Design of a mSUD Taiwan Taigi Treebank Aligned on Mandarin and Teochew Translations

Pierre Magistry and **Ilaine Wang**
ERTIM – Inalco
2, rue de Lille
75007 Paris, FRANCE
pierre.magistry@inalco.fr

Chen Siman 陈思漫 **Chen Xinlei** 陈新蕾 **Li Zhongjie** 李仲杰 **Wen Yu-Chieh** 溫侑潔 **Zhang Weiqi** 张炜祺

Abstract

This paper presents the design choices and the first (preliminary) release of a trilingual treebank of Taigi sentences aligned on translations into Mandarin and Teochew. The three languages come with morphosyntactic annotations following the joint morphology and Surface syntactic Universal Dependencies (mSUD) scheme. We provide 54 annotated and validated sentences as a first release for open discussion, with the objective of annotating the full set of 420 sentences from the examples in Taiwan Ministry of Education's Dictionary of Frequently-Used Taiwan Minnan.

Keywords: Treebank, Syntax, Morphology, Taigi, Mandarin, Teochew, mSurface Syntactic Universal Dependency

1 Introduction

This paper presents a syntactic treebank for Taiwan Taigi (hereafter "Taigi") aligned with two other Sinitic languages: Taiwan Mandarin and Teochew. Our motivation is to provide manually annotated data which will allow for linguistic description, cross linguistic comparisons and evaluation data needed to describe more precisely the knowledge encapsulated into various large language models.

We chose to follow the Universal Dependencies (UD) framework to ease alignment and cross linguistic comparisons. More specifically, we adopt the mSUD scheme to provide a description in terms of surface syntax (Guillaume et al., 2024) and a morph-level annotation to tackle the issues related to word segmentation more precisely.

2 Languages of the Treebank

Taigi is a Sinitic language spoken in Taiwan, with a rich history and a complex sociolinguistic situation. Taigi is closely related to other Southern Min languages spoken in the South of China and in numerous diasporan communities in South East Asia, but it also has unique properties resulting from the history of Taiwan, especially language contacts with Austronesian languages and Japanese. It is the most widely used language in Taiwan after Mandarin. After being prohibited until the late 20th century, it is now receiving political support and is undergoing the process of standardization by the Ministry of Education (MOE) in Taiwan.

Our dataset is based on an official release of a set of example sentences contained in the Dictionary edited by the MOE and published as Open Data under a Creative Commons license[1]. The text in Taigi is provided both in sinograms (commonly called *Chinese characters*) and in Tâi-lô romanization. The dictionary also includes translation in Mandarin for every sentence.

The Mandarin in this corpus (hereafter MSM, for Modern Standard Mandarin) corresponds to the official variant of Mandarin used by the Republic of China (ROC, Taiwan). It may differ from MSM in use in China, especially in terms of lexical choices, but also for some syntactic constructions. Taiwanese MSM is more subject to the influence from Taigi and this characteristic is probably slightly amplified by the translation nature of this specific corpus in which the text is originally written in Taigi.

[1] https://sutian.moe.edu.tw/zh-hant/siongkuantsuguan/

Teochew is a language originally spoken in the extreme South East of the Guangdong province in China, next to the Fujian province border. It is thus expected to be quite distant from Taigi but closer than MSM. Teochew is also spoken in many places by diasporan communities and is subject to variation from one community to another. This work is part of a broader project based in Paris where one of those communities lives, allowing us to work closely with Teochew people. Incidentally, this study contributes to a larger endeavor to characterize Teochew varieties. This will enable us to draw a more comprehensive picture, make more detailed comparisons and conduct transfer learning experiments.

According to our own observations of native speakers, Taigi and Teochew nevertheless allow for a certain degree of mutual intelligibility, but requiring noticeable efforts from both sides (and maybe a sense of Sinitic languages phonology). One of the participants of this project is a native speaker of Teochew from China who was in charge of providing translations (see below for more details). In contrast with Taigi, Teochew is not ongoing a standardization process. The choice of scripts is therefore less straightforward in this case, as we will explain in the next section.

2.1 Scripts of the Treebank

The languages of our corpus can be written using a variety of scripts.

sinograms The three languages can be written in sinograms. Following the original dataset, we use traditional variants for Mandarin and the official characters advocated by the MOE for Taigi without any modification of the original data. Selecting sinograms for Teochew is less easy as there is no clear official recommendation. For the sake of consistency we avoid simplified characters and limit our choice to traditional characters. This may be unusual for Teochew in China, but this practice is also attested in diasporan communities, so this choice is not too specific to our treebank. When unsure, we rely on two dictionaries.[2]

romanizations The three languages can also be written using the Latin script. MSM is typically romanized with *hànyú pīnyīn*, but the transcription was not included in the data released by the MOE. Regarding Teochew, the *Guangdong Peng'im* has been designed in 1960 by the Guangdong province government, and later slightly modified in the diasporan communities and turn into *Gaginang Peng'im* which is more suited for US keyboards. Converting between the two systems is straightforward using formal grammars[3] so we plan to release the treebank in both versions. Examples given in this paper are romanized with Guangdong Peng'im. Different romanizations are possible for Taigi. The most widely used are the historical *Pehoe-ji* and the now official Tâi-lô. The use of the latter is advocated for by Taiwan's Ministry of Education and the dataset we use includes this romanization. Although conversion between the two system can also be achieved fully automatically if needed, we chose to follow the recommendation and use Tâi-lô.

2.2 Organization of the Work

This work is part of a project which aims to study the limitations of current methods and tools in Natural Language Processing (NLP) when used for digital humanities research on Sinitic language resources. For this purpose, we first need resources covering the three axes of variation we wish to study: diachrony, diatopy, and grapholinguistics.

This specific part on the treebanks was conducted by a team of 5 MA students in NLP, advised by two researchers. The students have a variety of linguistic background which greatly contributed to the success of this work. All the students are proficient in MSM (either Taiwan Mandarin or China *putonghua*). One was also a native speaker of Taigi from Kaohsiung, another a speaker of Amoy Hokkien, and another a native speaker of Teochew from China. The two researcher have experience in Sinitic languages processing and linguistics. One has a decent command of Mandarin and knowledge

[2] http://www.czyzd.com/ and https://play.google.com/store/apps/details?id=com.tcknow.whattcsay3&hl=en_AU&gl=US

[3] A conversion tool was already available here: https://github.com/learn-teochew/parsetc

of Taigi grammar, while the other is an heritage speaker of the Teochew diaspora.

Students worked in pairs, issues were discussed during regular meeting and all the final trees are reviewed by one researcher.

For the Teochew part, the native speaker was asked to do the translation from the MSM version to limit the influence of Taigi in her lexical and grammatical choices.

3 Related Works

Despite increasing efforts to build corpora for Taigi, recent works focus more on large amount of raw data to train language models and speech corpora (Liao et al., 2020). Resources with morphosyntactic annotations for Taigi are still very scarce. Noticeable exceptions include the early T3 corpus (Chou, 2006) which was not publicly released, and an attempt to build a Part of Speech (POS) tagged corpus semi-automatically, based on Mandarin resources (Iunn et al., 2009). Tsay (2007) describes what appears to be the only corpus manually annotated in morphosyntax, but it is from a very different genre (spontaneous child speech) and includes only POS tagging (no dependency syntax). To our knowledge, our treebank is thus the first of its kind.

On the other hand, various syntactic treebanks for MSM are available. The UD website contains no less than 6 different corpora for MSM. Two of them come with a mSUD version: Chinese Beginner[4] and Chinese PatentChar[5]. Our work starts from their guidelines and we do our best to stay compatible in terms of linguistic descriptions. Another important related treebank is the UD conversion of the Sinica Treebank (Hsieh et al., 2022), which is closer to the MSM in our corpus, but it only uses the UD scheme and was converted from the original Sinica Treebank (Huang et al., 2000) with a rule-based system that introduces a small amount of errors.

Universal Dependencies (Nivre et al., 2020) is a well known project which aims at building multilingual treebanks following consistent guidelines across different languages, to ease NLP and linguistic comparison. The Surface-Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018) is an annotation scheme which follows distributional criteria closer to more traditional dependency syntax theories than UD (favoring syntactic rather than semantic heads). It still aims at being fully compatible with UD guidelines through automatic conversion using graph rewriting rules.

The mSUD scheme (Guillaume et al., 2024) is an extension to SUD allowing for the joint annotation of syntax and morphology in the same formalism by using morphs as leaves of the dependency trees. It has been used for MSM in Li et al. (2019). Similar discussions can be found in the literature of Chinese processing advocating for character-level parsing (independently from the UD project) such as Zhang et al. (2014, 2013). Conversion from mSUD to word-level SUD is also straightforward with graph-rewriting tools.

Our work also relies on the tooling offered by the UD and SUD community, namely Arborator-Grew (Guibon et al., 2020) for online annotation and Grew (Guillaume, 2021) for graph queries and transformations.

4 Annotation Layers

4.1 Word Segmentation

Word Segmentation is an expected issue when addressing textual annotation for Sinitic languages. A typical treebanking workflow for such languages written in *scriptio continua* is to start with defining segmentation guidelines as a prerequisite before starting the syntactic analysis. But word segmentation decisions and syntactic analysis are closely related, and it is more convenient to address both jointly. We do so by adopting the mSUD scheme which allows us to somehow delay the word segmentation analysis while conducting the full syntactic analysis of each sentence.

Another helpful feature of this Taigi corpus is that it comes with a Tâi-lô romanization. Taigi has a long history of digraphia, being written either in sinograms for centuries or romanized since the end of the 19th century. The use of the Latin script has long in-

[4] https://universaldependencies.org/treebanks/zh_beginner/index.html

[5] https://universaldependencies.org/treebanks/zh_patentchar/

troduced the use of a non *scriptio continua* and a word-level writing. This script used to be more widely adopted and has been an actual writing system for publication (contrary to hànyú pīnyīn and 注音符號 zhùyīn fúhào for Taiwan MSM which are only used as phonetic transcription, mostly for educational purposes). As a result, we can rely on the Tâi-lô text to provide a first good approximation of a word-level tokenization.

Just like any language written with the Latin script and spaces, the correspondence between orthographic words and syntactic units is not perfect – typically when dealing with multiwords expressions.

Segmentation is discussed in the appendix of the guidelines for official Taigi romanization followed by the editors of the dictionary[6], especially some ambiguous cases where different segmentations correspond to different meanings. Such examples include *âng-hue* vs. *âng hue* (sinograms: 紅花, lit. red+flower), which correspond respectively to a medicine plant name (non compositional meaning) or to a red flower. In our treebank this would result in different morphosyntactic relations, resp. `/m` or `mod`. The same goes for *oo niau* vs. *oo-niau*, (sinograms 烏貓) resp. a black cat or a fashionable girl. Longer frozen expressions also exist such as *tshenn-mê-gû* 青盲牛 for illiterate, non compositionnal meaning from *thsenn-mê* *'blind'* and *gû 'cow'*[7].

Such ambiguities can also concern other relations and affect parts of speech, for example 風吹 (lit. *wind+blowing*) may mean *the wind blows* with a `subj` relation, in which case it is written *hong tshue*, or it may mean *a kite* (non compositional meaning) in which case it is written *hong-tshue* and annotated with a `/m` relation.

In the case of our Taigi treebank, specific structures such as Verb-Object constructions can lead us to introduce some discrepancies with the Tâi-lô version as illustrated in Figure 1. This confirms the benefits of using the mSUD scheme even when annotating corpora in Latin script.

4.2 Part of Speech Tags

One major difficulty to tackle when building a new UD treebank is to follow their standard POS tagset. For each treebank, the annotation guidelines must detail how the limited and fixed tagset of only 17 tags[8] was adapted.

In our case, the most striking example of such difficulties is the lack of a specific tag for nominal classifiers. Therefore, we decided to follow the common practice for MSM treebanks to annotate classifiers as a nominal elements (`NOUN` tag) and to mark the classifier function as a `clf` syntactic relation with the head noun. Another issue is the question of the adjectives and the distinction between stative and action verbs. As we only have a single `VERB` tag for all subtypes of verbs, we use the `ADJ` tag for predicative adjectives and keep `VERB` for action verbs.

More generally, we adopt the following strategy:

1. We first try to follow the same guidelines as the MSM corpora annotated in mSUD[9]. Most of the discussions are relevant for Sinitic languages with only little adaptation for Taigi.

2. We discuss unclear cases during meetings or through github issues.

3. We check categories used in the MOE's Dictionary of Frequently-Used Taiwan Taigi[10]

An interesting issue arises from the differences of tagsets between UD and the MOE dictionary. The latter adopts the categories from traditional Chinese lexicography. We did not try to define a systematic mapping between the two tagsets as we expect difficult cases of ambiguities. However, it is interesting to observe the correspondences resulting from our annotation afterwards.

We project the categories from the dictionary on our corpus by assigning the list of possible traditional categories to the `xpos`

[6] https://language.moe.gov.tw/001/Upload/FileUpload/3677-15601/Documents/tshiutsheh_1081017.pdf p.55

[7] in such case syntactic structure may change, a *'blind cow'* would more likely be translated as *tshenn-mê ê gû* 青盲的牛.

[8] https://universaldependencies.org/u/pos/all.html

[9] https://guidelines.surfacesyntacticud.org/docs/language/mandarin_chinese/

[10] https://sutian.moe.edu.tw/zh-hant/

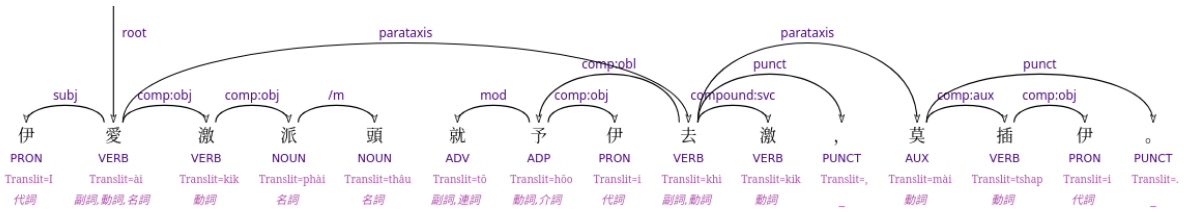

Figure 1: *He likes to put on airs, just let him put on airs and ignore him.*
Example of discrepancy between syntactic units and orthographic word segmentation. Here *kik-phài-thâu* is written in a single word (meaning *to put on airs*, lit. *to arouse, stimulate* (kik) *a dignified air* (phài-thâu), but despite its non compositional meaning, the Verb-Object structure is clear and required to explain the repetition of *kik* without *phài-thâu*.

property of each tokens. In order to do so, we build the words from the graph by extracting subtrees connected by morphological (`/m`) relations and try to match the sinograms and pronunciation to a dictionary entry (See Figure 1 for an example of a fully annotated sentence). Since the corpus is based on dictionary examples, the coverage is almost perfect. On the other hand, we do not try to resolve the ambiguities and simply record the list of possible categories for each word. Subword tokens are assigned the `xpos` of the word they appear into.

This allows us to draw the following Sankey diagrams. Such illustration is helpful not only to understand and document our annotation choices, but also to compare traditional categories and UD tagset and for error mining, combined with the power of ArboratorGrew query language.

Figure 2 is based on the whole corpus and contains both clear cases of good correspondences at the bottom, such as 代詞 – `PRON`, 副詞 – `ADV`, 動詞 – `VERB`, 名詞 – `NOUN`, 形容詞 – `ADJ`, 助詞 – `PART` and more complex cases which deserve more discussion are shown.

Figure 3 is based on the same data, with the aforementioned correspondences removed to focus on the less obvious cases.

The following observations are examples of what we can infer from the two figures:

- for Sinitic languages, we expect a high level of lexical ambiguity as 0-derivation is often possible to change the POS of a word, but we also see that a large part of the vocabulary is actually very stable;
- the idioms (熟語) in the dictionary are a subclass of verbal expressions;
- position words (方位詞) are split into a subclass of nouns and a subclass of adverbs (with possible lexical ambiguity);
- prepositions (介詞) are split between actual prepositions (`ADP`) and `VERB`s. This is to be expected as 介詞 are usually grammaticalized verbs, resulting in lexical ambiguity;
- as explained earlier, we treat classifiers (量詞) as a special case of nominal items and tag them as `NOUN`;
- temporal expressions (時間詞) are mostly nominal expressions, aside from an occasional adverbial usage;
- 助詞, some kind of particles seems to cover not only what we tag `PART`, but also encompass different kinds of items. It is likely to be more related to actual functions of the word than it is to POS which definitely deserves more investigation;
- our class of `AUX` (auxiliary) is clearly a subclass of verbs, but with many lexical ambiguities. Some are also classified as 副詞 (closer to adverbs) in traditional analysis, but we consider auxiliaries as syntactic heads, so the `ADV` tag would not be appropriate. In UD, `AUX` must be a closed list of items, so further investigation is required to strictly define this list based on our preliminary annotations.
- the 代詞 category is often translated as "pronoun", but actually includes determiners.

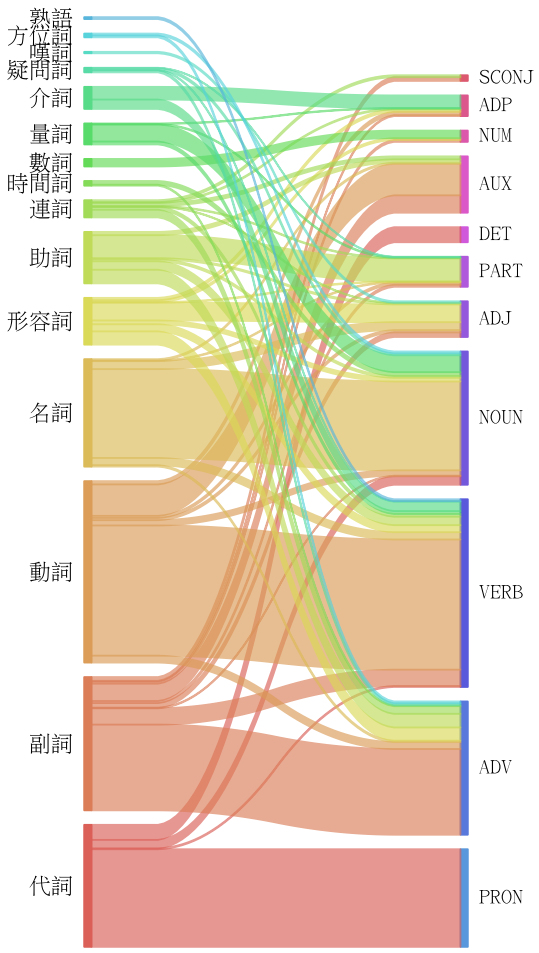

Figure 2: Sankey diagram to show correspondences and mismatches between the (S)UD POS tagset and categories from the MOE Dictionary.

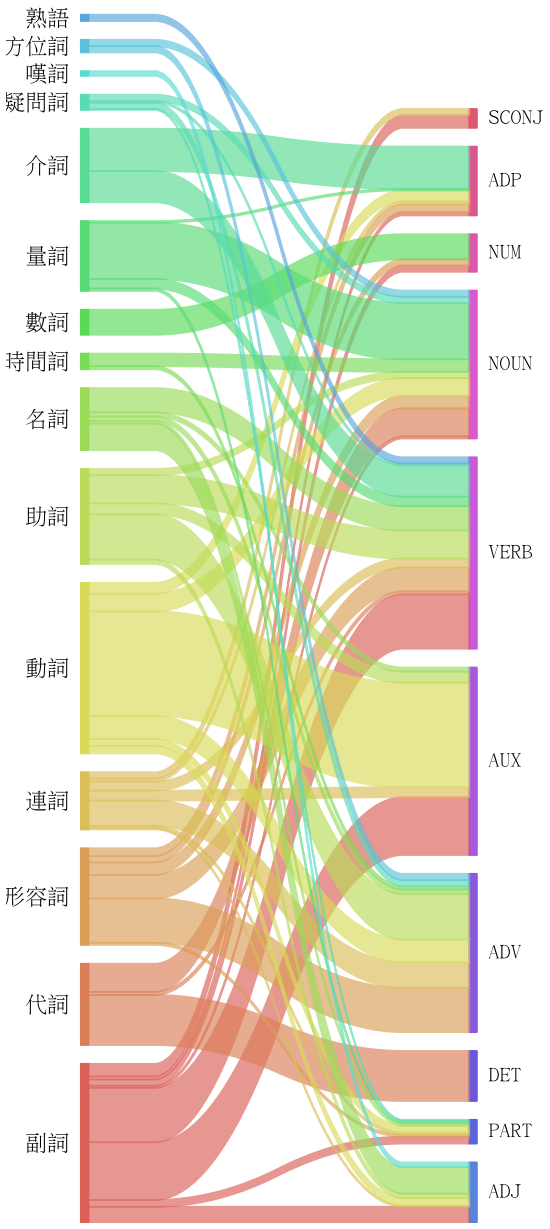

Figure 3: Sankey diagram to show correspondences and mismatches between the (S)UD POS tagset and categories from the MOE Dictionary, with naively expected mapping removed.

4.3 Syntactic Relations

We used the following syntactic relation (sorted by frequency):

mod for modifiers
punct for punctuation marks
comp:obj for verb – (direct) object relations
subj for verb – subjects relations
comp:aux for complements of auxiliary verbs
parataxis between heads of clauses without explicit syntactic or discourse connector
compl:obl for oblique (indirect) objects
discourse:sp for sentence final particles
det for noun phrase determiners
comp:res for resultative constructions
comp:pred for predicative constructions with 是
clf for classifiers
comp:svc for serial verbs constructions
comp:dir for directional complements
cc for coordinations
conj for conjunctions
conj:redup for reduplications
comp:periph for peripheral complements (such as topicalized arguments)
aspect for aspect markers
subj:periph for peripheral subjects
dislocated

5 Current status

5.1 Statistics

At the time of writing, about 100 Taigi sentences have been annotated and 54 went through the double-check validation process.[11]

These 54 sentences contain 609 tokens (sinograms) and 487 words (obtained by merging morphological subtrees) in Taigi. For Mandarin, we have 649 tokens and 540 words. As for Teochew, 38 sentences have been translated for now, with a total of 403 tokens and 347 words.

5.2 Examples

Here are some examples of annotated sentences from our treebank.

Figure 4 shows the syntactic tree of the sentence 句話盤來盤去 *gu3 uê7 tuang5 lai5 tuang5 ke3*。in Teochew. The syntactic structures of that sentence in the three languages are identical except for the nominal subject: while we have 句話 in Teochew, the numeral appears both in MSM (一句話傳來傳去。) and in Taigi (一句話盤來盤去。*Tsit kù uē puânn lâi puânn khì.*). This example highlights the fact that in Teochew, classifiers can have a determinative function thus making the NUMeral optional.

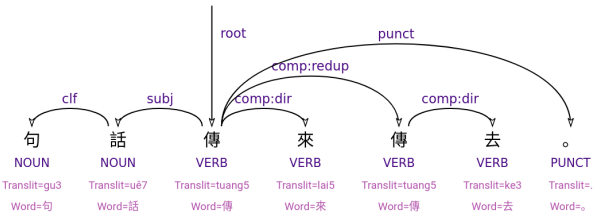

Figure 4: *A word is passed around* (Teochew)

In Figure 5, we can see that while Teochew uses a serial verb construction (and so does MSM), the co-verb 共 *kā* is completely grammaticalized in Taigi.

In our last example in Figure 6, we can see that the trees are similar but there are some lexical differences and the relations are also a little different, due to the nature of the root word: an ADJ for Taigi and MSM, but an AUX for Teochew.

6 Conclusion

The parallel treebank presented in this paper is still at an early stage of compilation. However, rather than waiting to have the whole set of sentences fully annotated, we decide to do an early release of the data on a gitlab repository . We believe it can already be an open place for discussions through the system of *issues* and pull requests provided by gitlab. We hope it becomes the place of lively exchange about Taigi grammar, mSUD annotations.

Acknowledgments

This work is supported by the French National Research Agency as part of the DiLSi-HN ANR project (ANR-23-CE38-0004-01).

[11] As this is a work in progress, these figures will be updated between the submission and the conference.

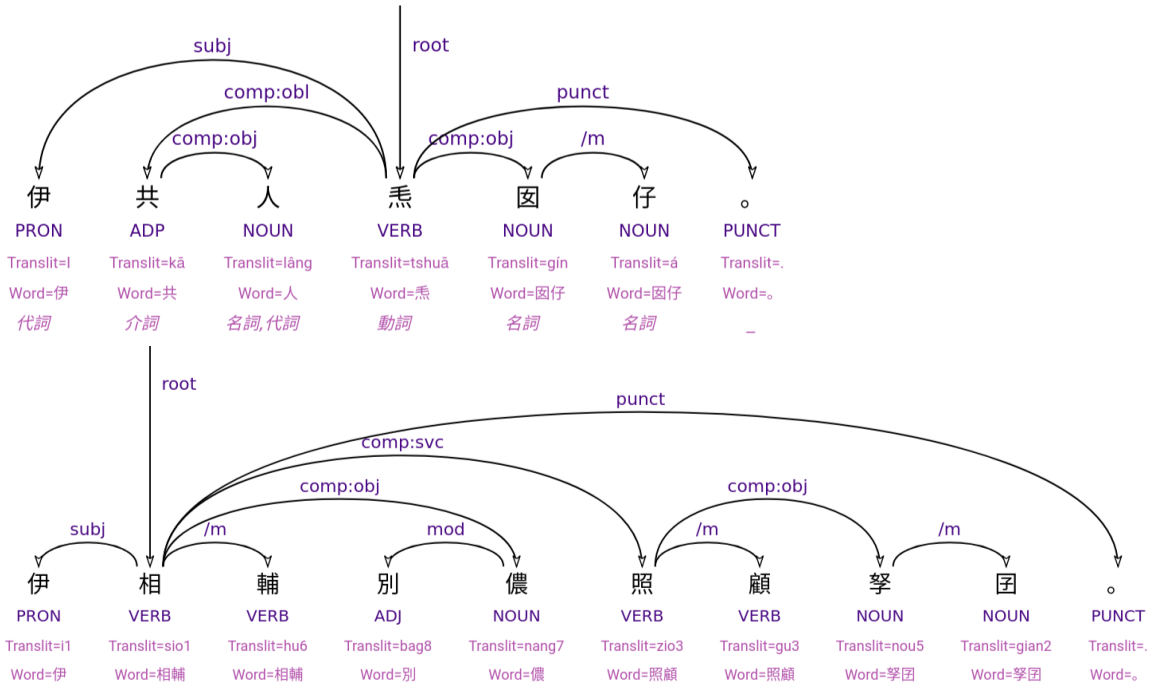

Figure 5: *He takes care of other people's children* (Taigi, Teochew)

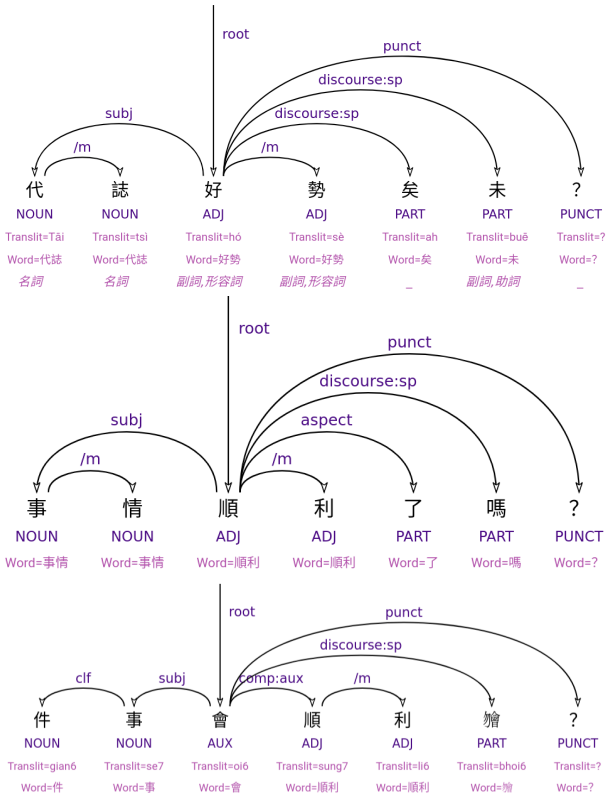

Figure 6: *Did things go smoothly ?* (Taigi, MSM, Teochew)

References

S. Y. Chou. 2006. T3 taiwanese treebank and brill part-of-speech tagger. Master's thesis, National Tsing Hua University.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. 2024. Joint annotation of morphology and syntax in dependency treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and*

Evaluation (LREC-COLING 2024), pages 9568–9577, Torino, Italia. ELRA and ICCL.

Yu-Ming Hsieh, Yueh-Yin Shih, and Wei-Yun Ma. 2022. Converting the Sinica Treebank of Mandarin Chinese to Universal Dependencies. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 23–30, Marseille, France. European Language Resources Association.

Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-ming Gao, and Kuang-Yu Chen. 2000. Sinica Treebank: Design criteria, annotation guidelines, and on-line interface. In *Second Chinese Language Processing Workshop*, pages 29–37, Hong Kong, China. Association for Computational Linguistics.

Un-Gian Iunn, Jia-hung Tai, Kiat-Gak Lau, Cheng-yan Kao, and Keh-jiann Chen. 2009. Modeling Taiwanese POS tagging using statistical methods and Mandarin training data. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 3, September 2009*.

Yixuan Li, Gerdes Kim, and Dong Chuanming. 2019. Character-level annotation for Chinese surface-syntactic Universal Dependencies. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 216–226, Paris, France. Association for Computational Linguistics.

Yuan-Fu Liao, Chia-Yu Chang, Hak-Khiam Tiun, Huang-Lan Su, Hui-Lu Khoo, Jane S. Tsay, Le-Kun Tan, Peter Kang, Tsun-guan Thiann, Un-Gian Iunn, Jyh-Her Yang, and Chih-Neng Liang. 2020. Formosa speech recognition challenge 2020 and taiwanese across taiwan corpus. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 65–70.

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.

Jane S. Tsay. 2007. Construction and automatization of a minnan child speech corpus with some research findings. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 4, December 2007: Special Issue on Speech and Language Processing for Taiwanese Minnan, Hakka, and Mandarin*, pages 411–442.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2013. Chinese parsing exploiting characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 125–134, Sofia, Bulgaria. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Character-level Chinese dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1326–1336, Baltimore, Maryland. Association for Computational Linguistics.