ShUD: the First Shanghainese Universal Dependency Treebank

Qizhen Yang
Shanghai World Foreign Language Academy
Shanghai, China
qzyang.main@gmail.com

Abstract

This paper introduces **ShUD**[1], the first Universal Dependencies (UD) treebank for Shanghainese, a Wu Chinese variant spoken by approximately 14 million people but severely under-resourced in NLP. The treebank is built through a scalable annotation pipeline that exploits grammatical parallels between Shanghainese and Mandarin. Our pipeline also provides a practical strategy for bootstrapping resources for other Chinese dialects. We documented syntactic phenomena unique to Shanghainese within the UD framework and fine-tuned a dependency parser[2] using our annotated treebank, contributing a foundation to both NLP tool development and cross-linguistic syntactic research.

1 Introduction

Shanghainese, the largest branch of Wu Chinese spoken by about 14 million of the overall 83 million Wu Chinese speakers (Pan et al., 1991; Xie, 2011), remains severely under-resourced in computational linguistics – an issue common among non-Mandarin Chinese varieties. To date, no annotated corpora exist for Shanghainese or any other Wu Chinese variety, in stark contrast to the growing availability of resources for Mandarin.

This lack of data hinders the development of NLP tools and limits linguistic research on Shanghainese. In this paper, we introduce **ShUD**, the first UD treebank for Shanghainese[3]. ShUD contains 983 sentences and 8,584 tokens. We design a sustainable annotation pipeline that leverages grammatical parallelism with Mandarin to improve annotation efficiency and quality. Our approach may extend to other Chinese varieties exhibiting similar syntactic traits. We document Shanghainese-specific constructions within the UD framework (De Marneffe et al., 2021). We also fine-tune a biaffine dependency parser using our data.

2 Related Work

2.1 Chinese Variants in the UD Project

Chinese is underrepresented in the UD project in both volume and variety. Existing treebanks focus almost exclusively on formal Mandarin (e.g., Poiret et al., 2023; Zeman et al., 2017), with rare exceptions for Cantonese (Wong et al., 2017) and Classical Chinese (Yasuoka, 2019; Yasuoka et al., 2022). To date, no other Chinese dialects have been included (Nivre et al., 2020).

2.2 Shanghainese and its NLP Resources

Although Shanghainese lacks annotated corpora, its grammar has been the subject of extensive linguistic study, and several grammar books (Qian, 1997; Zhu, 2006) and dictionaries are available. For instance, the *Shanghainese Dictionary* lists over 20,000 entries (Grayson, 2025). The Wu Chinese Society offers a comprehensive dictionary covering modern and historical Wu Chinese (Wu Chinese Society, 2009), including entries specific to Urban Shanghainese. Wu Chinese expressions are also available on Wiktionary (Wiktionary Contributors, 2024), though many reflect non-Shanghainese varieties.

Recent progress in speech processing has led to the release of a few colloquial Shanghainese corpora. Notably, Magic Data has published two speech corpora for Shanghainese conversation (Magic Data, 2021a,b).

2.3 Spoken Language Treebanks

The UD framework has expanded to include spoken language resources. These encompass di-

[1] https://github.com/UniversalDependencies/UD_Shanghainese-ShUD

[2] https://huggingface.co/q1zhen/ShUD

[3] Shanghainese includes several geographical and historical variants. We focus on Middle and New Period Urban Shanghainese; Old Period Shanghainese, used around a century ago, is no longer spoken.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 186–193
August 27, 2025 ©2025 Association for Computational Linguistics

verse languages including Slovenian (Dobrovoljc and Nivre, 2016), Cantonese (Wong et al., 2017), etc., with many representing the only available UD resources for low-resource languages (for details, see Dobrovoljc, 2022). These treebanks exhibit considerable variation in transcription approaches, annotation principles, and treatment of speech-specific phenomena such as fillers, disfluencies, and repairs (Dobrovoljc, 2022).

3 Data Source and Features

We use the open-source *Scripted Chinese Shanghai Dialect Daily-use Speech Corpus* (ASR-SCShhiDiaDuSC, hereafter **A.-S.**) as the data source for our treebank. The corpus focuses on daily-use speech, providing an accessible and representative sample of contemporary Shanghainese[4] (Magic Data, 2021b).

The A.-S. corpus contains 4.23 hours of transcribed Shanghainese speech, totalling 4,819 utterances from 10 speakers. While originally created for speech processing, the corpus is well-suited for syntactic annotation. Speakers read Mandarin Chinese prompts aloud in Shanghainese, adapting vocabulary and structure naturally, which is the most common practice for native speakers to read the language from text in everyday life (due to the lack of standardised orthography and formal texts in Shanghainese). Given the high lexical overlap between the two languages in colloquial contexts, such adaptations produce fluent Shanghainese expressions rather than literal translations. Since Shanghainese is primarily used in colloquial contexts, the corpus is particularly well-suited to represent the language, and sentences are typically short. We use only the textual transcriptions for our annotation. An example instance from the dataset is shown below:

Prompt	这点机会也可能没有。
[Mandarin]	(zhe dian ji hui ye ke neng mei you)[5]
Transcription	搿眼机会啊可能没了。[6]
[Shanghainese]	(geh ngae ci ue a khu nen meh leh)[7]
Gloss	This bit chance also possible no.
Translation	*There may not be any chance of this.*

4 Treebank Construction

4.1 Pipeline Overview

Although Shanghainese and Mandarin share many syntactic properties, off-the-shelf UD parsers perform poorly on Shanghainese due to major lexical differences, especially in the use of particles[8]. Leveraging the strong performance of existing Mandarin parsers, we designed a hybrid annotation pipeline: Shanghainese utterances are first manually transliterated into Mandarin vocabulary, then automatically tokenised and parsed by a parser. The original Shanghainese vocabulary is then restored, followed by thorough manual verification and correction at each stage. Each mapping is also saved for future reference to reduce manual workload. Since we are performing dependency annotations (in contrary to constituency), the relations are transferrable between the languages as the they do not contain structural information (De Marneffe and Nivre, 2019). We utilise Stanza (Qi et al., 2020) `v1.10.1`, which's Mandarin Chinese parser is trained upon the GSDSimp treebank[9].

This approach is especially advantageous. It significantly reduces the workload of annotators and boosts efficiency by reusing pre-trained models. The consistency can be improved because fatigue-related errors by annotators can be reduced.

Figure 1 shows the detailed pipeline. Section 6 presents the evaluations on the effects of different steps of the pipeline.

4.2 Inter-annotator Agreement

Annotation was conducted by two annotators fluent in both Shanghainese and Mandarin, each ex-

[4] Licensed under Creative Commons BY-NC-ND 4.0. Additional permission for derivative research was granted by Beijing Magic Data Technology Co., Ltd.

[5] In this paper, Mandarin Chinese words will be transcribed in Hanyu Pinyin without tone marks since precise tonal values are not essential in this project.

[6] There are standardised Shanghainese orthographies proposed by some scholars, but they are never widely accepted or used for a limited usage in writing. A common practice (which is also used by the corpus) is to use Mandarin words with similar pronunciations. In this paper, we will present the corpus texts as is and list out as many possible transcriptions as possible in theoretical discussion contexts.

[7] Similarly, existing standardised Romanisation schemes for Shanghainese are also rarely used, especially that pronunciations are rapidly evolving over time due to the influence of Mandarin. In this paper, the pronunciations, not guaranteeing their accuracy, will be mostly taken from Wu Chinese Society's dictionary; precise pronunciations are not essential in this project.

[8] For example, the Shanghainese word 伐 *veh* can function either as a negator (also transcribed as 勿 or 弗, equivalent to Mandarin 不 *bu*) or as a question particle without semantic content. In contrast, in Mandarin, the same character 伐 *fa* primarily means "to cut down [wood]" and is treated as a verb by standard parsers, resulting in incorrect POS tagging when applied to Shanghainese.

[9] `https://universaldependencies.org/treebanks/zh_gsdsimp/index.html`

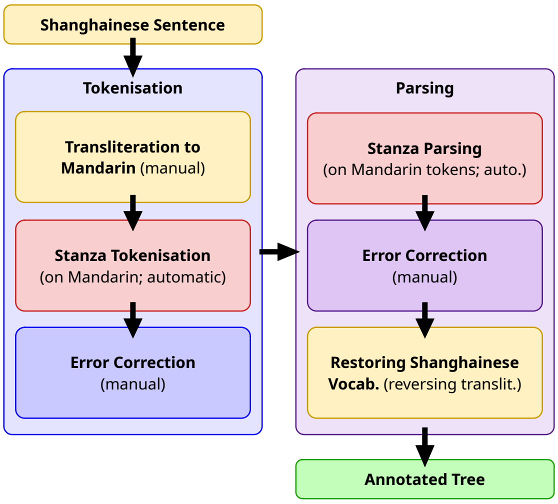

Shanghainese Sentence
Tokenisation
Transliteration to Mandarin (manual)
Stanza Tokenisation (on Mandarin; automatic)
Error Correction (manual)
Parsing
Stanza Parsing (on Mandarin tokens; auto.)
Error Correction (manual)
Restoring Shanghainese Vocab. (reversing translit.)
Annotated Tree

Figure 1: Overview of the annotation pipeline.

tensively trained in UD with a focus on Chinese-specific guidelines.

The first 100 sentences were independently annotated twice. Inter-annotator agreement reached 92.09% for UPOS tagging, 87.82% for Unlabelled Attachment Score (UAS), 81.89% for Labelled Attachment Score (LAS), and 99.09% for tokenisation. Most disagreements arose from differing interpretations of lexical ambiguity in contextually underspecified sentences. An example is provided in Appendix A. These cases are rare and do not substantially affect the overall annotation quality.

5 Annotation Guideline

This section outlines the principles and guidelines followed in the annotation process.

5.1 Text Segmentation and Tokenisation

Our tokenisation scheme follows the principles of the Penn Chinese Treebank (Xue et al., 2005), where a word, defined as one or more characters forming a lexical unit, serves as the basic unit of annotation.

Function words are treated as separate tokens, even when phonologically or morphologically attached to verbs. These include items such as 了 *leh*, marking the perfective aspect, and 勒 *leh*, indicating the continuation of an action.

Following the UD guidelines for Chinese[10], we treat compound words and multi-word expressions in Shanghainese similarly to idiomatic expressions in Mandarin Chinese (e.g., Chengyus). While Mandarin Chengyus are typically fossilised expressions from Classical Chinese, many idiomatic phrases in Shanghainese derive from Wu Chinese or from historical transliterations and reinterpretations of European words. These expressions function as lexicalised units and are not analysable by Shanghainese grammar.

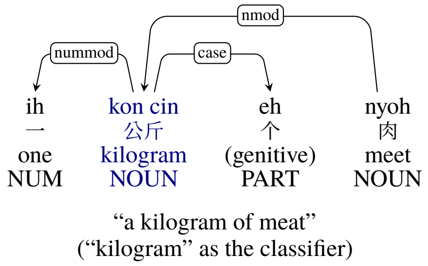


Figure 2: A classifier with genitive 个 *eh*.

If such multi-word expressions are mistakenly segmented during preprocessing, each component is assigned the same part of speech as the full expression and connected using the `goeswith` relation. In the final version of the treebank, they are merged and presented as a single token.

5.2 Linguistically Motivated Guidelines

This section highlights annotation decisions that differ from or are particularly noteworthy relative to the UD for Mandarin.

Nouns. Like Mandarin Chinese, words tagged as `NOUN` include regular nouns, classifiers, temporal nouns, position words, and localisers.

Classifiers can be pre-modified directly by `NUM` and `DET`. They have the feature `NounType=Clf`. In the case of having a numeral or determiner, the classifier is attached to it with a `clf` relation, and the numeral or determiner is then attached to the head noun. However, if the classifier does not come with a numeral or determiner, then the classifier would be the indefinite determiner with the noun as the head.

If there is a genitive 个 *eh* (also transliterated as 呃 or 额) between the classifier and the noun, then the classifier (with the numeral and genitive attached as a phrase) would be a `nmod` dependent of the head noun, as shown in Figure 2.

Temporal nouns, despite typically being the adjunct of verbs, are always tagged as a noun. They would have a `nmod` relation from the verb.

Verbal Polarity. Verbs can be negated by markers such as 勿 *veh* (also 伐, 弗) and 没 *meh*. Negat-

[10] https://universaldependencies.org/zh/

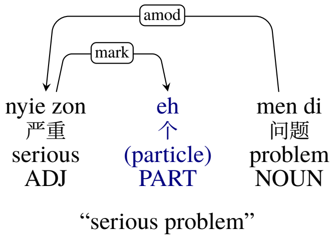


Figure 3: Particle 个 *eh* following an adjective.

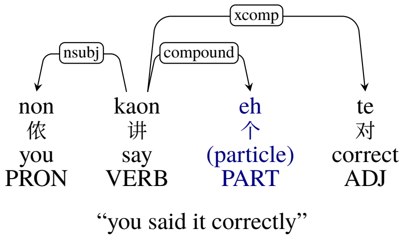


Figure 4: Particle 个 *eh* marking a complement.

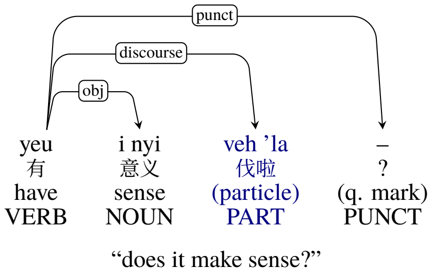


Figure 5: Sentence-final particles.

ors are excluded from lemmatisation, and the token is marked with `Polarity=Neg`.

Particles. The multifunctional particle 个 *eh* (also 呃, 额) corresponds to Mandarin 的, 地, and 得 (*de*), functioning as a genitive, relativise, nominalise, or adverbialiser. It is annotated with the `mark` relation when introducing dependent clauses or modifiers. See Figure 2 for its genitive use and Figure 3 for broader functions.

When 个 *eh* is used in an extent or descriptive construction (corresponding to Mandarin 得 *de*), a `compound` relationship is used, as specified in UD Chinese guidelines (`compound:ext`). In this case, it follows a verb, adjective, or adverb, then followed by a complement part. The complement is treated as a `xcomp` or `ccomp` dependent (depending on its subject) of the configuration's head. The complement serves as the adverb if it is after a verb; it is similar to the latter clause in English's "so… that…" construction if it is after an adjective or adverb. Figure 4 shows an example of this usage.

Another group of particles different from Mandarin Chinese are sentence-final particles. Common ones include 伐 *vah*, 了 *leh* (also 嘞), and 啦 *'la*. Combined use is also very common in Shanghainese, especially rhetorical questions, such as 伐啦 *veh 'la* (Myers, 2015). We treat them as a single token; however, if they are syntactically different (e.g., one indicating the end of a sentence and the other marking the question), then they would still be separated. Sentence-final particles are attached to the sentence via a `discourse` relationship. Figure 5 shows this usage.

	Singular	Plural
1st	吾 *ngu*	阿拉 *ah 'la*
2nd	侬 *non*	㑚 *na* (also 拿)
3rd	渠 *yi* (also 伊)	渠拉 *yi 'la* (also 伊拉)

Table 1: Personal pronouns in Shanghainese.

Pronouns. In Middle/New Period Urban Shanghainese, there are no polite forms of pronouns as in Mandarin. Table 1 shows the personal pronouns. Possessive case of the personal pronouns are constructed by appending the genitive particle 个 *eh*.

There are two demonstrative pronouns in Shanghainese. The proximal demonstrative is 搿 *geh* (also 葛; "this/these" or "here"). The distal demonstrative is 埃 *i* (also 伊; "that/those" or "there"). They also have some derived forms, such as 搿搭 *geh teh* "here", 埃搭 *i teh* "there", 埃面搭 *i mie teh* "there", etc. We treat these words as single tokens.

Other rules. Other syntactic features are almost identical with Mandarin Chinese, and we thus primarily reference to Chinese UD guidelines. For words that do not have an exact correspondence in Mandarin, we consider their Mandarin synonyms with the same POS or structurally similar Mandarin constructions to determine the relationships.

6 Statistics and Pipeline Evaluation

Statistics of the treebank. The current treebank contains 983 sentences with 8,584 tokens. An example in ConLL-U format can be found in Appendix B. More sentences will be annotated in the next UD release.

34 relations and 15 UPOS tags are in the treebank. Among all tokens, 3,356 of them (approximately 40%) are mapped during the annotation using our pipeline, and we have collected 374 pairs of Shanghainese-to-Mandarin lexicon correspondences for 296 Shanghainese words. Appendix C shows detailed statistics on the treebank.

Experiment (stage of manual steps applied)	Segmentation		Parsing	
	Tokens	UPOS	UAS	LAS
Raw Sentence	70.69	52.29	30.48	24.97
+Tokenisation	100.00	61.07	51.84	39.15
+Tok. +Lexicons	100.00	92.66	81.54	73.97

Table 2: Evaluation of the effects of each stage of the pipeline, compared with the golden data (segmentation is evaluated using percentage accuracy). The CoNLL 2017 UD evaluation script[11] is used.

Pipeline evaluation. To evaluate the effectiveness of our pipeline, we assess the automatic parses on Shanghainese, produced by Stanza's Mandarin Chinese parser, at different processing stages using the first 100 sentences. The manually corrected annotations are treated as the golden data. Table 2 shows the results.

Feeding the unsegmented Shanghainese sentences directly to Stanza's Mandarin model (*Raw Sentence*) produces poor tokenisations, which in turn drags down UPOS tagging and dependency parsing. Correcting tokenisations (+*Tokenisation*) slightly improves the results; parsing is still directly on Shanghainese tokens by the Mandarin parser, and accuracies are still low. Replacing Shanghainese word forms with their Mandarin equivalents before parsing (+*Tokenisation* +*Lexicons*) yields a large improvement.

Manual tokenisation and lexical mapping substantially improve the accuracy, but the performance of the Stanza parser on Shanghainese remains considerably low, underscoring the need for a dedicated Shanghainese treebank to support the development of more accurate parsers.

Nonetheless, the pipeline still makes use of existing resources and improves annotation efficiency, especially in the early phase. The average annotation efficiency is around 50 sentences per hour, with almost a quarter of the sentences requiring no manual corrections other than lexical mapping.

7 Model Fine-tuning

We fine-tune a graph-based dependency parser on our annotated treebank using SuPar[12]'s implementation (Zhang et al., 2020) of the biaffine-based dependency parser by Dozat and Manning (2018). The model couples a biaffine scorer over head–dependent hidden states with `XLM-RoBERTa-large`, the pretrained multilingual model, as the contextual encoder.

We trained the model over 50 epochs with AdamW at a base learning rate of 5×10^{-5} with 10% warm-up, 20× scheduled decay, and gradient clip of 5. Batches are formed by sentence length bucketing with a maximum of 5,000 tokens per update and a 20-token fixed positional window. The treebank is randomly split into train, dev, and test with ratio of 80% (786 sentences), 10% (98 sentences), and 10% (99 sentences), respectively.

At the final checkpoint, the model reaches UAS of 75.61 and LAS of 64.91. The fine-tuned parser shows strong capacity in learning a robust representational foundation and significantly outperforms Stanza's Mandarin parser without manual correction. However, its capability in generalisation to unseen data is still limited, possibly due to the small size of training data. Appendix D shows more details of training.

8 Conclusion and Future Work

In this paper, we present the first UD treebank for Shanghainese, named **ShUD**. We propose a scalable annotation pipeline that leverages the strong performance of existing parsers and the substantial syntactic overlap between Shanghainese and Mandarin Chinese. We also fine-tuned a dependency parser and achieved considerable parsing accuracy.

Our parser could serve as a primitive foundation for future annotations and further automate the pipeline to reduce reliance on Mandarin parsers.

The treebank is still limited by its size and genre, with current data source solely based on scripted speech in everyday context. In future work, we plan to extend the treebank, by more annotations and possible expansions to social media text, written literature, news podcasts, etc., to develop more accurate models. Universal morphological features, beyond syntactic dependencies, can also be added to better support downstream tasks.

References

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 1–54.

Marie-Catherine De Marneffe and Joakim Nivre. 2019. Dependency Grammar. *Annual Review of Linguistics*, 5(1):197–218.

[11] https://github.com/ufal/conll2017/blob/master/evaluation_script/conll17_ud_eval.py

[12] https://github.com/yzhangcs/parser

Kaja Dobrovoljc. 2022. Spoken Language Treebanks in Universal Dependencies: an Overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573, Portorož, Slovenia. European Language Resources Association (ELRA).

Timothy Dozat and Christopher D. Manning. 2018. Simpler but More Accurate Semantic Dependency Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Joel Grayson. 2025. Shanghainese Dictionary.

Magic Data. 2021a. ASR-CShhiDiaCSC: A Chinese Shanghai Dialect Conversational Speech Corpus.

Magic Data. 2021b. ASR-SCShhiDiaDuSC: A Scripted Chinese Shanghai Dialect Daily-use Speech Corpus.

Ethan Myers. 2015. Sentence final particles in Shanghainese: Navigating the left periphery. *Open Access Theses*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Wuyun Pan, SF Zhengzhang, RJ You, and Lien Chinfa. 1991. An introduction to the Wu dialects. *Journal of Chinese Linguistics Monograph Series*, (3):235–291.

Rafaël Poiret, Tak-Sum Wong, John Lee, Kim Gerdes, and Herman Leung. 2023. Universal Dependencies for Mandarin Chinese. *Language Resources and Evaluation*, pages 1–38.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Nairong Qian. 1997. 上海话语法 *(Shanghainese Syntax)*. 上海人民出版社 (Shanghai People's Press).

Wiktionary Contributors. 2024. Category:Wu language - Wiktionary, the free dictionary.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John SY 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.

Wu Chinese Society. 2009. 吴音小字典·吴语小词典 (Wu-Language Dictionary / Wu-Language Lexicon).

Yuwei Xie. 2011. Language and Development of City: The Linguistic Triangle of English, Mandarin, and the Shanghai Dialect. *Language*, 1:1–2011.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Koichi Yasuoka. 2019. Universal Dependencies treebank of the four books in Classical Chinese. In *DADH2019: 10th International Conference of Digital Archives and Digital Humanities*, pages 20–28. Digital Archives and Digital Humanities.

Koichi Yasuoka, Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, Shigeki Moro, and Kazunori Fujita. 2022. Designing Universal Dependencies for Classical Chinese and Its Application. *Journal of Information Processing*, 63(2):355–363.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

Yu Zhang, Zhenghua Li, and Zhang Min. 2020. Efficient Second-Order TreeCRF for Neural Dependency Parsing. In *Proceedings of ACL*, pages 3295–3305.

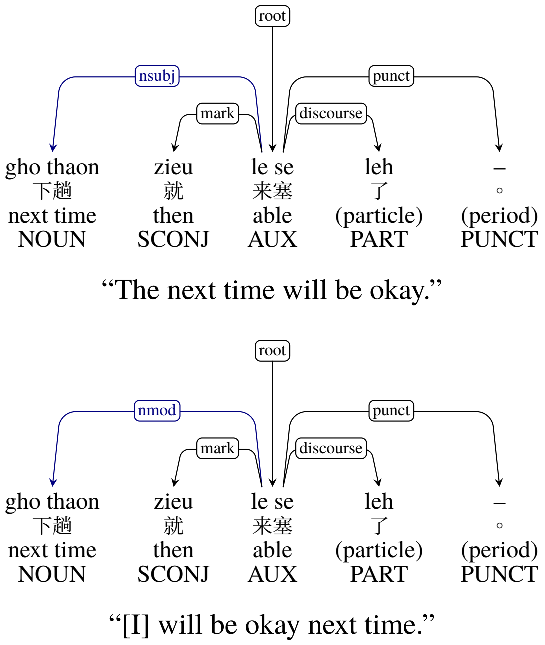
X. Zhu. 2006. *A Grammar of Shanghai Wu*. LINCOM studies in Asian linguistics. LINCOM Europa.

A Ambiguous Example

gho thaon	zieu	le se	leh	–
下趟	就	来塞	了	。
next time	then	able	(particle)	(period)
NOUN	SCONJ	AUX	PART	PUNCT

Figure 6: An example of an ambiguous sentence.

Take the sentence in Figure 6 as an example. The tokenisation and UPOS tagging are unanimous. However, when it comes to understanding the sentence, there could be a deviation. The sentence could be treated as a complete sentence, where the word 下趟 *gho thaon* "next time" is treated as the subject; it would have a `nsubj` dependency with the root. On the other hand, in a colloquial context, it could also be interpreted as a sentence omitting its subject, probably "I"; in this case, the word would be a temporal noun and has a `nmod` dependency with the root. This is shown in Figure 7.



"The next time will be okay."

"[I] will be okay next time."

Figure 7: Two possible interpretations that lead to different dependency labels of the ambiguous sentence.

B Example in ShUD Treebank in ConLL-U Format

Figure 8 shows an example.

C Detailed Treebank Statistics

The treebank spans across sentences with lengths from 2 tokens to 18 tokens, with an average of 8.73 tokens per sentence. Figure 10 shows the distribution of sentence lengths.

34 out of UD's 37 dependency relations are found in the treebank, excluding `expl`, `list`, and `fixed`. Figure 9 illustrates the distribution of all relations in our treebank.

In all 17 UPOS tags, 15 of them are found in the treebank, excluding `X` and `SYM`. Figure 11 shows the distribution of UPOS tags.

D Detailed Fine-tuning Results

The model is fine-tuned for 50 epochs. Figure 12, 13 shows the UAS and LAS, respectively, in each epoch. All of them show large generalisation gaps, with the model fitted almost perfectly on training data. They all show an early stopping at around epoch 15. This suggests a typical overfitting behaviour that the model lacks generalisation improvements, highlighting the need for more training data.

Another observation is that unlabelled metrics are generally around 10%–20% better than labelled metrics. The immediate focus could also be on the parser's relation labelling quality.

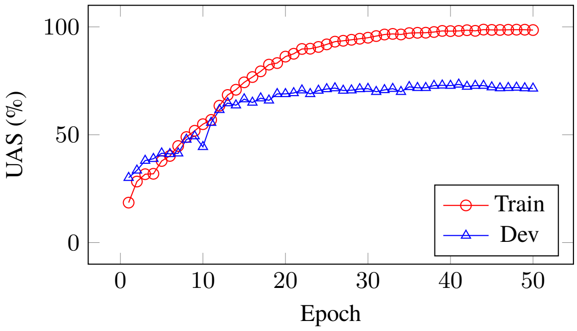


Figure 12: Unlabelled Attachment Score (UAS).

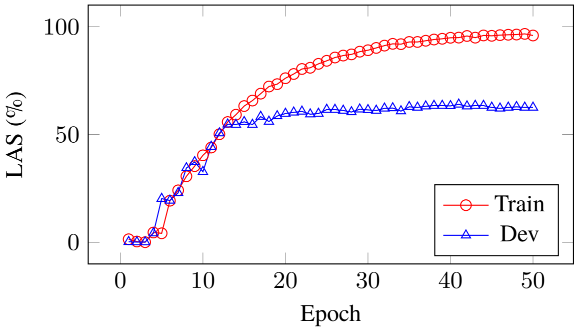


Figure 13: Labelled Attachment Score (LAS).

```
# sent_id = 3
# text = 吾伐会的随便讲吾爱侬呃。
# text_cmn = 我不会随便讲我爱你的。
1  吾      吾      PRON   PRP  Person=1      4  nsubj      _  SpaceAfter=No
2  伐会的  伐会的  AUX    MD   Polarity=Neg  4  aux        _  SpaceAfter=No
3  随便    随便    ADV    RB   _             4  advmod     _  SpaceAfter=No
4  讲      讲      VERB   VV   _             0  root       _  SpaceAfter=No
5  吾      吾      PRON   PRP  Person=1      6  nsubj      _  SpaceAfter=No
6  爱      爱      VERB   VV   _             4  ccomp      _  SpaceAfter=No
7  侬      侬      PRON   PRP  Person=2      6  obj        _  SpaceAfter=No
8  呃      呃      PART   UH   _             4  discourse  _  SpaceAfter=No
9  。      。      PUNCT  .    _             4  punct      _  SpaceAfter=No
```

Figure 8: An example in ShUD treebank in ConLL-U format. Translation: *I won't say I love you casually.*

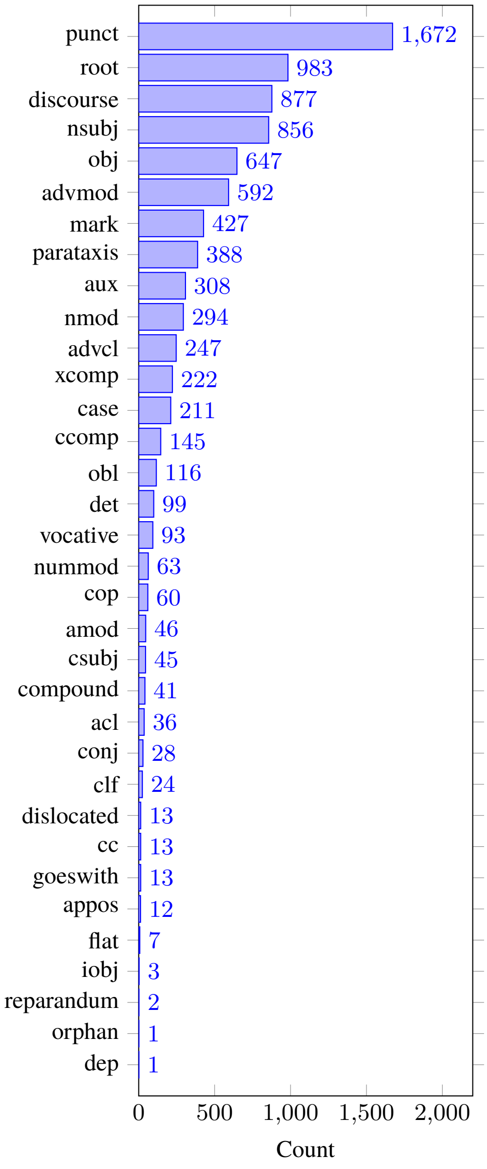


Figure 9: Distribution of relations.

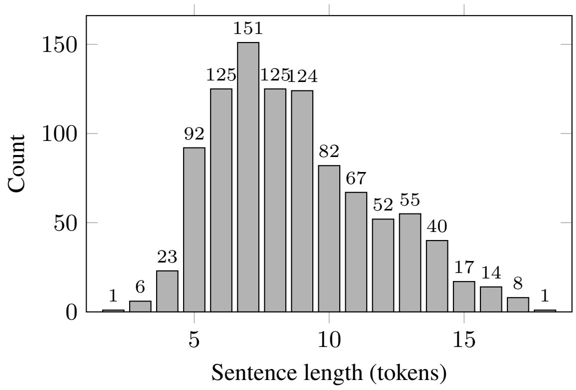


Figure 10: Distribution of sentence lengths.

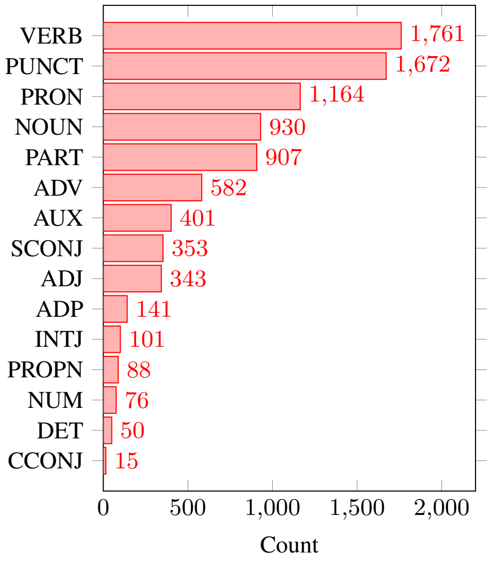


Figure 11: Distribution of UPOS tags.