Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) @LREC 2022, pages 23–30
Marseille, 24 June 2022

Converting the Sinica Treebank of Mandarin Chinese to Universal Dependencies

Yu-Ming Hsieh[1], Yueh-Yin Shih[2], Wei-Yun Ma[2]
Research Center for Humanities & Social Sciences, Academia Sinica
Institute of Information Science, Academia Sinica
morris@gate.sinica.edu.tw, {yuehyin, ma}@iis.sinica.edu.tw

Abstract

This paper describes the conversion of the Sinica Treebank, one of the major Mandarin Chinese treebanks, to Universal Dependencies. The conversion is rule-based and the process involves POS tag mapping, head adjusting in line with the UD scheme and the dependency conversion. Linguistic insights into Mandarin Chinese alongwith the conversion are also discussed. The resulting corpus is the UD Chinese Sinica Treebank which contains more than fifty thousand tree structures according to the UD scheme. The dataset can be downloaded at https://github.com/ckiplab/ud.

Keywords: Sinica Treebank, Universal Dependencies, conversion.

1. Introduction

The recent surge of interest in using a unified tagset and annotation guideline for treebanks of many languages has led to the speedy growing of the Universal Dependencies (UD) Project (Nivre et al., 2016). The project aims to facilitate the development of parsing technologies, enabling the use of techniques such as cross-lingual transfer. The UD version 2.9 consists of 217 treebanks in 122 languages with contributions from 477 researchers around the world.

Apart from developing treebanks by manual parsing or manual correction of automatic parsing, a UD treebank can also be automatically converted from an existing treebank, which uses a different annotation scheme (Arnardóttir et al., 2020). The present work is to convert the Sinica Treebank, in which the thematic relation between a predicate and an argument is marked in addition to grammatical category, to a UD approach treebank.

There are already 5 Mandarin Chinese UD corpora on the UD website. However, compared to other major languages, the data size for Chinese is quite small. The Sinica TreeBank has been a major Traditional Chinese Treebank developed in Taiwan and has made contribution to many NLP tasks. We hope to enlarge the usage of the Sinica treebank by converting it to the UD format and also gain some insights along with the conversion to share with the community.

The paper is structured as follows. Section 2 introduces the design and contents of the Sinica TreeBank. Section 3 describes the conversion process. The resulting corpus and the comparison with other UD Chinese treebanks are presented in Section 4. Section 5 is the conclusion and future work.

2. Sinica Treebank

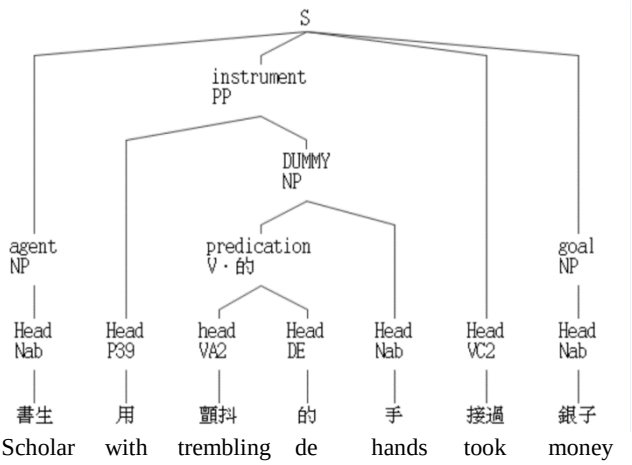
The Sinica Treebank[1] has been developed and released to public since 2000 by the Chinese Knowledge Information Processing (CKIP) group at Academia Sinica. It is one of the first structurally annotated corpora in Mandarin Chinese. Current version 3.0 (6 files) contains 61,087 structural trees and 361,934 words in Chinese. The textual material is extracted from the tagged Sinica Corpus[2] so the issues of word segmentation and category assignment are previously resolved. Based on ICG grammar (Information-based Case Grammar), the contexts are parsed by an automatic parser (Chen 1996) before human post-editing.

The structural frame of the Sinica Treebank is based on the Head-Driven Principle ; that is, a sentence or phrase is composed of a core Head and its arguments, or adjuncts. The Head defines its phrasal category and relations with other constituents. Each structural tree is annotated with words, part-of-speech of words, syntactic structure brackets, and thematic roles. The POS tagset and themantic roles are defined and explained in the CKIP technical report 93-05 (CKIP 1993) and 13-01 (CKIP 2013) respectively. An example Sinica Tree annotation and the strutural tree are presented below.

(1) 書生用顫抖的手接過銀子

“The scholar took the money with trembling hands”

S(agent:NP(Head:Nab:書生)|instrument:PP(Head:P39:用|DUMMY:NP(predication:V·的(head:VA2:顫抖|Head:DE:的)|Head:Nab:手))|Head:VC2:接過|goal:NP(Head:Nab:銀子))



In (1), the Head of the sentence is 接過 “take over” which is a transitive verb classified as VC2 in the CKIP tagset. It

[1] http://turing.iis.sinica.edu.tw/treesearch/

[2] http://asbc.iis.sinica.edu.tw/

takes two core arguments *agent* (scholar) and *goal* (silver, used as money) and an ajunct *instrument* (hand) in this case.

There are six primary phrasal categories annotated in the Sinica Treebank. S is a complete tree headed by a predicate. VP, NP and PP are phrases headed by verb (V), noun (N) and preposition (P) respectively. GP is a phrase headed by locational noun (Nc) or locational adjunct (Ng). XP is a conjunctive phrase headed by a conjunction (C) but the actual category depends on the conjoined elements.

Other non-terminal categories are phrases including structural DE, represented as {A, N, V, S, DM, GP, NP, PP, VP, ADV}·{的, 地} or 得·{V, VP, S}. In (1), for example, the phrase V·的 is the modifier of the DUMMY NP.

DUMMY is the semantic role marked on the locally undecidable categories. It is the semantic head so inherits the semantic role from the upper level. The syntactic head (Head) is distinguishable from the semantic head (head) by the first letter capitalization.

3. The Conversion

Our method of converting the Sinica Treebank to a UD corpus consists of three steps. First, we map the original POS tags of the Sinica Treebank to the UD tags. Some dependency relations also correspond to Parts of speech. Then we examine and adjust the head marking before transferring the dependencies. Finally, we replace the semantic relations with the UD dependencies by transferring rules.

3.1 POS Tag Conversion

To obtain the UD tags, a word's original tag from the Sinica Treebank is used along with transferring rules, which map each original tag to a corresponding UD tag. The correspondent POS tags between two systems are shown in Table 1. There are 15 out of 17 UD tags adopted in our system. Two UD tags, SYM (symbol) and PUNCT (punctuation), are excluded due to the design of the Sinica TreeBank. Texts were cleaned up by removing non-word symbols before annotating. Foreign words do exist but are annotated according to their actual usage. For example, CNN (Cable News Network) gets a "Nba" POS tag. As a result, SYM is useless in our system. As for PUNCT, the discussion is in the subsection below.

3.1.1 Punctuation

In the Sinica Treebank, texts are segmented not only by period, question marks and exclamation marks, but also commas, colons and semicolons, resulting the possible incompleteness of sentences. That is, there are sentence trees as well as phrase trees in the corpus. Most punctuations are not included in our system except "、" Dun Hao, a special Chinese punctuation which is also translated as comma, just like the ",". To avoid the confusion between the two distinct punctuations in Chinese texts, we use "Dun Hao" instead of "comma" in this paper. Dun Hao is indeed a punctuation but functions as coordinating words like "and" and "or". Therefore, the Sinica POS category for Dun Hao is "Caa" (coordinating conjunctions) and be transferred to CCONJ in the resulting UD corpus.

3.1.2 POS and Dependency Correspondence

As shown in Table 1, some dependencies are also available according to their lexical categories. The direct mapping conducts mainly for modifier words and function words. Words which belong to the lexical categories A (adjective), C (conjunction), D (adverbial), P(preposition), I (interjection), T (sentence -final particle), and some sorts of Nouns in the CKIP 93-05 yield direct dependencies according to their POS. For example, conjunction words (Caa) such as 和 "and" and 或 "or" always have the dependency *cc* to their governors/heads and copula 是 (VH_11) is always marked as a *cop*.

SINICA POS	UPOS	Dependency
A	ADJ	amod
Caa	CCONJ	cc
Cab (等、等等、之類)	X	conj
Cbaa, Cbab, Cbba, Cbbb, Cbca, Cbcb	SCONJ	mark
Da, Dbb,Dbc, Dc, Dd Dfa,Dfb, Dg,Dh,Dj,Dk,	ADV	advmod
Dbaa, Dbab	AUX	aux
Di		aux:aspect
P02 (in short BEI construction)		aux :pass
Naa, Nab, Nac, Nad, Naea, Naeb, Ncb, Ncc, Ncda, Ncdb, Ndaaa, Ndaad, Ndaba, Ndabb, Ndabc, Ndabe, Ndabf, Ndca, Ndcb, Ndcc Nv1, Nv2, Nv3, Nv4	NOUN	
DM		det/nummod
Nfa, Nfb, …, Nfi		clf
Nba, Nbc, Nca, Ndaab, Ndaac	PROPN	
Nep, Nes	DET	det
Neqa, Neu	NUM	nummod
Neqb	NUM	nummod:post
Ng	ADP	case:post
Nhaa,Nhab, Nhac,Nhb,Nhc	PRON	
I	INTJ	discourse
P01, P02, P03, ...P66	ADP	case
VA*,VB*, VC*, VD*, VE*, VF*, VG*, VH*, VI*, VJ*, VK*, VL* (* =1 or 2 digit numbers) V_12, V_2	VERB	
V_11	AUX	cop
Ta, Tb, Tc, Td	PART	discourse:sp
DE (的、地、之、得)	PART	case:de mark:adv mark:relcl mark:comp

Table 1: The mapping table of UPOS and Dependencies

Both interjection (I) and sentence-final particle (T) are discourse elements and the dependency relation is *discourse*. The difference between the two categories is that sentence-final particle (T) has an extra feature *sp* to

distinguish it from interjection. That is, (I) maps to *discourse* and (T) to *discourse:sp*.

DM is the determiner measurement compound and its dependency varies according to different conditions. For the detailed discussion, please refer to section 4.2.

3.2 Head Adjustment

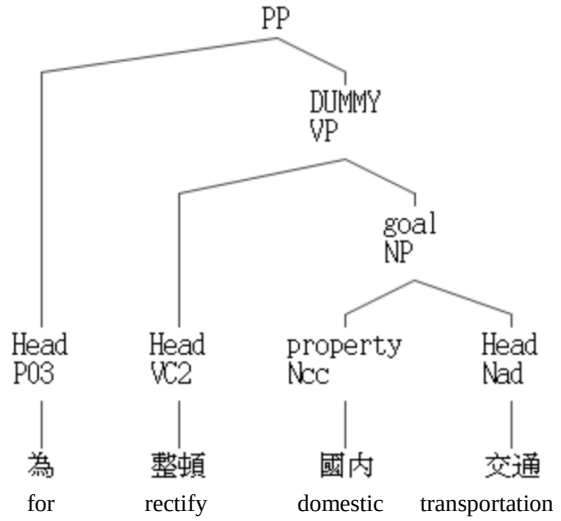
As mentioned in Section 2, the head of each phrase is already marked in the Sinica structural trees. However, the principles for determining the heads of phrases in the UD project are somewhat different from the Sinica Treebank. Some adjustments have to be made before converting. The different aspects are described below.

3.2.1 Content over Function

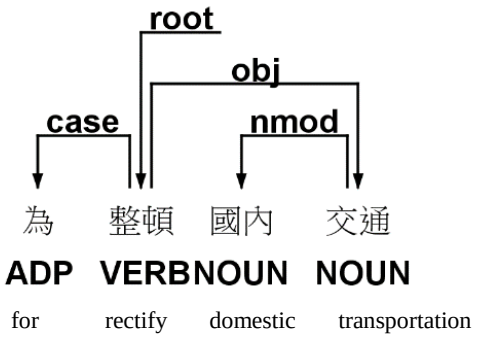
In the UD, function words attach to the content words they further specify (Nivre et.al 2016). In the Sinica Treebank, this principle also fits for the NP or VP clauses but diverges in other grammatical structures. There are two kinds of head markers in Sinica. "Head" indicates a syntactic head and "head" reveals a semantic head. In an endocentric phrasal category like NP or VP, the syntactic head and the semantic head are identical. However, in PP, GP, XP or "VP-de" constructions, the syntactic Head doesn't carry sufficient semantic information so the original Sinica annotations violate the UD principles. The conversion from Sinica to UD need to reversely choose the semantic heads as the governors of these structures. We use a PP phrase 為整頓國內交通 'for rectifying the domestic transportation'to illustrate the head ajustment.

In (2a), the original Sinica corpus marked the preposition 為 'for' to be the head. However, the UD version should reversely choose the DUMMY VP and find the VP head 整頓 'rectify' as the head.

(2a)



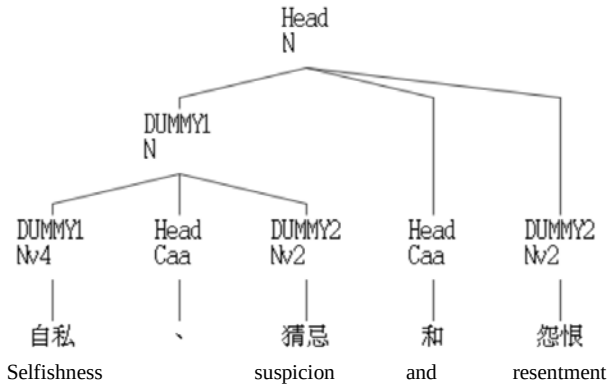
(2b)



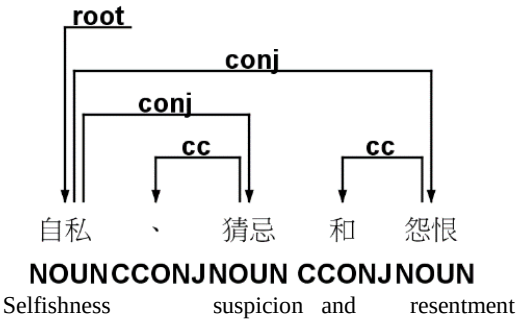
3.2.2 First Head is the Parent in Parallel-Head Constructions

The UD in principle assumes conjuncts of the coordinate structure have equal status as syntactic heads. However, the dependency tree format does not allow this analysis to be encoded directly, so the first conjunct in the linear order is by convention always treated as the parent of all other conjuncts. On the other hand, two conjuncts in Sinica are conjoined by means of a conjunction to form a new DUMMY and continue to conjoin with the right-side conjunct until achieving the rightmost one. The rightmost conjunction word is the head of the whole coordinating structure. An example demostrate in (3a) and the conversion involving both content-oriented and leftmost-dominated is shown in (3b).

(3a) 自私、猜忌和怨恨 'selfishness, suspicion, and resentment'

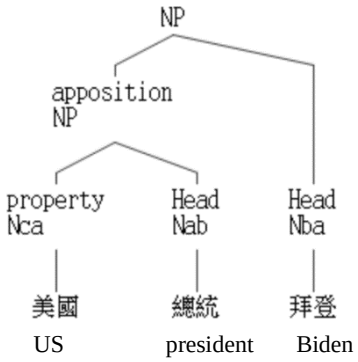


(3b)

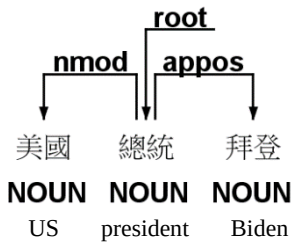


Apposition is another case of parallel-head constructions. In UD, the *appos* relation is also strictly left to right, meaning the first nominal is treated as the head. However, in Sinica, the apposition relation represents as head-final formalism and needs to reverse the head selection. The Sinica tree for the NP phrase 美國總統拜登 "U. S. President Biden" presents in (4a) and the UD version with reversed head is shown (4b).

(4a)



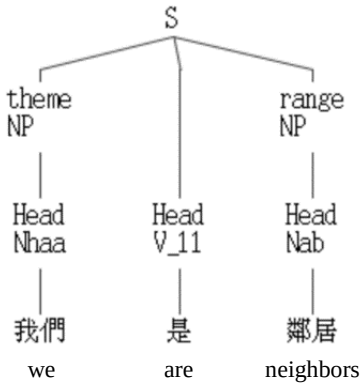
(4b)



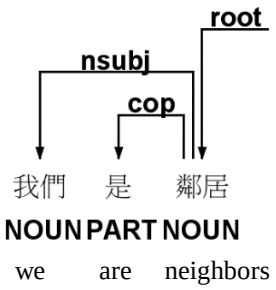
3.2.3 Copula

In the Sinica Treebank, the copula word 是 'SHI' is classified as a verb and its POS tag is "V_11". It functions as the head of a clause just like other verbal predicates and takes two arguments which are "theme" and "range". An example is shown in (5). In the UD scheme, however, the head shifts to the range NP.

(5a)



(5b)

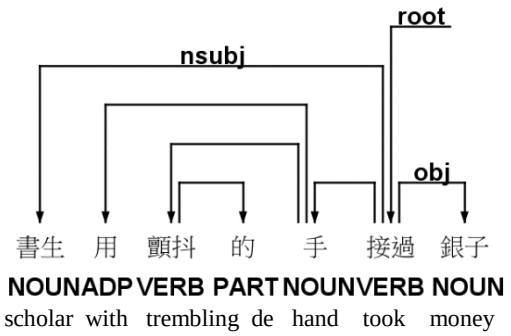


3.3 Dependency Conversion

Since the dependent relations, which are semantic roles, have already existed in the Sinica treebank, our challenge is to convert semantic-based relations to syntactic-based UD dependencies. As stated in section 3.1.2 and shown in Table 1, for most function words and modifiers, assigning dependecies according to the lexical categories shows a better result than a direct relation-dependency mapping. For example, "quantifier" sometimes transfers to *det* but sometimes to *nummod*. By mapping Nep, Nes, and DM to *det* and Neu and Neqa to *nummod*, the ambiguilty is solved.

As for the core arguments, it depends on the root of a tree to assign the proper dependencies. For example, in (1) the root is 接過 "take over" and the arguments a*gent* and *goal* should be converted to *nsubj* and *obj* respectively, as shown in (6).

(6)



3.3.1 Conversion According to the Sentence Patterns of Each Category

However, it is possible to convey a concept with different surface structures. Consider the sentences (7) and (8) below:

(7) 老師罵學生們 'The teacher scolded students'

(8) 學生們被老師罵 'Students are scolded by the teacher'

They both convey the same meaning and 老師 'teacher'is the *agent* and 學生們'students' is the *goal* of the scolding event. Clause (8) is the passive way of saying clause (7) and the object (*goal*) of the active clause becomes the subject (*goal*) of the passive clause. As a result, we also need rules to deal with such situation and a dependency *nsubj:pass* is introduced to mark the subjects of passive clauses. Thanks to the earlier work that has been done and recorded in CKIP 93-05, the possible sentence patterns for each verb category have been analysized in detail and we can make use of the analysis to produce the transferring

rules. We take VC2 for example to demostrate the conversion.

There are five sentence patterns for VC2. The first is the active sentence pattern in which *agent* is in the subject position and *goal* is in the object position. The second is the BA (把) construction of Chinese in which the *goal* is led by a preposition and be put right after the subject. In other word, the BA construction turns the word order from SVO to SOV. The third is BEI (被) construction with the reverse order of *agen*t and *goal*. The last two can be seen as the special cases of BA/BEI constructions (把/被句) with an extra argument *theme* which is a part of arguemnt *goal*. The sentence patterns and corresponding UD dependencies are listed below:

1. AGENT[{NP,PP}]〈 *〈GOAL[NP]
   → (case) nsubj〈 root〈 obj
2. AGENT [NP,PP]〈GOAL [PP]〈 *
   → (case) nsubj 〈 case obl:patient 〈 root
3. GOAL[NP]〈 AGENT [{PP,P}]〈 *
   → nsubj:pass〈case obl:agent / aux〈root
4. GOAL[NP]〈AGENT[{PP,P]〈 *〈 THEME [NP]
   →nsubj:pass〈 case obl:agent / aux〈root〈obj
5. AGENT [NP,PP]〈 GOAL [PP]〈 *〈 THEME [NP]
   → (case) nsubj〈 case obl:patient〈 *〈 obj

3.3.2 Conversion of Phrases including Structural Particle DE

The structural particle {的、地、得}DE are widely used and can be classified as four types:

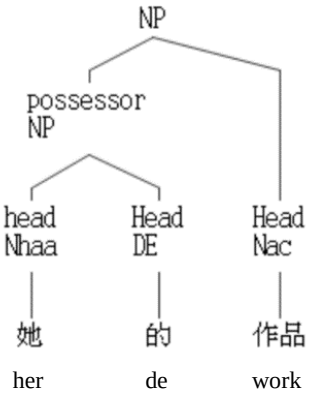
1. possessive DE : 她的作品 'her works'
2. attributive DE :美妙 的 聲音 'beautiful sound'
3. adverbial DE 大聲地叫 'shout loudly'
4. complement DE :踩得水花四濺 'stamp one's feet to make water splashes'

By analysing the sinica tree structures of DE phrases, we can gain the following converting rules.

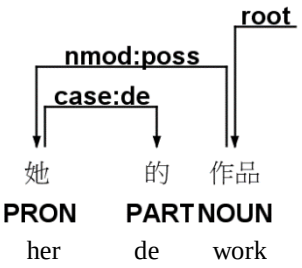
- possessor:{N,NP}·的
  → nmod :poss < case:de
- property:{N,NP, GP, PP, DM}·DE
  → nmod < case :de
- {property,predication}:{A, V, VP, S, VP, ADV}·DE
  → acl < mark:relcl
- manner:{A, ADV, DM, V,VP}·DE}
  → advcl < mark:adv
- complement:得·{V, VP, S}
  → mark:comp <xcomp (if V orVP)
  → mark:comp <ccomp (if S)

Examples of possessive DE 她的作品 'her works' shows the orginal Sinica Tree in (9a) and the converted UD tree in (9b). Similarly, complement DE 踩得水花四濺 'stamp one's feet to make water splashes' also presented in (10a) and (10b) below.

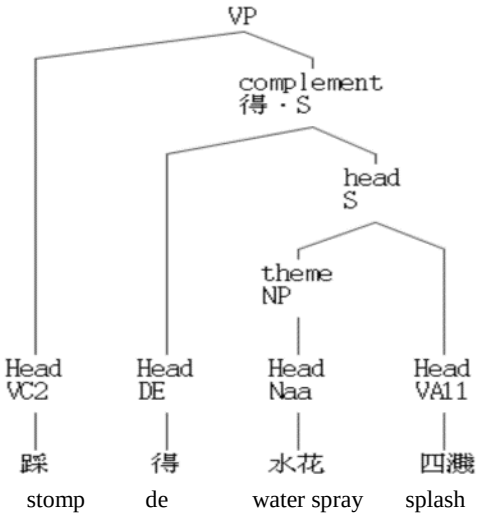
(9a)



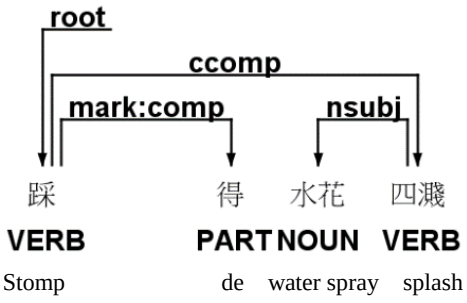
(9b)



(10a)



(10b)



3.4 Evaluation

To evaluate accuracy of the automatic conversion, we manually annotated a selection of 183 trees from the Sinica Treebank. There are 62 sentences manually selected to cover a wide range of syntactic categories which yield different constructions. The remaing 121 trees which have consecutive id numbers in our corpus are also selected to check the accuracy of the conversion. In total, the selection includes 966 tokens and the average tokens per tree is 5.28.

The evaluation results are showed in Table 2. The 121 trees with a lower average tree length (4.66 tokens per tree) reach 92.55% accuracy. However, the result of the 62 hand-picked sentences drops down to 83.83% accuracy because of the longer tree length (6.48 tokens per tree) and more complex sentence structures. The overall accuracy is 0.89 for all 966 tokens. Since the average tree length for the Sinica Treebank is 6.28 tokens per tree (shown in Table 3), which is lower than the hand-picked trees in the evaluation, we expect the accuracy of the whole converted corpus might be slightly higher than 83.83%.

	121 sentences	62 hand-picked	all selected sentences
Aver. tree length	4.66	6.48	5.28
All token numbers	564	402	966
Right-converted	522	337	859
Accuracy	92.55%	83.83%	89.93%

Table 2 : The evaluation results for 183 selected trees

4. The UD Chinese Sinica Treebank

After the conversion process mentioned above has done, the output of resulting corpus in the CoNLL-U format is illustrated in (11).

(11) The scholar took the money with trembling hands.

1	書生	_	N	Nab	_	NOUN	6	_	nsubj	_
2	用	_	P	P39	_	ADP	5	_	case	_
3	顫抖	_	V	VA2	_	VERB	5	_	acl	_
4	的	_	DE	DE	_	PART	3	_	mark:relcl	_
5	手	_	N	Nab	_	NOUN	6	_	obl	_
6	接過	_	V	VC2	_	VERB	0	_	root	_
7	銀子	_	N	Nab	_	NOUN	6	_	obj	_

While the UD POS tags we adopt is similar to other UD Chinese corpora, we have a smaller set of dependencies. The reason is that the textual material of the Sinica treebank is extracted from the tagged Sinica Corpus which has undergone post-editing and compound words and multiword expressions are in principle treated as a unit before trees are drawn, regardless the internal structure of these elements. Also, shorter sentence length results in simpler dependency relations. Currently, there are 23 UD main dependencies as well as 13 subtypes. The alphabetical list of Sinica UD dependency relations explains as follows.

- acl (clausal modifier of noun)
- advcl (adverbial clause modifier)
- advmod (adverbial modifier)
- amod (adjectival modifier)
- appos (appositional modifier)
- aux (auxiliary)
- aux:aspect (aspect auxiliary)
- aux:pass (passive auxiliary)
- case (case marker)
- case:de (case marker for possessive DE)
- case:post (localizer)
- cc (coordinating conjunction)
- ccomp (clausal complement)
- clf (classifier)
- conj (conjuct)
- cop (copula)
- csubj (clausal subject)
- csubj:pass (passive clausal subject)
- det (determiner)
- discourse (discourse element)
- discourse:sp (sentence-final particle)
- dislocated (dislocated/topicalized element)
- iobj (indirect object)
- mark (subordinating marker)
- mark:relcl (mark relative clause)
- mark:adv (mark adverbial)
- mark:comp (mark complement caluse)
- nmod (nominal modifier)
- nmod:poss (possessive nominal modifier)
- nmod:tmod (temporal modifier)
- nummod (numeric modifier)
- nummod:post (post quantifier)
- obj (object)
- obl (oblique)
- obl:agent (agent modifier)
- obl:patient (patient modifier)
- root (root)
- xcomp (open clausal complement)

Because of the complexity of language phenomena, it is impossible to have rules covering all the circumstances of linguistic expression. About 2% of Sinica trees cannot be fully transferred to the UD style annotation. Post-editing is needed for these unsuccessful trees. The list of dependency relations still keeps on updating.

Although this is still an ongoing work, we have found some aspects worth discussing. In comparison with two other UD Chinese research teams, namely Google and City University of Hong Kong, the following issues are described below.

4.1 Tree Length

Due to the phrase/sentence segmentation principle of the Sinica Treebank, some trees have one word only so just have the root without any branches (dependencies). We, therefore, remove the one-word trees and the remaining tree number is 53,548. The sentence length is 6.28 tokens on average, which is a lot shorter than other UD Chinese treebanks.

Table 3 is the comparison of tree length (tokens per tree) among UD Chinese corpora. .Obviously, the sentence segmentation principle of the Sinica Treebank is the main reason to gain shorter sentences. However, for the sake of natural language understanding, shorter sentences are easier to process both for humans and for machines. By looking into the existing UD corpora, the annotations for super long sentences are quite often questionable.

Moreover, the punctuation usage in Chinese is not so strict, resulting in the misuse between comma and period in texts. Some relations should be treated in the discourse level rather than in the syntactic level.

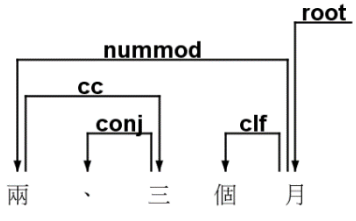
Corpus	Sentences	Tokens	Average Sentence Length
GSD	4,997	123,291	24.67
PUD	1,000	21,415	21.41
CFL	451	7,256	11.09
HK	1,004	9,874	9.83
Sinica	53,548	336,281	6.28

Table 3 : The comparison of tree length among UD Chinese corpora

4.2 Classifier

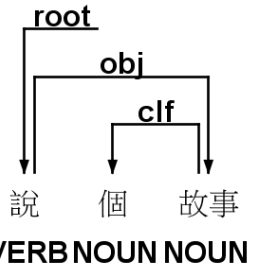
Classifiers are a special lexical category in Chinese. They are often obligatory in a noun phrase with a numeral modifier and optional with a demonstrative. The two UD Chinese research teams treat classifiers differently. We consider both approaches and make our own choice to fit the tree structures of the Sinica Treebank in which DM is the determiner measure compound/phrase. For the simple (a numeral/demostrative and a classifier) construction, the two elements are grouped together to form a DM. The whole DM depends on the Head of an NP it belongs to. The relation is either *det* or *nummod*, decided by the determiner types. However, if there are more than one numerals as in (12) or zero numeral as in (13), the classifier itself has a *clf* relation to its head Noun.

(12)



two three classifier months

(13)



say classifier story

Differences in marking the relation of classifiers and its surrounding elements between each reseach team are shown in Table 4.

Team	condition	Determiner	Classfier
Google	$1^{(+)}$determiner	H: classfier R:nummod/det	H: head of NP R: clf
	0 determiner	n/a	H: head of NP R: clf
HK	$1^{(+)}$determiner	H: head of NP R:nummod/det	H: determiner R: clf
	0 determiner	n/a	H: head of NP R : det
Sinica	DM	H : head of NP R :nummod or det	
	$2^{(+)}$determiners	H : head of NP R :nummod	H : head of NP R :clf
	0 determiner	n/a	H : head of NP R :clf

Table 4: The comparison of classifiers between 3 research teams

5. Conclusion and Future Work

We have presented the process of converting the Sinica Treebank to the UD annotation scheme. It is an attemp to create the Chinese language resource in a universally accepted format so that the long-standing Sinica Treebank can be more usable for a variety of multiligual NLP tasks. The conversion was a challenging task and there is still quite a few works to be done.

More complete converting rules have to be discovered and added to the current system. Features are not included in this version of converting corpus. We will consider the necessity of adding features and make this corpus more compatible to other UD corpora. Finally, to make this corpus more competitive to others, more complete sentences are required. Since sentences are composed of phrases, the methods of finding adjacent phrases and the replacement of some dependency relations due to the composition are worth investigating.

6. Bibliographical References

Chen, K.-J., Huang C.-R, Chang, L.-P. Hsu, H.-L. (1996). Sinica Corpus: Design Methodology for Balanced Corpora. Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation (PACLIC II), Seoul Korea. p. 167–176.

Chen, K.-J., Luo, C.-C. Chang, M.-C., Chen, F.-Y., Chen, C.-J., Huang, C.-R., Gao. Z.-M. “Sinica Treebank: Design Criteria, Representational Issues and Implementation”. In Book “Treebanks — Building and Using Parsed Corpora”, Ch. 13, pp. 231–248, 2003.

CKIP (1993). Lexical Category Analysis of Mandarin Chinese. CKIP Technical Report 93-05

CKIP (2013). Semantic roles of Sinica Treebank. CKIP Technical Report 13-01.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016): 1659-1666.

Nivre, J., M.-C. de Marneffe, Ginter, F., Hajič, Manning, C., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D.

(2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the 12th International Conference on Language Resources and Evaluation. (LREC 2020)

Poiret, R., Wong, T S., Lee, J. et al. Universal Dependencies for Mandarin Chinese. Lang Resources & Evaluation (2021). https://doi.org/10.1007/s10579-021-09564-2

Arnardóttir, Þ., Hafsteinsson, H., Sigurðsson, E., Bjarnadóttir, K., Ingason, A., Jónsdóttir, H., Steingrímsson, S. (2020). A Universal Dependencies Conversion Pipeline for a Penn-format Constituency Treebank. In Proceedings of the Fourth Workshop on Universal Depencies. (UDW 2020)