UNIVERSAL DEPENDENCIES TREEBANK OF THE FOUR BOOKS IN CLASSICAL CHINESE

Koichi YASUOKA
Ph. D / Professor of Digital Humanities
Institute for Research in Humanities, Kyoto University
yasuoka@kanji.zinbun.kyoto-u.ac.jp
Kyoto 606-8265 JAPAN

ABSTRACT

Classical Chinese is an isolating language without notational inflection, and its texts are continuous strings of Chinese characters without spaces or punctuations between words or sentences. In order to apply Universal Dependencies for classical Chinese, we need several "not-universal" treatments and enhancements. In this paper such treatments and enhancements are revealed.

KEYWORDS

Classical Chinese treebank; Part-Of-Speech tagging; dependency parsing; sentence segmentation.

1. INTRODUCTION

On May 15, 2019, the author and his colleagues released UD_Classical_Chinese-Kyoto treebank as a part of Universal Dependencies 2.4 (Nivre et al., 2019). The treebank consists of full texts of the Four Books (孟子, 論語, 大學, and 中庸, taken from Kanseki Repository (Wittern, 2016)) with POS (Part-Of-Speech) tags and manually-annotated dependency relations. Classical Chinese is quite different from modern Chinese, thus different approach was needed to develop the classical Chinese treebank. In this paper the author briefly mentions Universal Dependencies for classical Chinese, especially "not-universal" treatments and enhancements on tokenisation, POS-tagging, lemmatisation, morphological features, dependency relations, and sentence segmentation.

2. DEVELOPING UNIVERSAL DEPENDENCIES FOR CLASSICAL CHINESE

The UD treebanks are stored in UTF-8 text files under CoNLL-U format, in which every word is represented by a line containing the following tab-separated fields:

1. ID: Word index, integer starting at 1 for each new sentence.
2. FORM: Word form or punctuations symbol.
3. LEMMA: Lemma or stem of word form.
4. UPOS: Universal POS-tag.
5. XPOS: Language-specific POS-tag.
6. FEATS: List of morphological features from the universal feature inventory or from a defined language-specific extension.
7. HEAD: Head of the current word, which is either a value of ID or zero (0).
8. DEPREL: Universal dependency relation to the HEAD ("root" iff HEAD = 0) or a defined language-specific subtype of one.
9. DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.
10. MISC: Any other annotation.

Our UD_Classical_Chinese-Kyoto treebank uses nine of the ten fields shown above: ID (see Sections 2.1 and 2.6), FORM (see Section 2.1), LEMMA (see Section 2.3), UPOS (see Section 2.2), XPOS (see Section 2.1), FEATS (see Section 2.4), HEAD (see Sections 2.5 and 2.6), DEPREL (see Section 2.5), and MISC (see Section 2.3). We do not use DEPS field (filled by underscore).

2.1 Tokenisation

Classical Chinese texts do not have any spaces or punctuations between words or between sentences. They consist of continuous strings of Chinese characters from the start to the end. The analysis of classical Chinese texts has to begin with finding out word-boundaries, i.e. tokenisation. Morioka et al. (2013) and Yasuoka et al. (2018) have developed MeCab-Kanbun, a powerful (dictionary-based) tokeniser and POS-tagger for classical Chinese texts. We utilise MeCab-Kanbun to tokenise our texts, and then we check them while we annotate dependency relations manually. After tokenisation, we tentatively fill ID field in sequence from 1 to the number of words in each paragraph, and fill FORM field with tokenised words. We also use the four-level word-class system of MeCab-Kanbun to fill XPOS field of each word.

2.2 POS-Tagging

The predicate-object-final structure of very early Chinese texts had only three categories of words: predicate, object, and final. Here in our linguistic model we tentatively call them "verb" "noun" and "particle" (instead "v" "n" and "p" of MeCab-Kanbun) respectively. Several words were specialised to be used as verbs, several as nouns, but most of them had been used in two or three categories around Zhou (周) dynasty.

At that era, we can observe very early modifier usages of verbs. Several verbs were specialised to be used as adverbial modifiers, afterwards caused adverbs. In between verbs and adverbs, auxiliary verbs were almost specialised to auxiliary uses, but incidentally used as verbs. Adjective usages of verbs were not specialised as adjectives at that era, on the other hand, some caused prepositions.

For POS-tagging of classical Chinese texts in UD, we use "VERB" "ADV" "AUX" "ADP" and "SCONJ" to fill UPOS field of each verb-origin word, following the overview of modifier usages mentioned above. For noun-origin words we use "NOUN" "PROPN" "PRON" "NUM" and "ADV" (noun-origin adverbs including 何), categorising them in rather nowadays point of view. For particle-origin words we use "PART" "CCONJ" and "INTJ", keeping up with the guideline of UD v2. We do not use "ADJ" "DET" "SYM" "PUNCT" or "X" for UD_Classical_Chinese-Kyoto in UD 2.4.

2.3 Lemmatisation and Gloss

In classical Chinese texts we often observe variants, such as 青 and 靑. Though 青 and 靑 have the same meaning *'blue'* and have the same pronunciation, their codepoints in UTF-8 are different. We choose one of the variants as the canonical character to lemmatise variants, and fill LEMMA field with the canonical character. For 青 and 靑, based on Pulleyblank (1991), we fill LEMMA field with 青, and we fill MISC field with "Gloss=blue" to indicate their approximate meaning.

2.4 Morphological Feature

We use 15 types of morphological features to fill FEATS field. Among them 11 features are universal and follow the guideline of UD v2: "Aspect" "Case" "Degree" "Mood" "Person" "Polarity" "PronType" "Reflex" "Tense" "VerbForm" and "Voice". The other four features are language-specific: "AdvType" "NameType" "NounType" and "VerbType". We use AdvType=Cau|Deg|Tim to annotate adverbs of cause, degree, or time, respectively. We use NameType=Geo|Nat|Sur|Giv|Prs to annotate proper nouns of geographical, nationality, surnames, given names, or other kind of personal names, respectively. We use NounType=Class to annotate classifier nouns. We use VerbType=Cop VerbForm=Conv or VerbForm=Part to annotate copula, adverbial, or adjective usage of verbs, respectively, though they have no notational inflection.

Table 1. Dependency relations used in classical Chinese treebank

	Nominals	**Clauses**	**Modifier Words**	**Function Words**
Core arguments	nsubj ↪nsubj:pass obj iobj	csubj ccomp xcomp		
Non-core dependents	obl ↪obl:tmod ↪obl:lmod vocative expl dislocated	advcl	advmod discourse ↪discourse:sp	aux cop mark
Nominal dependents	nmod nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat ↪flat:vv compound ↪compound:redup	list parataxis	orphan	root

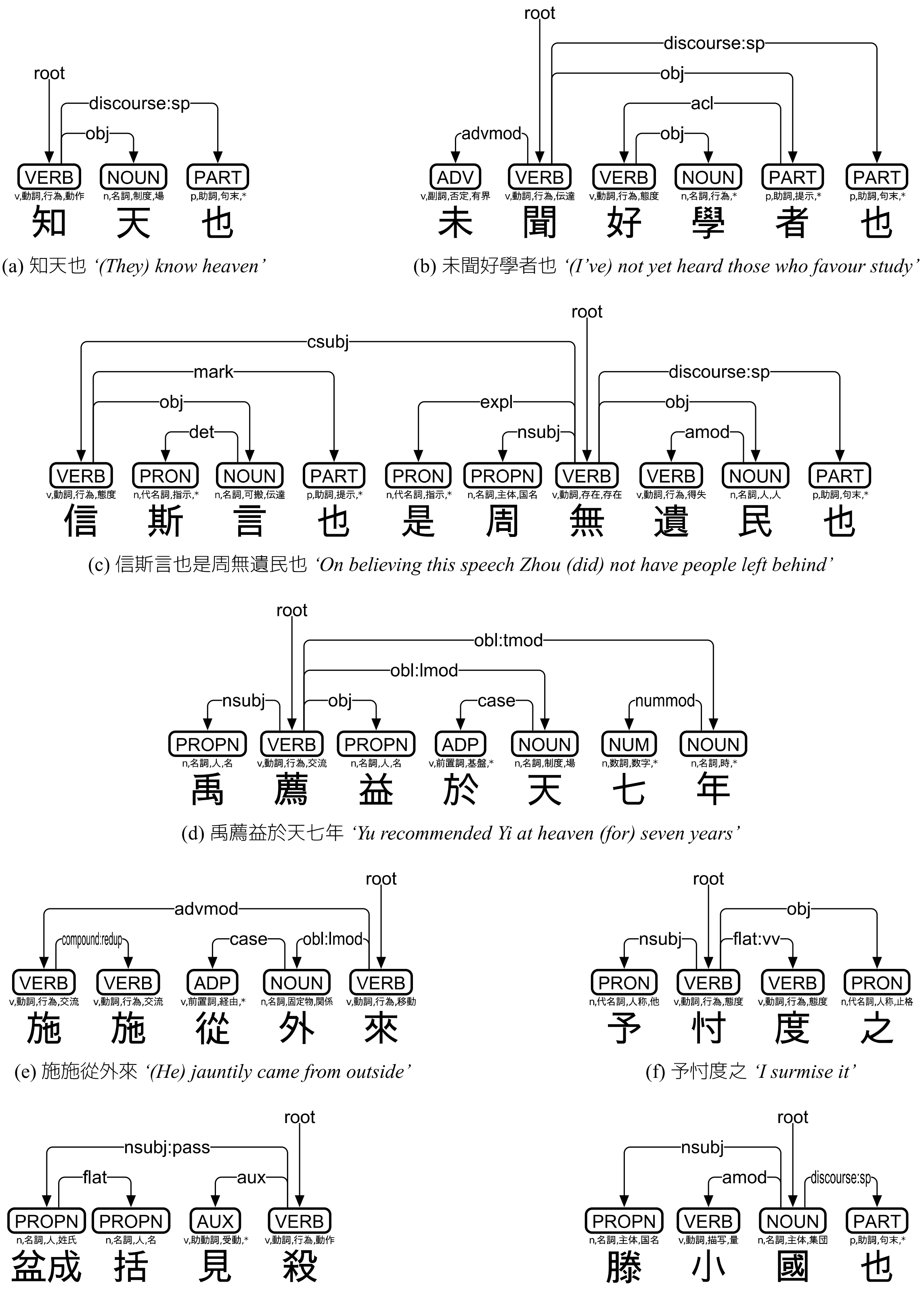
2.5 Dependency Relation

We manually annotated dependency relations on our classical Chinese texts, filling DEPREL field with 38 tags shown in Table 1. Six tags are language-specific: discourse:sp obl:tmod obl:lmod compound:redup flat:vv and nsubj:pass, and the other 32 tags are from UD v2. We tentatively fill each HEAD field with ID (1 to the number of words in each paragraph) of its head word, or with 0 when its DEPREL is root. We allow multiple roots in a paragraph, thus multiple 0's in HEAD field may occur. We keep our treebank planar and projective, as a result any links may not cross one another.

We borrowed discourse:sp from UD_Cantonese-HK treebank (Leung et al., 2016) to annotate final sentence particles in the predicate-object-final structure. For a simple sentence 知天也 (Figure 1(a)) we annotate discourse:sp for 也, which is a final sentence particle. For a sentence 未聞好學者也 (Figure 1(b)) we annotate discourse:sp for 也 also. Though 好學者 '*those who favour study*' forms a predicate-object-final structure, we do not annotate discourse:sp for 者, since we regard 者 as a noun-like particle (not a sentence particle) here. For a complex sentence 信斯言也是周無遺民也 (Figure 1(c)), which includes two 也's, we annotate discourse:sp for the latter 也 and mark for the former 也. The predicate-object-final structure 信斯言也 '*believe this speech*' constitutes a subject clause (annotated by csubj) in the sentence, therefore we treat 也 of 信斯言也 as a marker particle (not a sentence particle).

We use obl:tmod and obl:lmod to annotate temporal and locational oblique nominals, respectively. For an example sentence 禹薦益於天七年 (Figure 1(d)), we annotate obl:tmod for 年 and obl:lmod for 天. We use compound:redup and flat:vv to annotate serial verb constructions, compound:redup for reduplicated compounds (Figure 1(e)) and flat:vv for other kinds of serial verbs (Figure 1(f)). We use nsubj:pass to annotate passive nominal subjects (Figure 1(g)). Additionally, we use case to annotate verb-origin prepositions on the links to them from nouns (Figure 1(d)(e)), following that Nivre (2015) treats adpositions as dependents of nouns.

In a copula sentence of classical Chinese texts, we regard the final nominal is the predicate and the first nominal is the subject, keeping up with the guideline of UD v2. For example, in the copula sentence 滕小國也 (Figure 1(h)), we regard 國 '*country*' is the predicate and 滕 '*Teng*' is the subject, then we link to 滕 from 國 by nsubj. However, the guideline is rather complicated when the predicate consists of a clause. For example, in the copula sentence 信斯言也是周無遺民也 (Figure 1(c)), the clause 信斯言也 is the subject, 是 is the



(a) 知天也 *‘(They) know heaven’*

(b) 未聞好學者也 *‘(I’ve) not yet heard those who favour study’*

(c) 信斯言也是周無遺民也 *‘On believing this speech Zhou (did) not have people left behind’*

(d) 禹薦益於天七年 *‘Yu recommended Yi at heaven (for) seven years’*

(e) 施施從外來 *‘(He) jauntily came from outside’*

(f) 予忖度之 *‘I surmise it’*

(g) 盆成括見殺 *‘Pencheng Kuo was killed’*

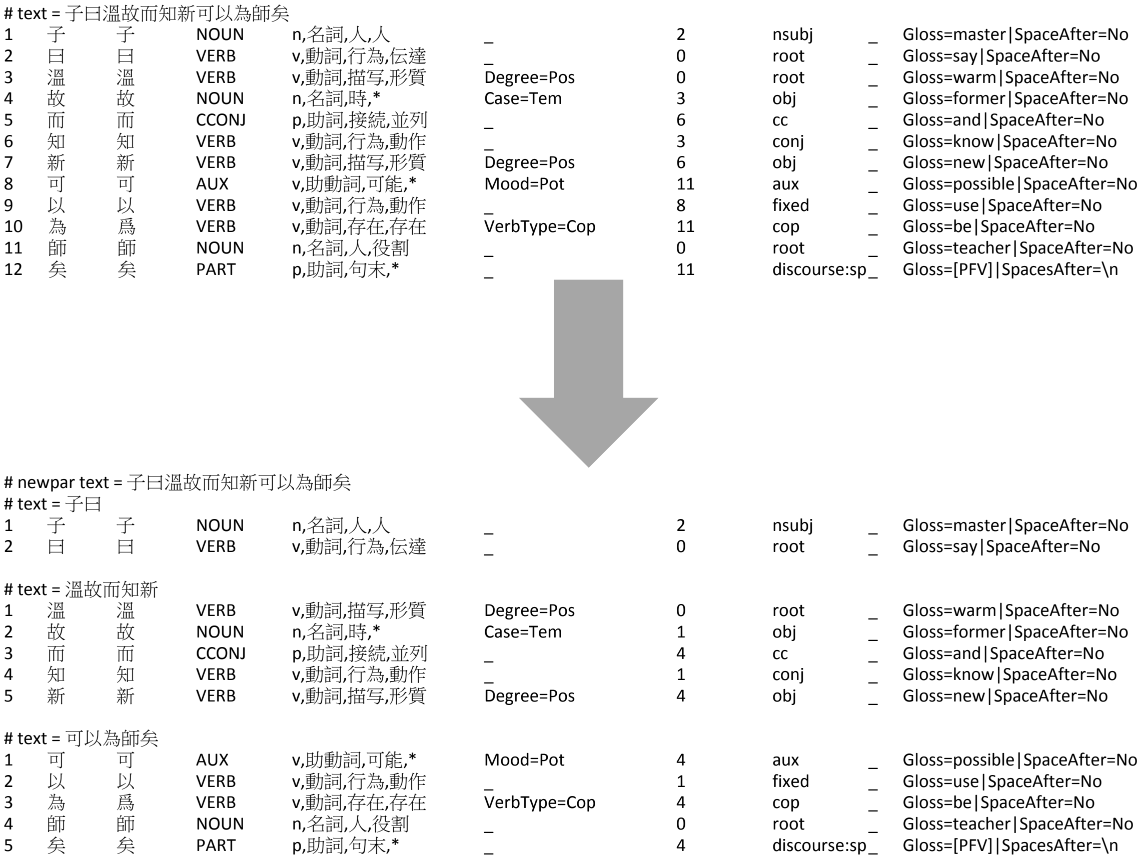
(h) 滕小國也 *‘Teng (is a) small country’*

Figure 1. Example sentences with manually-annotated dependency relations

expletive subject, and the clause 周無遺民 is the “predicate”. Then we link to 信 (the head word of 信斯言也) from 無 (the head word of 周無遺民) by csubj. On the other hand, inside of 周無遺民, 周 is the subject of 無. Then we link to 周 from 無 by nsubj. As a result two subject links (csubj and nsubj) and one expletive link are from 無.

2.6 Sentence Segmentation

Although we allow multiple roots in our classical Chinese treebank, the guideline of UD v2 does not. Thus, for the release of UD 2.4, we divide each paragraph into sentences. We renumber ID and HEAD, so that each sentence includes one and only root. Figure 2 shows a short paragraph 子曰溫故而知新可以為師矣 segmented into three sentences.



```
# text = 子曰溫故而知新可以為師矣
1    子   子   NOUN    n,名詞,人,人          _              2    nsubj          _   Gloss=master|SpaceAfter=No
2    曰   曰   VERB    v,動詞,行為,伝達      _              0    root           _   Gloss=say|SpaceAfter=No
3    溫   溫   VERB    v,動詞,描写,形質      Degree=Pos     0    root           _   Gloss=warm|SpaceAfter=No
4    故   故   NOUN    n,名詞,時,*           Case=Tem       3    obj            _   Gloss=former|SpaceAfter=No
5    而   而   CCONJ   p,助詞,接続,並列      _              6    cc             _   Gloss=and|SpaceAfter=No
6    知   知   VERB    v,動詞,行為,動作      _              3    conj           _   Gloss=know|SpaceAfter=No
7    新   新   VERB    v,動詞,描写,形質      Degree=Pos     6    obj            _   Gloss=new|SpaceAfter=No
8    可   可   AUX     v,助動詞,可能,*       Mood=Pot       11   aux            _   Gloss=possible|SpaceAfter=No
9    以   以   VERB    v,動詞,行為,動作      _              8    fixed          _   Gloss=use|SpaceAfter=No
10   為   爲   VERB    v,動詞,存在,存在      VerbType=Cop   11   cop            _   Gloss=be|SpaceAfter=No
11   師   師   NOUN    n,名詞,人,役割        _              0    root           _   Gloss=teacher|SpaceAfter=No
12   矣   矣   PART    p,助詞,句末,*         _              11   discourse:sp_      Gloss=[PFV]|SpacesAfter=\n

# newpar text = 子曰溫故而知新可以為師矣
# text = 子曰
1    子   子   NOUN    n,名詞,人,人          _              2    nsubj          _   Gloss=master|SpaceAfter=No
2    曰   曰   VERB    v,動詞,行為,伝達      _              0    root           _   Gloss=say|SpaceAfter=No

# text = 溫故而知新
1    溫   溫   VERB    v,動詞,描写,形質      Degree=Pos     0    root           _   Gloss=warm|SpaceAfter=No
2    故   故   NOUN    n,名詞,時,*           Case=Tem       1    obj            _   Gloss=former|SpaceAfter=No
3    而   而   CCONJ   p,助詞,接続,並列      _              4    cc             _   Gloss=and|SpaceAfter=No
4    知   知   VERB    v,動詞,行為,動作      _              1    conj           _   Gloss=know|SpaceAfter=No
5    新   新   VERB    v,動詞,描写,形質      Degree=Pos     4    obj            _   Gloss=new|SpaceAfter=No

# text = 可以為師矣
1    可   可   AUX     v,助動詞,可能,*       Mood=Pot       4    aux            _   Gloss=possible|SpaceAfter=No
2    以   以   VERB    v,動詞,行為,動作      _              1    fixed          _   Gloss=use|SpaceAfter=No
3    為   爲   VERB    v,動詞,存在,存在      VerbType=Cop   4    cop            _   Gloss=be|SpaceAfter=No
4    師   師   NOUN    n,名詞,人,役割        _              0    root           _   Gloss=teacher|SpaceAfter=No
5    矣   矣   PART    p,助詞,句末,*         _              4    discourse:sp_      Gloss=[PFV]|SpacesAfter=\n
```

Figure 2. Sentence segmentation for 子曰溫故而知新可以為師矣 under CoNLL-U format

3. IMPLEMENTATIONS AND RESULTS

In order to annotate UD_Classical_Chinese-Kyoto treebank, we developed CoNLL-U Visualiser and Editor. The Visualiser and Editor are based upon the techniques of SVG (Scalable Vector Graphics), they interactively work with JavaScript on Web-browsers (Figure 3), and they store all data of the treebank into our UD-Kanbun GitLab server at https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun/ud-kanbun/. The server is open to public, so that anyone can download the treebank as well as the Visualiser and Editor.

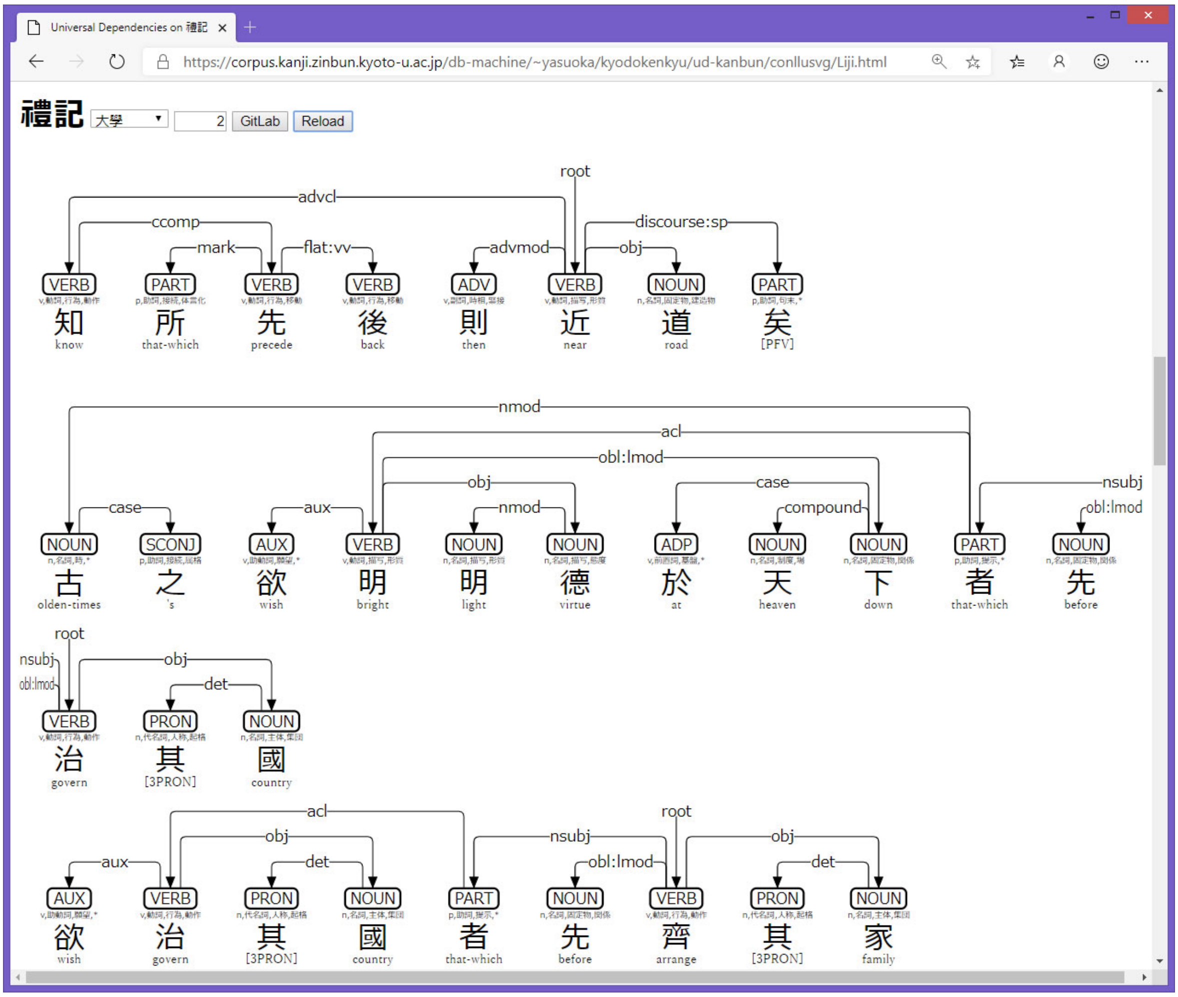


Figure 3. CoNLL-U Visualiser and Editor running on Web-browser

In order to analyse classical Chinese texts, the author has released a python-based natural language processor at https://pypi.org/project/udkanbun which is based upon MeCab-Kanbun, adding the techniques of Straka and Straková (2019). Straka and Straková's linguistic model "classical_chinese-kyoto-ud-2.4" is built up from our UD_Classical_Chinese-Kyoto treebank with the techniques of deep learning, and their model can be used for tokenisation, POS-tagging, lemmatisation, dependency parsing, and sentence segmentation (Figure 4). Its performance on the tokenisation of classical Chinese texts is in accuracy of 99.5% (F1 score). On POS-tagging 90.8% (correct UPOS ratio), on lemmatisation 99.4% (correct LEMMA ratio), on dependency parsing 56.3% (Morphology-aware Labeled Attachment Score), and on sentence segmentation 38.9% (F1 score). Since sentence segmentation needs more accuracy, the author has tentatively disabled automatic sentence segmentation on his python-based natural language processor.

4. CONCLUSION

In this paper we have discussed to apply Universal Dependencies for classical Chinese texts, especially "not-universal" treatments and enhancements on tokenisation, POS-tagging, lemmatisation, morphological features, dependency relations, and sentence segmentation. On the experimental implementation, we have obtained excellent results for automatic tokenisation and POS-tagging, good result for automatic dependency parsing, but rather poor result for automatic sentence segmentation. For the future work we will investigate

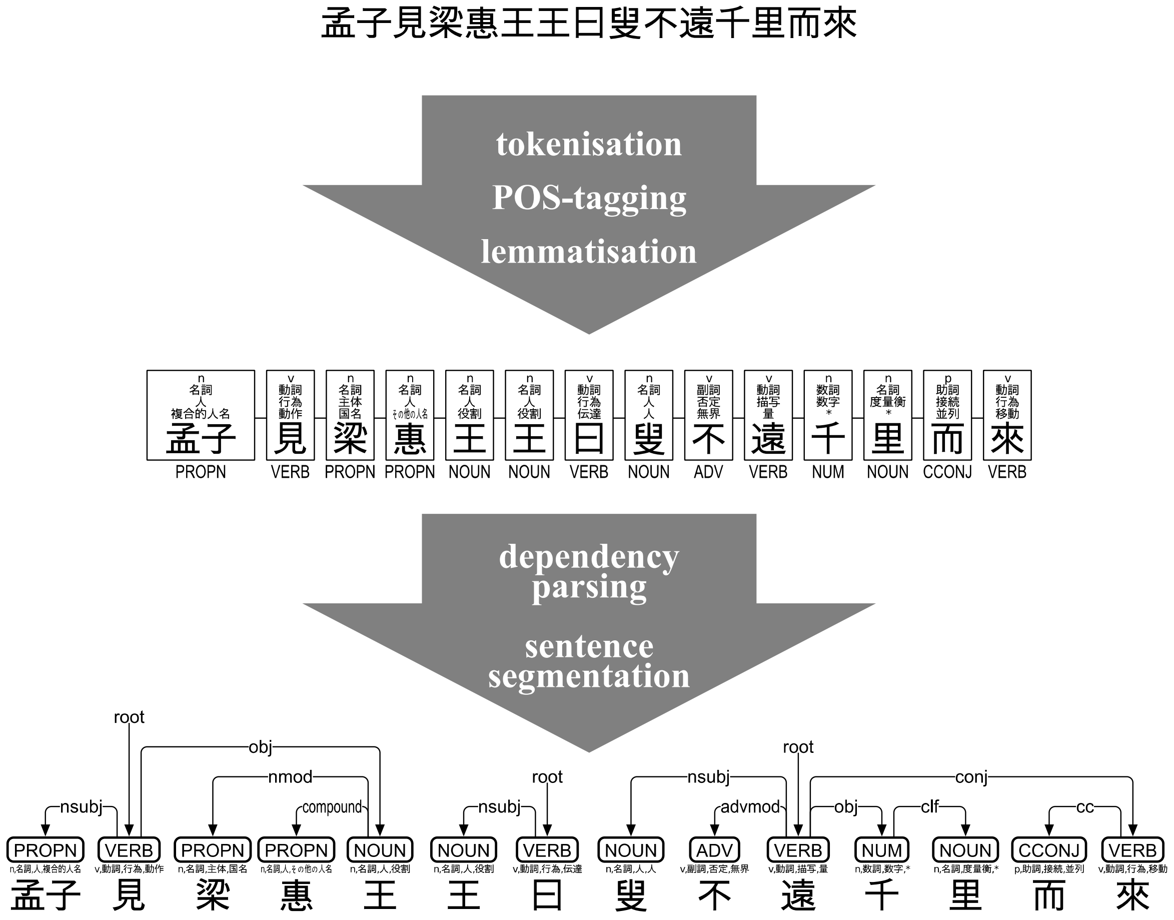

孟子見梁惠王王曰叟不遠千里而來
tokenisation
POS-tagging
lemmatisation
孟子 見 梁 惠 王 王 曰 叟 不 遠 千 里 而 來
PROPN VERB PROPN PROPN NOUN NOUN VERB NOUN ADV VERB NUM NOUN CCONJ VERB
dependency parsing
sentence segmentation
root
obj
nmod
nsubj
compound
root
nsubj
nsubj
root
advmod
conj
obj
clf
cc

Figure 4. Analysis of classical Chinese texts (overview)

automatic sentence segmentation of classical Chinese texts, applying other ideas such as 山崎直樹 (2019) or 胡韧奋 et al. (2019).

ACKNOWLEDGEMENT

The author would like to express his appreciation to the analysing team: Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, Shigeki Moro, Yuan Li, Hiroyuki Shirasu, and Kazunori Fujita. This research is supported by the Grant-in-Aid for Scientific Research No. B17H01835.

REFERENCES

胡韧奋, 李绅, 诸雨辰. 2019. 基于深层语言模型的古汉语知识表示及自动断句研究. *CCL 2019: 18th China National Conference on Computational Linguistics*.

Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing Universal Dependencies for Mandarin Chinese. *Proceedings of the 12th Workshop on Asian Language Resources*, 20–29. https://www.aclweb.org/anthology/W16-5403

Tomohiko Morioka, Christian Wittern, Koichi Yasuoka, and Naoki Yamazaki. 2013. A Study of Linguistic Analysis for Classical Chinese Texts. *International Conference on Culture and Computing 2013*, 143–144. https://doi.org/10.1109/CultureComputing.2013.37

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. *CICLing 2015: 16th International Conference on Intelligent Text Processing and Computational Linguistics*, 3–16. https://doi.org/10.1007/978-3-319-18111-0_1

Joakim Nivre, Daniel Zeman, et al. 2019. *Universal Dependencies 2.4*. LINDAT/CLARIN digital library at the Institute of Formal Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague. http://hdl.handle.net/11234/1-2988

Edwin G. Pulleyblank. 1991. *Lexicon of Reconstructed Pronunciation in Early Middle Chinese, Late Middle Chinese, and Early Mandarin*. UBC Press, Vancouver.

Milan Straka and Jana Straková. 2019. *Universal Dependencies 2.4 Models for UDPipe*. LINDAT/CLARIN digital library at the Institute of Formal Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague. http://hdl.handle.net/11234/1-2998

Christian Wittern. 2016. Special Issue: Kanseki Repository. *CIEAS Research Report 2015*, Kyoto University, Kyoto. http://hdl.handle.net/2433/210140

山崎直樹. 2019. 古典中国語のテクストをいかに切り分けるか. 開篇, *37*:111-119.

Koichi Yasuoka, Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, and Shigeki Moro. 2018. Morphological Analysis of Classical Chinese Texts and Its Application. *IPSJ Journal, 59*(2):323–331. http://id.nii.ac.jp/1001/00185742