CompLex-ZH: A New Dataset for Lexical Complexity Prediction in Mandarin and Cantonese

Le Qiu[1], Shanyue Guo[1], Tak-sum Wong[1],
Emmanuele Chersoni[1], John S. Y. Lee[2], Chu-Ren Huang[1]

[1]The Hong Kong Polytechnic University, [2]City University of Hong Kong
Correspondence: emmanuele.chersoni@polyu.edu.hk

Abstract

The prediction of *lexical complexity* in context is assuming an increasing relevance in Natural Language Processing research, since identifying complex words is often the first step of text simplification pipelines. To the best of our knowledge, though, datasets annotated with complex words are available only for English and for a limited number of Western languages.

In our paper, we introduce *CompLex-ZH*, a dataset including words annotated with complexity scores in sentential contexts for Chinese. Our data include sentences in Mandarin and Cantonese, which were selected from a variety of sources and textual genres. We provide a first evaluation with baselines combining hand-crafted and language models-based features.

1 Introduction

In psycholinguistics and Natural Language Processing (NLP) research, the notion of *complexity* relates to the difficulty faced by a speaker in reading and understanding specific linguistic productions (Blache, 2011; Chersoni et al., 2016, 2017, 2021; Sarti et al., 2021; Iavarone et al., 2021; Xiang et al., 2021), and its assessment has important applications in education technology, such as the simplification of text for second language learners and/or populations with special needs (Štajner, 2021; North et al., 2023). One major source of complexity is depending on word choice, it corresponds to the difficulty that one may encounter in understanding a specific word in context, which could be solved with the help of NLP systems by i) automatically identifying the complexity of the target word (*lexical complexity in context*); ii) proposing simpler and more familiar words as replacements (*lexical simplification*).

Although the problem of lexical complexity received increasing attention in the NLP community in the last few years (Shardlow et al., 2020; Štajner et al., 2022; Ai, 2022; Yang et al., 2023), with the introduction of new benchmark datasets and the organization of several shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021; Saggion et al., 2023; Shardlow et al., 2024), the evaluation in this task has been so far limited to a small number of Western languages.

Our research effort aims at filling this gap, by introducing **CompLex-ZH**, the first evaluation benchmark for lexical complexity prediction in Chinese. CompLex-ZH includes data annotated for word complexity in context by native speakers, and has been built by carefully sampling sentences from different sources and text genres. In addition to Mandarin Chinese, we also provide lexical complexity data for Cantonese: Cantonese is a major variety of Chinese with a large population of speakers worldwide (more than 85 million of speakers, according to recent estimates by Eberhard et al. (2022)) but having a low-resource status in terms of availability of NLP models, corpora and resources, and thus it might prove more challenging to handle for LLMs trained on standard Chinese (Xiang et al., 2024). Our initial evaluation results show that a baseline regressor based on a combination of handcrafted

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 20–26
November 15, 2024 ©2024 Association for Computational Linguistics

features and contextualized embeddings only reaches a moderate accuracy in predicting Chinese lexical complexity. [1]

2 Related Work

An early shared task on lexical complexity in English was organized in 2016 by Paetzold and Specia (2016), with complexity defined as a binary variable: raters and automatic systems had to decide whether a word was complex/difficult to understand or not. Of course, this is a simplifying and problematic assumption, as there are many situations in which word complexity cannot be determined in a clear-cut way, and it is better described by a continuous value. Another shared task in 2018 (Yimam et al., 2018) also focused on complexity prediction as a binary decision, but it included an additional regression subtask in which the systems, given a target word in context and a specific annotator, had to predict the likelihood that the annotator would have considered the target as complex.

The Task 1 at SemEval-2021 (Shardlow et al., 2021) was the first one treating lexical complexity prediction as a regression task. As a gold standard, this shared task used the CompLex corpus (Shardlow et al., 2020, 2022), which includes words in sentential contexts from three different genres, i.e. the Bible, the proceedings of the European Parliament and biomedical articles, and the scores are mean complexity ratings between 0 and 1. Moreover, this benchmark features not only single words as targets, but also multiword expressions.

Notice that the identification of complex words is only the first step of pipelines aiming at the lexical simplification of a text (Saggion and Hirst, 2017). Additional steps generally require the generation of simpler substitution words and their ranking. Over the years, several studies have been dedicated to lexical simplification in English (Paetzold and Specia, 2017; Qiang et al., 2020), Chinese (Qiang et al., 2021), Portuguese (North et al., 2022) and Spanish (Ferrés and Saggion, 2022), and a shared task has been organized in co-location with EMNLP 2022 (Saggion et al., 2023). In many of these studies, a target word has already been identified as complex and the systems have to focus on the substitute generation and ranking component: for example, the selection of the target words in the Chinese dataset by Qiang et al. (2021) only includes words classified as "high-level" (meaning, understandable only by advanced speakers) in the Chinese HSK Vocabulary (Zhao et al., 2003). Our current work is focusing instead on the previous step of lexical complexity detection, and aims at providing the first benchmark for the Chinese language with words of varying degrees of complexity.

3 Dataset Creation

In order to create a challenging benchmark, we decided to include not only data for lexical complexity in Mandarin Chinese, which is the standard variety of Chinese, but also for Yue Chinese or Cantonese. Cantonese is commonly used in colloquial scenarios (e.g., daily conversation and social media) and it exhibits different vocabulary, grammar, and pronunciation compared to Mandarin. It is natively spoken by a large number of speakers in Hong Kong, Macao, Guangdong and part of Guangxi, and in many overseas Chinese communities in South-East Asia, North America, and Western Europe (Sachs and Li, 2007; Yu, 2013; Xiang et al., 2024).

We believe its inclusion is an interesting feature of our dataset, as it will allow to test the robustness of Chinese language models to different Chinese varieties. This is useful because, despite being a mainly spoken variety, Cantonese can also be used in some written contexts, such as the Legislative Council of the Hong Kong Special Administrative Region, in medical document transcriptions, or in sections of special interests of local newspapers (Xiang et al., 2024), and thus there might be the need for simplification of Cantonese texts.

3.1 Target Selection and Sentence Sampling

In constructing our dataset, we collected most Mandarin data from the Chinese Wikipedia (*Zh-Wikipedia*), *Weibo* and *People's Daily*, and most Cantonese data from the Cantonese Wikipedia (*Yue-Wikipedia*) and from the *LIHKG* dataset for topic classification. [2]

[1] Code and data will be made available at https://github.com/Laniqiu/CompLex-ZH.

[2] https://github.com/toastynews/lihkg-cat-v2.

People's Daily stands as one of the most authoritative newspapers in China, while Weibo is a popular micro-blogging site in mainland China, and LIHKG is a Reddit-like forum based in Hong Kong. We also sourced supplementary materials (categorized as *Other*) from the BCC corpus (Xun et al., 2016) for Mandarin, from a counseling corpus (Lee et al., 2020) and from the PolyU Corpus of Spoken Chinese (Hong Kong Polytechnic University, 2015) for Cantonese. By incorporating these varied sources, we managed to cover a wide range of topics, including daily life, sports, public health, politics, and so on.

The raw materials have been tokenized, using Jieba[3] for Mandarin and PyCantonese (Lee et al., 2022) for Cantonese. We examined the vocabulary of our corpora and identified target words. We primarily found high-frequency content words that are more colloquial and frequently encountered in everyday conversations from Weibo, People's Daily, LIHKG, etc.. Low-frequency content words were also included from BCC and Wikipedia, offering a broader spectrum of vocabulary beyond colloquial expressions. We then sampled at least 1 sentence for each target. Multiword expressions, following Shardlow et al. (2021), could also be chosen as targets (They constitute 1.97% of Mandarin targets and 2.69% of Cantonese targets.). These targets and the sentences including them made up then for our unrated datasets. Finally, it should be mentioned that the number of target words we could sample is much lower for Cantonese, as our Cantonese corpora were much smaller (i.e. with lower frequencies for the candidate target words) and with less variety of textual genres.

3.2 Rating Collection

For rating collection, we created about 300 questionnaires for both varieties, using the data from section 3.1. Participants are requested to evaluate the difficulty in understanding the given words within the given contexts. The provided options are designed in a 5-point Likert scale, ranging from 1 indicating *Very easy* to 5 *Very difficult*. Each questionnaire consists of about 100 rating questions and 2 validation questions. The validation samples are prepared with *gold* answers. Before the questionnaire distribution, we conducted a small pilot study, and identified some questions where participants highly agreed on the options (i.e., *gold answers*). These are then inserted in the instruction messages, and in the questionnaires. The annotators whose answers significantly deviated from the gold answers for those samples were considered as non-reliable raters and their responses were rejected. Concretely, we had examples where all the pilot study participants gave very low/easy scores, such as the Cantonese 呢份試卷好**簡單**呀！(*This exam is too easy*), or very high/difficult scores, such as the Mandarin 她谈着她婚后的**睽离**和甜蜜的生活 (*She talked about her detached and sweet life after marriage*): an annotator's answer were discarded if easy validation questions were rated higher than 3 (the mid point of the scale), and difficult ones were rated lower than 3.

Our raters (refer to the Appendix for more information) have mostly been recruited in Hong Kong, where Cantonese is the principal vernacular language and Mandarin Chinese is one of the official languages. Each rater was paid 100 HKD (≈ 12.8 USD) for a single questionnaire.

Each sample has been rated by at least 5 raters. The complexity score of a sample is then the average score assigned by all the raters, while the complexity score of a target word is the average of the scores of all its samples. We provided some examples in Table 1. The statistics of the dataset can be found in Table 3 in the Appendix.

Context	Score
... 忽然变得澄清见底，翳障 全无。 ...it turns crystal, without obstacles in sight.	.213
此前有团队 已经在粪便里发现新冠病毒。 The team had found coronavirus in feces.	.893
... 感受到被失蹤、被跟蹤的實在... ...I truly felt disappeared and stalked...	.588
點解講GOOD JOB 佢反而又呆哂... Why he acts so dumb and ... when you said GOOD JOB?	.200

Table 1: Some examples with high/ low complexity scores. The first 2 are in Mandarin and the last 2 in Cantonese. Target words are underlined.

[3] https://github.com/fxsjy/jieba

	Feat.	MAE	R^2	ρ
Mand.	HC	.065	.186	.091
	Stroke	.065	.083	.107
	WLen	.065	.055	.082
	LogF	.065	.201	.061
	Emb	**.059**	**.355**	**.338**
	Comb.	.060	.086	.322
Canto.	HC	**.060**	.051	.191
	Stroke	.063	-.001	.008
	WLen	.063	.0184	.158
	LogF	.061	.022	.149
	Emb	.061	**.056**	.353
	Comb.	.061	.045	**.354**
Joint	HC	.065	.047	.135
	Stroke	.066	-.002	-.015
	WLen	.066	-.002	-.109
	LogF	.066	.040	.116
	Emb	**.062**	.131	**.329**
	Comb.	**.062**	**.136**	.326

Table 2: Summary of evaluation results. We investigated overall and individual HC features, embedding features and the combination of the most influential LogF feature and of the word embeddings (Comb.). The metrics are: mean absolute error (MAE), R-squared value (R^2), and Spearman correlation coefficient (ρ). Notice that the metrics are not directly comparable across language settings, given the different number of items in the test sets.

4 Evaluation Experiments

We ran some preliminary evaluation experiments using a Ridge Regression model with handcrafted features (HC) and contextualized word embeddings (Emb) as predictors and the complexity score for each sentence as the target variable. The data were splitted in training, validation and test set with a 8:1:1 split percentage, and we ran separate evaluations for the two varieties and for the two feature types to assess their impact. Handcrafted features of the target word include its logarithmic frequency (LogF), extracted via the Wordfreq Python package (Speer, 2022); the number of characters (WLen) and the number of strokes[4], which are well-known visual complexity indexes (Tse et al., 2017; Sun et al., 2018). For the contextualized embeddings in the two varieties, we used CINO (Yang et al., 2022), a RoBERTa-based architecture trained on texts both in standard Chinese and in several minority languages of China, in order to obtain vector representations for both Mandarin and Cantonese by means of a single model.

The most common regression metrics have been calculated on the test sets (324 instances for Mandarin, 250 for Cantonese, 574 for a joint dataset of both) and are shown in Table 2. We can notice that, in each corpus partition, contextualized embeddings contribute to improve significantly over the results of the out-of-context HC features, and among those features, it can be seen that logarithmic frequency is predictive of lexical complexity in Mandarin, but it performs much more weakly in the settings including Cantonese data, which might be due to a more limited coverage of frequency norms in this variety. Compared to the original results in Shardlow et al. (2020) on English, it is interesting to observe that on our data HC features are not consistently more informative than embeddings. This could be due to differences in the embeddings type: static embeddings from GloVe (Pennington et al., 2014) and sentence embeddings from InferSent (Conneau et al., 2017) in the previous work, contextualized, token-level embeddings from a more recent RoBERTA-based architecture in our present evaluation.

We can also observe that correlation values and MAE in Mandarin and Cantonese are similar, but explained variance in Cantonese is much lower, confirming that the Cantonese data pose a non-trivial challenge for Chinese NLP. Scores in general are relatively low, suggesting the need for more sophisticated approaches to improve the modeling of lexical complexity in Chinese.

5 Conclusion

In this paper, we have introduced CompLex-ZH, the first dataset for evaluating predictions of lexical complexity in context for two major Chinese varieties, Mandarin and Cantonese. We have sampled target words in context from a variety of text genres and collected ratings from speakers in Hong Kong.

Our preliminary evaluation shows that the contextualized embeddings of a language model trained on multiple Chinese varieties significantly help in improving the prediction over handcrafted, out-of-context features. However, the accuracy is not high - as suggested by the limited amount of explained variance and by the weak-to-moderate correlation scores, leaving space for future improvements.

[4] Source: https://github.com/WuChengqian520/

Lay Summary

The first step that a computer has to take to simplify a text and make it more accessible is to identify difficult words and expressions. So far, datasets to train machine learning systems to recognize the complexity of understanding words in context (lexical complexity) have been available only for English and a few other Western languages.

In our work, we put together a dataset of human complexity judgements for words in context in Chinese, and we included two different Sinitic varieties: Mandarin and Cantonese. We carry out a first test for predicting lexical complexity in Chinese, and obtained our best results with the features extracted from CINO, a language model trained on multiple Chinese dialects. On the other hand general accuracy remains moderate, as the models seem to struggle with the more rare and data scarce Cantonese variety.

Acknowledgements

EC acknowledges the financial support from the start-up fund project “Building and Predicting Neurocognitive-Motivated Lexical-Semantic Norms for Mandarin Chinese” (1-BE8G), sponsored by the Faculty of Humanities of the Hong Kong Polytechnic University. JL acknowledges support from a Strategic Research Grant (7006037) at City University of Hong Kong.

References

Haiyang Ai. 2022. Automating Lexical Complexity Measurement in Chinese with WeCLECA. *International Journal of Asian Language Processing*, 32(01):2250011.

Philippe Blache. 2011. Evaluating Language Complexity in Context: New Parameters for a Constraint-based Model. In *Proceedings of the International Workshop on Constraints and Language Processing*.

Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2016. Towards a Distributional Model of Semantic Complexity. In *Proceedings of the COLING Workshop on Computational Linguistics for Linguistic Complexity*.

Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Logical Metonymy in a Distributional Model of Sentence Comprehension. In *Proceedings of *SEM*.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2021. Not All Arguments Are Processed Equally: A Distributional Model of Argument Complexity. *Language, Resources and Evaluation*, pages 1–28.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of EMNLP*.

David M Eberhard, Gary F Simons, and Charles D Fennig. 2022. *Ethnologue: Languages of the World*. Dallas: SIL International.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A Dataset for Lexical Simplification in Spanish. In *Proceedings of LREC*.

Department of English Hong Kong Polytechnic University. 2015. PolyU Corpus of Spoken Chinese.

Benedetta Iavarone, Dominique Brunato, and Felice Dell’Orletta. 2021. Sentence Complexity in Context. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Jackson Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. PyCantonese: Cantonese Linguistics and NLP in Python. In *Proceedings of LREC*.

John Lee, Tianyuan Cai, Wenxiu Xie, and Lam Xing. 2020. A Counselling Corpus in Cantonese. In *Proceedings of the Joint SLTU and CCURL Workshop (SLTU-CCURL)*.

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022. ALEXSIS-PT: A New Resource for Portuguese Lexical Simplification. *arXiv preprint arXiv:2209.09034*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 Task 11: Complex Word Identification. In *Proceedings of SemEval*.

Gustavo Paetzold and Lucia Specia. 2017. Lexical Simplification with Neural Ranking. In *Proceedings of EACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of EMNLP*.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. LSBert: A Simple Framework for Lexical Simplification. *arXiv preprint arXiv:2006.14939*.

Jipeng Qiang, Xinyu Lu, Yun Li, Yunhao Yuan, and Xindong Wu. 2021. Chinese Lexical Simplification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1819–1828.

Gertrude Tinker Sachs and David CS Li. 2007. Cantonese as an Additional Language in Hong Kong. *Multilingua*, 26(95):130.

Horacio Saggion and Graeme Hirst. 2017. *Automatic Text Simplification*, volume 32. Springer.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2023. Findings of the TSAR-2022 Shared Task on Multilingual Lexical Simplification. *arXiv preprint arXiv:2302.02888*.

Gabriele Sarti, Dominique Brunato, and Felice Dell'Orletta. 2021. That Looks Hard: Characterizing Linguistic Complexity in Humans and Language Models. In *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon-Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex–A New Corpus for Lexical Complexity Predicition from Likert Scale Data. In *Proceedings of the LREC Workshop on Tools and Resources to Empower People with REAding DIfficulties*.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of SemEval*.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting Lexical Complexity in English texts: The Complex 2.0 Dataset. *Language Resources and Evaluation*, 56(4):1153–1194.

Robyn Speer. 2022. rspeer/wordfreq: v3. 0. *Version v3. 0.2. Sept*.

Sanja Štajner. 2021. Automatic Text Simplification for Social Good: Progress and Challenges. *Findings of ACL-IJCNLP*.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical Simplification Benchmarks for English, Portuguese, and Spanish. *Frontiers in Artificial Intelligence*, 5.

Ching Chu Sun, Peter Hendrix, Jianqiang Ma, and Rolf Harald Baayen. 2018. Chinese Lexical Database (CLD) a Large-Scale Lexical Database for Simplified Mandarin Chinese. *Behavior Research Methods*, 50:2606–2629.

Chi-Shing Tse, Melvin J Yap, Yuen-Lai Chan, Wei Ping Sze, Cyrus Shaoul, and Dan Lin. 2017. The Chinese Lexicon Project: A Megastudy of Lexical Decision Performance for 25,000+ Traditional Chinese Two-character Compound Words. *Behavior Research Methods*, 49:1503–1519.

Rong Xiang, Emmanuele Chersoni, Yixia Li, Jing Li, Chu-Ren Huang, Yushan Pan, and Yushi Li. 2024. Cantonese Natural Language Processing in the Transformers Era: A Survey and Current Challenges. *Language Resources and Evaluation*, pages 1–27.

Rong Xiang, Jinghang Gu, Emmanuele Chersoni, Wenjie Li, Qin Lu, and Chu-Ren Huang. 2021. PolyU CBS-Comp at SemEval-2021 Task 1: Lexical Complexity Prediction (LCP). In *Proceedings of SemEval*.

Endong Xun, Gaoqi Rao, Xiaoyue Xiao, and Jiaojiao Zan. 2016. The Construction of the BCC Corpus in the Age of Big Data. *Corpus Linguistics*.

Cheng-Zen Yang, Jin-Jian Li, and Shu-Chang Lin. 2023. Lexical Complexity Prediction using Word Embeddings. In *Proceedings of ROCLING*.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese Minority Pre-trained Language Model. In *Proceedings of COLING*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the NAACL Workshop on Innovative Use of NLP for Building Educational Applications*.

Henry Yu. 2013. Mountains of Gold: Canada, North America, and the Cantonese Pacific. In *Routledge Handbook of the Chinese Diaspora*, pages 124–137. Routledge.

J Zhao, B Zhang, and J Cheng. 2003. Some Suggestions on the Revision of the Outline of the Graded Vocabulary for HSK. *Chinese Teaching in the World*.

A Statistics of the dataset

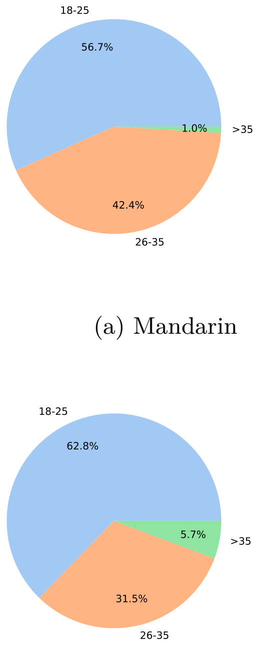
Table 3 presents the statistics of the target words, samples, and ratings of the dataset.

	Source	Sent.	Word	Sent./ Word	R./ Sent.	R./ Word	Complex.	STD
Mand.	Weibo	1600	770	2.08	8.28	17.21	.269	.061
	People's Daily	1228	713	1.72	8.36	14.41	.268	.064
	Other	412	255	1.62	6.61	10.67	.323	.164
	All	3240	1017	3.19	8.10	25.80	.283	.094
Canto.	LIHKG	1043	222	4.70	9.45	6.97	.284	.077
	Wiki	1037	219	4.74	6.97	32.99	.268	.073
	Other	425	129	3.29	9.10	29.98	.274	.067
	All	2505	260	9.63	8.36	80.58	.274	.065

Table 3: Statistics of CompLex-ZH. The original ratings have been normalized to a 0-1 range following Shardlow et al. (2020)'s convention: $1 \rightarrow 0, 2 \rightarrow 0.25, 3 \rightarrow 0.5, 4 \rightarrow 0.75, 5 \rightarrow 1$. Denotation: Mand. = Mandarin, Canto. = Cantonese, Sent. = Sentence, R./ Sent. = Ratings per Sent., R./ Word = Ratings per word, Complex. = average word-wise complexity score, STD = Standard deviation.

B Background information of raters

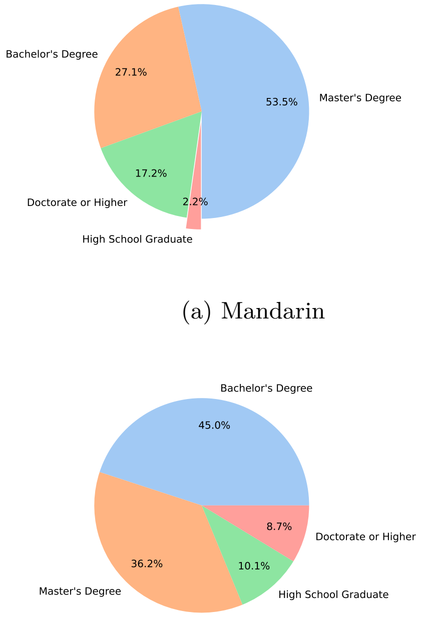
We recruited 318 raters for the Mandarin dataset, and 299 raters for Cantonese. After rejecting some annotators' answers, following the validation procedure in section 3.2, we eventually have 314 raters for Mandarin and 298 raters for Cantonese. As shown in Figure 1 and Figure 2, most of our raters are aged between 18 to 35, holding a bachelor or a higher level degree.



(a) Mandarin

(b) Cantonese

Figure 1: Age Distribution of annotators.



(a) Mandarin

(b) Cantonese

Figure 2: Education levels of annotators.