IberSPEECH 2022
14-16 November 2022, Granada, Spain

Neural Detection of Cross-lingual Syntactic Knowledge

Yongjian Chen[1], *Mireia Farrús*[1,2]

[1]Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona (UB), Barcelona, Spain
[2]Institut de Recerca en Sistemes Complexos (UBICS), Universitat de Barcelona (UB), Spain
yjchen@ub.edu, mfarrus@ub.edu

Abstract

In recent years, there has been prominent development in pre-trained multilingual language models, such as mBERT, XLM-R, etc., which are able to capture and learn linguistic knowledge from input across a variety of languages simultaneously. However, little is known about where multilingual models localise what they have learnt across languages. In this paper, we specifically evaluate cross-lingual syntactic information embedded in CINO, a more recent multilingual pre-trained language model. We probe CINO on Universal Dependencies treebank datasets of English and Chinese Mandarin for two syntax-related layer-wise evaluation tasks: Part-of-Speech Tagging at token level and Syntax Tree-depth Prediction at sentence level. The results of our layer-wise probing experiments show that token-level syntax is localisable in higher layers and consistency is shown across the typologically different languages, whereas sentence-level syntax is distributed across the layers in typology-specific and universal manners.

Index Terms: pre-trained multilingual language model, cross-lingual syntax, probing task

1. Introduction

In recent years, natural language processing has observed impressive development in pre-trained language models (PLMs), especially the advent of transformer-based PLMs like Bidirectional Encoder Representations from Transformers (BERT) [1], which learn contextual relations between words or sub-words in texts. BERT gradually displaces static word emebeddings like word2vec [2] and GloVe [3] thanks to their powerful contextualised representations and emerges as pre-trained encoders for other NLP models. The wide use of BERT models has yielded substantial improvements to numerous NLP tasks [1]. Following the great success of BERT, its multilingual variants, i.e., BERT-based models pre-trained on more than 100 languages have also been proposed, e.g., mBERT [1], XLM-R [4].

While BERT-based pre-trained language models have significantly boosted the state of the art on a great number of NLP applications, little is known about inner working of BERT-based PLMs. In order to gain better insights into their contextualised representations, BERTology comes into being, arousing a wave of interest in peeking into what linguistic knowledge is captured in these pre-trained contextualised representations [5]. One of the most prominent methods to explore the internal representations is *probing*, which aims to test what types of linguistic information are captured in different parts of PLMs through probing tasks. A probing task is to use contextualised representations derived from a PLM to train a classifier that predicts a linguistic property (e.g. part-of-speech, semantic role, etc.) for the input tokens. Early probing studies [6] [7] look at the 12 layers of BERT and suggests that linguistic knowledge is generally encoded in the same order of the traditional language processing pipeline, where surface-level features like word order are captured in lower layers, syntactic features more in middle layers and semantic features in higher layers [8]. The structural probing research [9] also suggests that syntactic knowledge that can reconstruct dependency trees is encoded in middle layers of BERT. However, not all PLMs derived from BERT localise linguistic knowledge in the same way as English BERT. Instead, they can exhibit different behaviours in encoding linguistic knowledge. For example, [10] evidences that morphosyntactic features are persistently distributed across all the layers of mBERT and XLM-R. This points to a belief that despite being based on BERT, different PLMS can still result in different contextualised representations and the behaviours in localising linguistic knowledge.

Of particular interest, we probe the syntactic properties of a multilingual language model extended from XLM-R called CINO [11], which is pre-trained on datasets of Chinese languages, including Standard Chinese, Cantonese and six other Chinese minority languages. The reason why we specifically carry out probing analysis on CINO is because as a further work, we seek to use appropriate multilingual representations to design a new metric in order to evaluate automatic speech recognition (ASR) output in English, Chinese Mandarin, and Chinese Cantonese from a syntactic perspective.

The structure of the current paper unfolds as follow. Section Related Work sheds light on two probing approaches, some background knowledge of CINO, and the widely used ASR evaluation metric. Section Experimental Setup describes the experiments performed in this work. Section Results and Discussion presents the layer-wise results and discuss the findings. Finally, Section Conclusions and Future Work summarises the findings and proposes the future work based on the limitations.

2. Related Work

2.1. Probing Approaches

As aforementioned, *probing* is a tool to measure how well linguistic properties are encoded in a PLM. In particular, the generic probing framework proposed by [6] [7] is called *edge probing* (See Figure 1). Edge probing can be applied to various prediction tasks, which can be represented as a labeled graph pertaining to spans in a sentence [6] [7]. Given a sentence of n tokens, edge probing formulates it as a set of k spans [$s^{(0)}$, $s^{(1)}$, $s^{(k)}$]. One or more tokens can be grouped in a span , so k is equal or inferior to n. As for multi-class single-label classification like part-of-speech tagging, a span contains one token. In contrast, as for multi-class single-label classification like constituent labelling, a span can contain more than one token as in Figure 1. Pre-trained contextualised representation(s) within the span is passed through a linear feedforward layer followed by a self-attention pooling that mainly contributes to transforming a variable-length sequence of span vectors into a

 10.21437/IberSPEECH.2022-31

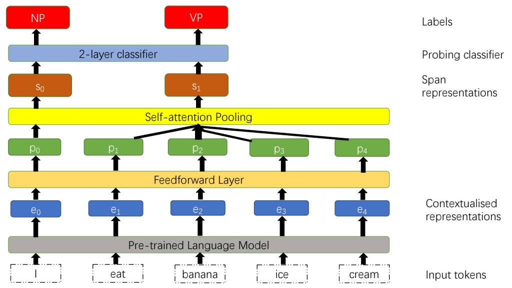

Figure 1: *Edge probing example for Top Constituent Labelling.*

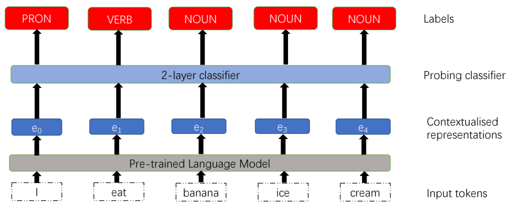

Figure 2: *Simple probing example for POS Tagging.*

single fixed-length span vector [6] [7]. This transformed span vector is fed into a 2-layer probing classifier that predicts a corresponding label [6] [7].

Alternative to edge probing, [12] proposes an easier probing approach (see Figure 2), which probes the pre-trained contextualised representations more directly. This simple probing approach only contains a classifier of a single linear layer without self-attention pooling operator, assuming that if a simple classifier can predict certain linguistic information from pre-trained contextualised representations, these representations are believed to encode this type of linguistic knowledge [12]. Therefore, based on this simple approach, rather than using intermediate representations as in edge probing, the probing classifier predicts a label immediately from the pre-trained contextualised representations. As a result, this simple approach can alleviate edge probing's shortcoming of introducing numerous new parameters that may conceal the source of success on probing tasks [13] This simple probing approach is particularly suitable for probing tasks that target at individual spans that have a single token, since fixed-length transformation through pooling is not necessary.

2.2. CINO

Instigated by the impressive success of BERT-based monolingual PLMs, the multilingual adaptations, i.e. the multilingual pre-trained language models (MPLMs) also advance rapidly. It is noted that MPLMs display an outstanding ability to understand a wide rage of languages and show prominent achievements in cross-lingual tasks [14]. CINO is a recently developed MPLM, which is transformer-based and shares the same network architecture with XLM-R [11]. CINO differs from XLM-R mainly in that CINO extends XLM-R's sentence-piece tokenizer and word embeddings by pre-training on a great vareity of Chinese languages including Chinese Mandarin and Chinese Cantonese [11]. In this work, we explore the CINO-base version, which has 12 layers, 768 hidden states, and 12 attention heads, with the vocabulary size 135,359, total model size 728MB, 68% size of XLM-R [11]. While CINO is suitable to make use of in our future work of new metric for ASR output since it can recognise English, Chinese Mandarin,and Chinese Cantonese scripts simultaneously, it also requires less computations costs.

2.3. ASR Evaluation Metric

As aforementioned, the motivation for us to probe CINO is that we seek propose a new metric for ASR performance at the syntactic level. Over the decades, Word Error Rate (WER) has been the predominant evaluation metric for ASR output [15]. WER relies on the word-level edit distance between the ground truth and predicted word, taking into account errors of substitution, insertion, and deletion [15]. However, there is an increasing interest in connecting ASR with NLP downstream tasks, especially the wide use of speech-enabled systems, e.g., voice assistants, spoken language translation system, etc.[16]. Since ASR output is fed as text input to the NLP models, so errors of ASR output can downgrade the performance on downstream tasks [16]. While syntactic correctness is a crucial part in the NLP pipeline, WER has limitations in evaluating syntactic aspect of ASR output. Therefore, it is worthwhile to design a novel evaluation metric for syntactic correctness of ASR output. To further the research into this novel metric, pobing syntactic information encoded in CINO is of great necessity.

3. Experimental Setup

To perform the evaluation tasks at different layers of CINO, we adopt the simple probing approach [12] to evaluate how well each layer of the pre-trained multi-lingual model captures syntactic information at both token level and sentence level.

3.1. Language and Dataset Selection

We include two languages in the probing experiments, i.e., English and Chinese Mandarin. In the first place, these two languages are the target ones in our future work. In addition, in terms of morphosyntax, English is a fusional language while Chinese Mandarin is an analytic language, and so these two languages exhibit a clear typological distinction. Chinese Mandarin and Chinese are of the same linguistic typology and share most syntactic aspects in common,mostly differing from each other in phonology, vocabulary, and utterance particles [17] [18]. Consequently, we exclude Chinese Cantonese from the probing experiments to save computational costs. By performing probing tasks with these two different languages, we can examine CINO's generalisation across typologically different languages in syntactic knowledge.

The Georgetown University Multilayer Universal Dependency Treebank (GUM-UDT) [19] is used for English and the simplified Chinese Universal Dependency Treebank (SC-UDT) [20] for Chinese Mandarin. Both datasets are split with provided partitions. GUM-UDT consists of 17 part-of-speech tags and 16 tree depth classes (tree depth values ranging from 0 to 15) , whereas SC-UDT consists of 15 POS tags and 9 tree depth classes (tree depth values ranging from 2 to 10).

3.2. Probing Tasks

Our selection of probing tasks is guided by two criteria. Firstly, the probing tasks should be representative of CINO's capability of syntax at different levels and thus should test syntax-related knowledge at both token level and sentence level. Secondly, the probing tasks should be workable with both English and Chinese Mandarin simultaneously. Guided by the criteria, the probing benchmarks comprise two syntactic evaluation tasks.

PART-OF-SPEECH TAGGING (POS) is a typical morphosyntactic task at token level, which seeks to assign word class label, e.g., noun, adjective etc. to each token of input sentences. While performance on POS is surface-level revelation of models' syntax-related ability, it makes sense to include it as the first probing benchmark in the sense that POS is a preliminary step in classic syntactic processing pipeline. Thus, POS can probe to what degree token-level syntax is captured in a pre-trained language model.

SYNTAX TREE-DEPTH PREDICTION (STDP) relates to the depth of the longest path from root to any leaf in a dependency tree. Predicting syntax tree-depth means categorising input sentences into the class relative to their depth. Being able to to perform accurate prediction suggests a model's ability to infer the hierarchical structure of sentences. Therefore, STDP can check how well sentence-level syntax is available in a pre-trained language model.

3.3. Probing Classifiers

Both probing tasks in this work target at single-token spans, i.e., each token being an individual span. Therefore, the simple probing approach is appropriate for our experiments. The probing classifier is then a single linear layer initialised with the same set of weights for all the models and tasks.

The input of the probing classifiers consists of pre-trained embeddings derived from CINO; that is, the output of each CINO layer (from layer 1 to layer 12). For each language, layer and task, we train one probing classifier, which allows us to quantify which layer encodes most information for a particular task and therefore is the most informative for the knowledge underpinning that task. During training, we freeze all CINO's parameters so that we can avoid the effect of fine-tuned CINO embeddings on the probing outcome and ascribe all learning to the probing classifiers. To describe the performance in the probing tasks, we report weighted F1-scores for both probing tasks.

Table 1: *Probing results across 12 layers of CINO.*

Layer	POS		STDP	
	EN	CN	EN	CN
Layer 1	0.8506	0.8231	0.3634	0.4137
Layer 2	0.9101	0.8744	**0.4253**	0.3971
Layer 3	0.9302	0.9003	0.3555	**0.4437**
Layer 4	0.9364	0.9086	0.3883	0.3796
Layer 5	0.9381	0.9137	0.3865	0.402
Layer 6	0.9399	0.9184	0.4081	0.3618
Layer 7	0.9465	0.9266	0.3952	0.3655
Layer 8	0.9567	0.9376	0.3683	0.3715
Layer 9	0.9611	0.9432	0.3818	0.3709
Layer 10	0.9618	0.9467	0.3333	0.4418
Layer 11	0.9618	**0.9486**	0.3497	0.4275
Layer 12	**0.9678**	0.948	0.3606	0.3723

4. Results and Discussion

We report our results in Table 1. As regards the token-level probing task (POS), CINO shows an overall increase in weighted F1-scores for both languages as the model's layer goes deeper (See Figure 3). The maximum F1-score was assigned to the last layer for English and to the penultimate layer for Chinese Mandarin. Yet, whether for English or Chinese Mandarin, the F1-scores show few differences from the last four layers (moderately over 0.96 for English and over 0.94 for Chinese Mandarin).

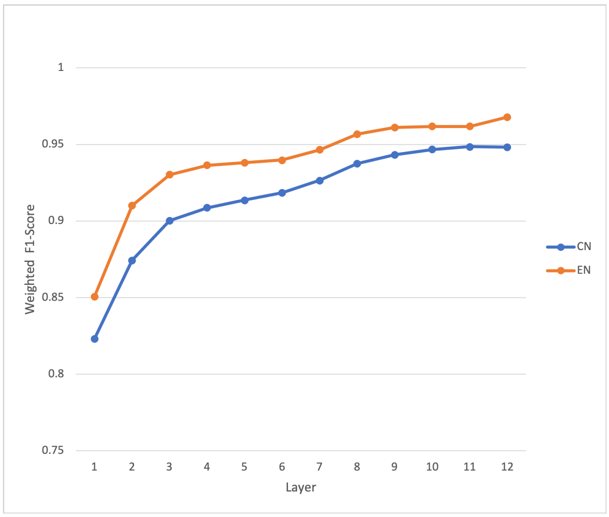

Figure 3: *Layer-wise weighted F1-scores for POS.*

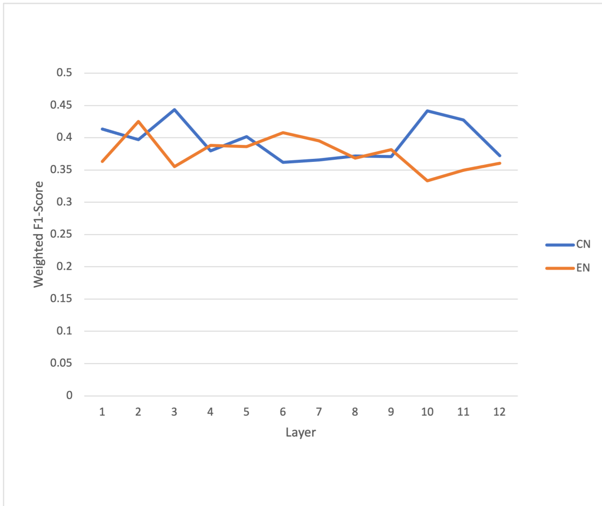

Figure 4: *Layer-wise weighted F1-scores for STDP.*

In contrast, regarding the sentence level probing task (STDP), CINO presents no consistent pattern but overall fluctuated weighted F1-scores across the 12 layers for both languages (see Figure 4). Interestingly, at the second layer, while the weighted F1-score climbs to the top for English, it reaches a local bottom for Chinese, but at the third layer, F1-score shows quite the opposite. This contrary pattern occurs similarly at the sixth and the tenth layers such that the F1-score displays a local maximum for one language and a local minimum for the other.

Our probing results indicate that for English and Chinese Mandarin, CINO carries more token-level information about POS in higher levels, which exhibits a different behaviour from what is encountered in English BERT where syntactic information is more prevalent in middle layers of a PLM [6] [7]. Besides, performance on POS is consistent across the two languages, and so provides evidence that CINO generalises token-level morphosyntactic knowledge across the two languages of distinct typological characteristics.

Unlike performance on POS, fluctuated scores of STDP are discerned in both English and Chinese Mandarin. We hypothesise that structural syntax is not located merely at some layers of the PLM, but instead, is spread across all the layers. This finding also shows inconsistency with the findings of classic syntactic processing pipeline [6] [7] [8]. Additionally, the layer-wise differences in local maxima and minima between English and Chinese are also interesting to shed light on. Throughout the first 12 layers of CINO, whereas English shows a nearly periodic rising-falling pattern, Chinese Mandarin depicts the contrary. Nevertheless, we also observe that the variance in F1-scores of STDP between English and Chinese is relatively small every four layers (layers 4, 8, 12). It seems possible that structural syntax encoded in some layers is typology-specific and that in the others is universal. For example, the second layer of CINO is more informative for sentence-level syntax pertaining to fusional languages like English, while the third layer is more responsible for that pertaining to analytic languages like Chinese. By contrast, hierarchical syntax encoded in the fourth layer of CINO tends to be universal across typologically different languages.

5. Conclusions and Future Work

In this work, we employ the simple edge probing approach with evaluation tasks to explore how capable the MPLM CINO is of token-level and sentence-level syntax at different layers. The experiments are performed with English and Chinese Mandarin. Our layer-wise probing results highlight that even though PLMS share similar architectures, they can still exhibit different behaviours in learning linguistic knowledge. In particular, our results reveal that CINO shows token-level syntactic localisation at higher layers (9+) for both of the typological distinct languages, thereby suggesting that pre-trained contextualised representations derived form CINO's higher layers are able to generalise token-level syntax across languages of typological differences. However, we reason that sentence-level syntax is distributed across the layers of CINO, but in two different manners. Some layers encode typology-specific hierarchical syntax while some layers encode the universal one. Furthermore, the probing findings can contribute to our future work where we seek to devise a novel metric for syntactic evaluation of ASR systems. The multilingual contextualised representations are suitable for the target languages of our ASR systems and features extracted from higher layers of CINO can be interpolated as token-level-syntax-aware cues into the evaluation metric.

On the other hand, the current work is limited to several aspects, mostly subject to the issue of diversity. Firstly, in this work, we only select one dataset for each language, which disallows us to control the potential bias caused by the datasets. Secondly, we perform experiments with only two languages, with each one representative of one linguistic typology (fusional versus analytic). With more languages from different language families used for each morphosyntactic typology, our probing results shall be more generalizable across a greater variety of languages. Last but not least, the probing tasks are not diverse enough, either, with one probing task relative to syntactic knowledge at a level. Therefore, as a continuation of the current work, we aim to improve the diverse coverage of datasets, languages, and probing tasks in order to further our research into probing the syntactic properties of CINO, especially peeking into whether and if yes, how typology-specific and universal encodings of sentence-level syntax co-exist in CINO.

6. Acknowledgements

The author J.C. would like to thank China Scholarship Council (CSS) for the financial support for his three year's PhD study.

7. References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[2] J. H. Lau and T. Baldwin, "An empirical evaluation of doc2vec with practical insights into document embedding generation," 2016, cite arxiv:1607.05368Comment: 1st Workshop on Representation Learning for NLP. [Online]. Available: http://arxiv.org/abs/1607.05368

[3] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[5] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.54

[6] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. V. Durme, S. R. Bowman, D. Das, and E. Pavlick, "What do you learn from context? probing for sentence structure in contextualized word representations," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=SJzSgnRcKX

[7] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4593–4601. [Online]. Available: https://aclanthology.org/P19-1452

[8] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3651–3657. [Online]. Available: https://aclanthology.org/P19-1356

[9] J. Hewitt and C. D. Manning, "A structural probe for finding syntax in word representations," 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). Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4129–4138. [Online]. Available: https://aclanthology.org/N19-1419

[10] R. Choenni and E. Shutova, "What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties," *ArXiv*, vol. abs/2009.12862, 2020.

[11] Z. Yang, Z. Xu, Y. Cui, B. Wang, M. Lin, D. Wu, and Z. Chen, "Cino: A chinese minority pre-trained language model," *arXiv preprint arXiv:2202.13558*, 2022.

[12] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith, "Linguistic knowledge and transferability of contextual representations," 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1073–1094. [Online]. Available: https://aclanthology.org/N19-1112

[13] T. Aoyama and N. Schneider, "Probe-less probing of BERT's layer-wise linguistic knowledge with masked word prediction," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, Jul. 2022, pp. 195–201. [Online]. Available: https://aclanthology.org/2022.naacl-srw.25

[14] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 833–844. [Online]. Available: https://aclanthology.org/D19-1077

[15] S. Kim, A. Arora, D. Le, C. feng Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Semantic distance: A new metric for asr performance analysis towards spoken language understanding," in *Interspeech*, 2021.

[16] T. Cui, J. Xiao, L. Li, X. Jiang, and Q. Liu, "An approach to improve robustness of nlp systems against asr errors," *ArXiv*, vol. abs/2103.13610, 2021.

[17] T.-s. Wong, K. Gerdes, H. Leung, and J. Lee, "Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank," in *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Pisa,Italy: Linköping University Electronic Press, Sep. 2017, pp. 266–275. [Online]. Available: https://aclanthology.org/W17-6530

[18] O. DALTE, "Key differences in mandarin and cantonese that complicate their mutual understanding," *I I*, p. 202052, 2020.

[19] A. Zeldes, "The GUM corpus: Creating multilayer resources in the classroom," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 581–612, 2017.

[20] P. Qi and K. Yasuoka, "Simplified chinese universal dependencies dataset."