla Yim and Ellen Bialystok. 2012. Degree of conversational code-switching enhances verbal task switching in cantonese–english bilinguals. *Bilingualism: Language and Cognition*, 15(4):873–883.

Bernard Yue and Andrew Gallant. 2014. *Hanzi Converter*. Version 0.2.2.

Xiaoheng Zhang. 1998. Dialect MT: A case study between Cantonese and Mandarin. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Ziqiong Zhang, Qiang Ye, Zili Zhang, and Yijun Li. 2011. Sentiment classification of internet restaurant reviews written in cantonese. *Expert Systems with Applications*, 38(6):7674–7682.